Electronic Journal of Statistics
Vol. 12 (2018) 4487–4516
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1514

Consistency of logistic classifier in abstract Hilbert spaces

Agne Kazakeviciute*

Brunel University London, United Kingdom

and

Malini Olivo

Agency for Science, Technology, and Research, Singapore, and the School of Physics, National University of Ireland, Galway, Ireland.

Abstract: We study the asymptotic behavior of the logistic classifier in an abstract Hilbert space and require realistic conditions on the distribution of data for its consistency. The number k_n of estimated parameters via maximum quasi-likelihood is allowed to diverge so that $k_n/n \to 0$ and $n\tau_{k_n}^4 \to \infty$, where n is the number of observations and τ_{k_n} is the variance of the last principal component of data used for estimation. This is the only result on the consistency of the logistic classifier we know so far when the data are assumed to come from a Hilbert space.

Keywords and phrases: Classification, consistency, functional data analysis, logistic classifier.

Received May 2018.

1. Introduction

Functional Data Analysis (FDA) is an active research area in statistics that includes a collection of theorems and methods for dealing with infinite-dimensional (functional) data (see Ramsay & Silverman (2002) and Ramsay & Silverman (2005) for an overview). Classification of functional data is one of the hottest topics in FDA and establishing consistency of various classifiers for functional data has been of a great research interest for more than a decade.

Most of classifiers assign an observation to the class with the largest estimated posterior probability. Consistency of such a classifier is then implied by the consistency of the estimate of that probability. If it depends on a finite number of unknown parameters, as in the logistic model in $\mathbb{R}^k$, then it suffices to consistently estimate all the parameters. For example, in the $\mathbb{R}^k$ case the logistic classifier has been proved to be consistent, strongly consistent (see, e.g. Chen et. al. (1999)) and even uniformly consistent Kazakeviciute & Olivo (2016).

The situation becomes more complicated if conditional probability is modelled by the infinite number of parameters, as in the logistic model in an infinite-dimensional Hilbert space E. In this case we are given independent observations

*e-mail: agne.kazakeviciute@brunel.ac.uk

$(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y), where X is E-valued random variable and $Y \in \{-1, 1\}$ is its associated class label. Usually, then the following 3-step procedure is used: (1) some orthonormal basis in E is chosen and the observations are replaced by their coefficients in that basis (a finite number, say, l of coefficients are retained), (2) the principal component analysis of the obtained $n \times l$ array of data is performed and the first k principal components are retained, (3) the usual logistic regression on the new $n \times (k+1)$ array of data is performed. From the mathematical point of view this means that we replace the original observations by their orthogonal projections in some k-dimensional subspace $E_k \subset E$ and find the estimate $\hat{\theta}_{kn}$ of the unknown parameter $\theta_0 \in E$, which maximizes the quasi-likelihood over all $\theta \in E_k$. Of course, if we want to analyze asymptotic properties of such an estimator (and of the corresponding classifier, based on that estimator), we should also assume that k depends on n, that is, the final estimator to be analyzed is $\hat{\theta}_{k_n n}$ for some sequence $k_n \to \infty$.

Note that if E_k is obtained by the procedure described above, then it is a random subspace of E (it depends on data). This makes the analysis of $\hat{\theta}_{k_n n}$ rather complicated. Therefore here we will analyze the simpler case where E_k are non random. Formally, this means that we omit the step of principal component analysis. This approach (call it naïve) is also known in the literature, but in some cases is not recommended for practical use. For example, Escabias et. al. (2007) argued that the naïve approach in the context of functional data introduces multicollinearity (strong dependence among predictors) which in turn causes inaccurate parameter estimates and increases their variance. However, the asymptotic results in the case where E_k are non-random in some situations are good, as we show later. Moreover, they show what can be expected in the general case because some required assumptions are likely to remain also in the general setting.

In this work we establish the consistency of the logistic classifier under the two sets of conditions. The first set consists of three conditions on the distribution of X that are rather simple and, nevertheless, sufficiently general. All three conditions are satisfied if X has a normal distribution in Hilbert space with zero mean and positive definite covariance form. The second set of conditions bound the growth rate of k_n: we require that $k_n/n \to 0$ and $n\tau_{k_n}^4 \to \infty$, where $\tau_k = \min_{\theta \in E_k, \|\theta\|=1} C(\theta, \theta)$ and C is the moment form of X defined by (3). As we later discuss, τ_k can be interpreted as the variance of the kth theoretical principal component. The first condition requires k to be asymptotically less than n which is almost necessary. The second condition suggests that the variance of the last theoretical principal component tends to 0 slower than $n^{-1/4}$, as $n \to \infty$. However, this condition can be relaxed, as our simulation study shows.

In the literature, there are limited attempts to study asymptotic behavior of logistic estimate when dimensionality k_n of data used for estimation diverges together with the sample size. For example, van de Geer (2008), Fan & Song (2010) and Wang (2011) studied related but slightly different problems, that is, models that include some kind of penalty on a parameter vector, such as Lasso. At first look it could seem that a very close attempt to solve the described problem was

the one of Liang & Du (2012), where they proved the asymptotic normality of the parameter estimate under mild conditions. However, the fundamental difference between their work and ours is that they did not consider covariates X to be random, while we do. In principle, the results for the model with non-random data can be applied also to the case where the data are random, provided that the assumptions used for non-random data are satisfied for each realization of random data. However, we cannot apply their result to solve our problem because one of their assumptions translates as $\inf_k \tau_k > 0$ which does not hold, if data come from a Hilbert space and follow normal distribution in Hilbert space. In such a situation we can always select basis system $\{e_j\}$ such that the coordinates of X are uncorrelated. Then $\sum_{j=1}^{\infty} C(e_j, e_j) = \sum_{j=1}^{\infty} \mathsf{E}X_j^2 = \mathsf{E}\|X\|^2 < \infty$. If $E_k = \left\{ \sum_{j=1}^{k} c_j e_j \mid c_1, \ldots, c_k \in \mathbb{R} \right\}$, then $\tau_k \leq C(e_k, e_k)$ and thus $\inf_k \tau_k = 0$.

The results nearest to ours are achieved in Müller & Stadtmüller (2005). In the paper, the authors studied generalized linear models with no penalty and established asymptotic normality for a properly scaled distance between the estimated and the true parameters. However, they assume (see assumption (M1)) that if $\mathsf{Var}^X Y = \sigma^2(\mathsf{E}^X Y)$ (where E^X, Var^X denote the conditional mean and conditional variance, given X, respectively) then the function σ is bounded away from 0: $\sigma^2(\mu) \geq \delta > 0$ for all μ. This is not the case for logistic regression model, where $\sigma^2(\mu) = \mu(1-\mu)$. This means that the results in Müller & Stadtmüller (2005) cannot be applied to prove the consistency of logistic classifier as considered in this work. Moreover, Müller & Stadtmüller (2005) approximated infinite-dimensional model by a finite-dimensional one, that is, they assumed that the distribution of Y depends on the projection of θ_0 onto some subspace E_k rather than on full $\theta_0 \in E$, and assumed that the error of such an approximation tends to 0. However, we could not find any proof of the latter rather complicated statement. No such approximation is involved in our work.

Our paper is organized as follows. In Section 2 we describe the statistical problem considered, explicitly state the assumptions, give some discussion on them, and state our main result. In Section 3 we provide a simulation study to check the necessity of the assumptions and we end this work with a brief discussion in Section 4. All proofs are left for Section 5.

2. Consistency

Let E be a separable Hilbert space with the inner product $\langle \cdot, \cdot \rangle$. Let $X \in E$ be a Hilbert space-valued random variable and Y a random variable, gaining values -1 and 1, with conditional probabilities (w.r.t. X) being $1 - p_{\theta_0}(X)$ and $p_{\theta_0}(X)$, respectively. Here $\theta_0 \in E$ is an unknown parameter and

$$p_\theta(x) = \frac{1}{1 + \mathrm{e}^{-\langle \theta, x \rangle}}, \ \theta, x \in E.$$

For example, if $E = \ell^2$, the space of all square-summable sequences, then $\langle \theta, x \rangle = \sum_{k=1}^{\infty} \theta_k x_k$. If $E = L^2([0,1])$, then $\langle \theta, x \rangle = \int_0^1 \theta(t)x(t)\mathrm{d}t$. Since E can be any Hilbert space, we will work with the general notation $\langle \theta, x \rangle$ instead.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from the distribution of (X, Y). For $\theta, x \in E$ and $y \in \{-1, 1\}$ define

$$m_\theta(x, y) = \log\left(1 + \mathrm{e}^{-y\langle \theta, x \rangle}\right)$$

and denote

$$M_n(\theta) = \overline{m_\theta(X, Y)} = \frac{m_\theta(X_1, Y_1) + \cdots + m_\theta(X_n, Y_n)}{n}, \quad M(\theta) = \mathsf{E} m_\theta(X, Y).$$

Note that

$$\begin{aligned} M_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \mathrm{e}^{-Y_i\langle \theta, X_i \rangle}\right) = \frac{1}{n}\log\prod_{i=1}^{n}\left(1 + \mathrm{e}^{-Y_i\langle \theta, X_i \rangle}\right) \\ &= -\frac{1}{n}\log\prod_{i=1}^{n} q_\theta(X_i, Y_i), \end{aligned}$$

where

$$q_\theta(X_i, Y_i) = \frac{1}{1 + \mathrm{e}^{-Y_i\langle \theta, X_i \rangle}}.$$

Obviously, $q_\theta(X_i, 1) = p_\theta(X_i)$ and $q_\theta(X_i, -1) = 1 - p_\theta(X_i)$. Also, for any bounded f,

$$\begin{aligned} \mathsf{E} f(X, Y) &= \int f(x, 1) q_\theta(x, 1)\mu(\mathrm{d}x) + \int f(x, -1) q_\theta(x, -1)\mu(\mathrm{d}x) \\ &= \int f(x, y) q_\theta(x, y)\mu(\mathrm{d}x)\nu(\mathrm{d}y), \end{aligned}$$

where ν is a counting measure in the set $\{-1, 1\}$. Therefore $q_\theta(x, y)$ is a density of (X, Y) w.r.t. the measure $\mu \times \nu$. Hence, since μ is unknown, $M_n(\theta)$ can be interpreted as the logarithm of the quasi-likelihood function, multiplied by $-1/n$.

Naturally, for various practical tasks it is of great interest to provide an estimate of p_θ.

Let (E_k) be some fixed sequence of the linear subspaces of the space E such that the following conditions are satisfied: (1) $\dim E_k = k$ for all k, (2) $E_k \subset E_{k+1}$ for all k, and (3) $\overline{\bigcup_k E_k} = E$. For any k and n define

$$\hat{\theta}_{kn} = \arg\min_{\theta \in E_k} M_n(\theta). \tag{1}$$

Note that taking $\theta \in E_k$ in the above expression introduces some approximation error. To force this error to tend to 0 as n diverges, fix some sequence (k_n) and set

$$\hat{\theta} = \hat{\theta}_{k_n n} \quad \text{and} \quad \hat{p} = p_{\hat{\theta}}. \tag{2}$$

We will call $\hat{p}$ the *logistic estimate* of the true conditional probability p_{θ_0}. For example, let $E = L^2(T)$ with the usual inner product

$$\langle \theta, x \rangle = \int_T \theta(t)x(t)\mathrm{d}t,$$

where $T \subset \mathbb{R}$ is an interval and L^2 is the space of square integrable real functions on T endowed with the usual inner product $\langle x_1, x_2 \rangle = \int_0^1 x_1(t)x_2(t)\mathrm{d}t$. The standard method for obtaining logistic estimate from a given sample (X_1, Y_1), $\ldots, (X_n, Y_n)$ is expanding X and θ via selected basis functions $\{e_j\}$

$$X_i(t) = \sum_{j=1}^{\infty} X_{ij} e_j(t), \quad \theta(t) = \sum_{j=1}^{\infty} \theta_j e_j(t),$$

choosing $k = k_n$ and then using (1), where

$$E_k = \left\{ \sum_{j=1}^{k} c_j e_j \mid c_1, \ldots, c_k \in \mathbb{R} \right\}.$$

We consider the following statistical task. We want to estimate the unknown true conditional probability p_{θ_0}, given the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the distribution of (X, Y). The quality of the estimate $\hat{p}$ is assessed by the risk $\mathsf{E}|\hat{p}(X) - p_{\theta_0}(X)|$. If the risk tends to 0, as $n \to \infty$, the estimate $\hat{p}$ is called *consistent*. It is well known that if $\hat{p}$ is consistent, then the empirical classifier, which assigns x to the class 1 whenever $\hat{p}(x) > 1/2$, is also consistent (see, e.g., van Ryzin (1966) or Kazakeviciute et. al. (2017)). Here we will consider the logistic estimate (2), where we suppose that $\hat{\theta}_{kn} = 0$, if the minimum is not attained or is not unique.

We will say that the distribution of X is *of full rank*, if $\mathsf{P}(\langle \theta, X \rangle = 0) = 0$, for all $\theta \neq 0$. Also, recall that any family of random variables (Z_s) is called *uniformly integrable*, if

$$\sup_s \mathsf{E}|Z_s| 1_{\{|Z_s| > c\}} \xrightarrow[c \to \infty]{} 0.$$

The consistency of the logistic estimate will be proved under the following assumptions on the distribution of X:

(FR) The distribution of X is of full rank.
(M) $\mathsf{E}\|X\|^4 < \infty$.
(UI) The family of random variables $(\langle \theta, X \rangle^2 / \mathsf{E}\langle \theta, X \rangle^2 \mid \|\theta\| = 1)$ is uniformly integrable.

Assumption (M) implies that the mean of X and the second moment form of X are correctly defined. The mean is the only such vector $\mathsf{E}X$ from E that $\langle\theta, \mathsf{E}X\rangle = \mathsf{E}\langle\theta, X\rangle$ for all $\theta \in E$. The second moment form is defined by

$$C(\theta_1, \theta_2) = \mathsf{E}\langle\theta_1, X\rangle\langle\theta_2, X\rangle. \tag{3}$$

If $\mathsf{E}X = 0$, (3) is called the *covariance form*. For example, if $E = L^2([0;1])$, then

$$C(\theta_1, \theta_2) = \mathsf{E}\int_0^1 \theta_1(s)X(s)\mathrm{d}s \int_0^1 \theta_2(t)X(t)\mathrm{d}t = \int_0^1 \mathrm{d}s \int_0^1 \theta_1(s)\theta_2(t)\tilde{C}(s,t)\mathrm{d}t,$$

where $\tilde{C}(s,t) = \mathsf{E}X(s)X(t)$ is a covariance function of the process X. If $E = \ell^2$ and x_i denote the coordinates of $x \in \ell^2$, then

$$C(\theta_1, \theta_2) = \mathsf{E}\sum_{i=1}^{\infty} \theta_{1i}X_i \sum_{j=1}^{\infty} \theta_{2j}X_j = \sum_{i,j} \theta_{1i}\theta_{2j}c_{ij},$$

where (c_{ij}) is a covariance matrix of the random vector X. Since E can be any abstract Hilbert space, we will work with the general notation $C(\theta_1, \theta_2)$.

The second moment form is a continuous bilinear form on E. Moreover, it is symmetric and positive semi-definite, that is, for all θ,

$$C(\theta, \theta) = \mathsf{E}\langle\theta, X\rangle^2 \geq 0.$$

Obviously, $C(\theta, \theta) = 0$ if and only if $\mathsf{P}(\langle\theta, X\rangle = 0) = 1$. This implies that $C(\theta, \theta) > 0$ if and only if $\mathsf{P}(\langle\theta, X\rangle = 0) < 1$. Recall that assumption (FR) is $\mathsf{P}(\langle\theta, X\rangle = 0) = 0$. Hence assumption (FR) is slightly stronger than the requirement that C is positive definite.

The required conditions are realistic and hold for a variety of real-life settings. For example, all three assumptions hold, if X is a normally distributed random vector with zero mean and positive definite covariance form. Indeed, then $\mathsf{E}\|X\|^s < \infty$, for all s, and

$$\sup_{\|\theta\|=1} \mathsf{E}\frac{\langle\theta, X\rangle^2}{\mathsf{E}\langle\theta, X\rangle^2} 1_{\left\{\frac{\langle\theta, X\rangle^2}{\mathsf{E}\langle\theta, X\rangle^2} > c\right\}} = \mathsf{E}Z^2 1_{\{Z^2 > c\}} \xrightarrow[c\to\infty]{} 0.$$

Here Z is a random variable that follows a standard normal distribution.

Denote

$$\tau_k = \min_{\substack{\theta \in E_k \\ \|\theta\|=1}} C(\theta, \theta). \tag{4}$$

Here C is the moment form of X, defined by (3). For example, if $E = \ell^2$, E_k satisfy the conditions mentioned above, $\mathsf{E}X = 0$, the coordinates of X are uncorrelated and the variances of them decrease, then τ_k is the variance of the kth coordinate. In other words, τ_k is the variance of the kth theoretical principal component.

Our main result is the following Theorem.

Theorem 1. *Suppose that assumptions (FR), (M) and (UI) hold. Moreover, suppose that*

$$k_n \to \infty, \quad \frac{k_n}{n} \to 0 \quad \textit{and} \quad n\tau_{k_n}^4 \to \infty,$$

as $n \to \infty$. Then the logistic estimate is consistent.

Note that the condition $n\tau_{k_n}^4 \to \infty$ requires that the data are such that the variance of the last principal component tends to 0 slower than $n^{-1/4}$, as $n \to \infty$. This in turn suggests that the data need to be such that it cannot be sufficiently explained only by a few principal components.

In statistics, the logistic model with an intercept is usually preferred over the one without it because useful model information might be incorporated in the intercept term. Theorem 1 implies the analogous result on the logistic estimate, when the model with an intercept is considered, that is, when the conditional probability that $Y = 1$, given $X = x$, is defined by

$$p_{\alpha,\theta}(x) = \frac{1}{1 + \mathrm{e}^{-\alpha - \langle \theta, x \rangle}} \quad \text{for } \alpha \in \mathbb{R} \text{ and } \theta, x \in E. \tag{5}$$

In this case, the assumption (FR) should be changed to

(FR') $\mathsf{P}(\langle \theta, X \rangle = \alpha) = 0$ for all $\theta \neq 0$ and $\alpha \in \mathbb{R}$.

We call $p_{\hat{\alpha},\hat{\theta}}$ the logistic estimate of (5), if

$$(\hat{\alpha}, \hat{\theta}) = \arg \min_{(\alpha,\theta) \in \mathbb{R} \times E_{k_n}} M_n(\alpha, \theta), \tag{6}$$

where

$$M_n(\alpha, \theta) = \overline{m_{\alpha,\theta}(X, Y)}, \quad m_{\alpha,\theta}(X, Y) = \log\left(1 + \mathrm{e}^{-Y(\alpha + \langle \theta, X \rangle)}\right).$$

We say that the logistic estimate is consistent, if $\mathsf{E}|p_{\hat{\alpha},\hat{\theta}}(X) - p_0(X)| \to 0$, as $n \to \infty$, where $p_0(x) = p_{\alpha_0,\theta_0}(x)$ in this case. As before, τ_k is defined by (4), where C is the covariance form of X. Our last result is the following Theorem.

Theorem 2. *Suppose assumptions (FR'), (M) and (UI) hold, and $\mathsf{E}X = 0$. Moreover, suppose that*

$$k_n \to \infty, \quad \frac{k_n}{n} \to 0 \quad \textit{and} \quad n\tau_{k_n}^4 \to \infty,$$

as $n \to \infty$. Then the logistic estimate is consistent.

3. Simulation study

To investigate the need of the conditions required for consistency, we performed a simulation study. We will give two examples: one, where all assumptions hold, and another one, where the assumption $n\tau_k^4 \to \infty$ does not hold.

Example 1. Since $X_i(t) = \sum_{j=1}^{\infty} C_{ij} e_j(t)$ for any selected basis system, it is enough to generate coefficients C_{ij}. To go in line with the (UI) assumption, we will generate C_{ij} as independent and normally distributed variables with zero mean and variances $\sigma_j^2 = 1/(1.1^j)$. Then $\tau_k = \sigma_k^2$. If we want that $n\tau_k^4 = n1.1^{-4k}$ tend to ∞, we have to take $k = \lceil c \log n \rceil$ with $c < 1/(4 \log 1.1) \approx 2.62$. In this example, we took $c = 2$, so that $n\tau_k^4 \to \infty$ and all assumptions would hold.

We took θ_0 with $\theta_{0i} = 1/(1.1^i)$ and calculated $p_{\theta_0}(X_i)$ up to the precision $\epsilon = 10^{-4}$. To this end we generated additional coordinates X_{ij} for $j \le l$, where l was the first index with $|\theta_{0l} X_{il}| < \epsilon$.

We generated $300, 500, 1000, 1500$ and 2000 observations, respectively, over 100 independent runs for each setting, and each time we approximated the distance

$$d(\hat{p}, p_0) = f(\hat{\theta}, \theta_0),$$

where

$$f(\theta, \theta_0) = \mathsf{E}|1/(1 + \mathrm{e}^{-U_1}) - 1/(1 + \mathrm{e}^{-U_2})|$$

with $U = (U1, U2)$ distributed according to the normal law with zero mean and covariance matrix

$$\Sigma = \begin{pmatrix} \sum_i \theta_i^2 \sigma_i^2 & \sum_i \theta_i \theta_{0i} \sigma_i^2 \\ \sum_i \theta_i \theta_{0i} \sigma_i^2 & \sum_i \theta_{0i}^2 \sigma_i^2 \end{pmatrix}.$$

We calculated f using the Monte Carlo method. We simulated 10000 independent copies of U, which gives, as preliminary testing shows, approximately 0.01 precision for d. We also reported the misclassification rate

$$\mathrm{MCR} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{\hat{y}_i \neq y_i\}},$$

where we set $\hat{y}_i = 1$, if $\hat{p}(x_i) >= 1/2$. Moreover, we reported the Bayes risk, where the probability of misclassification was calculated by

$$\mathsf{E} \min(p_0(X), 1 - p_0(X)) = \mathsf{E} \frac{1}{1 + \mathrm{e}^{|U|}}, \tag{7}$$

where $U \sim N(0, 1/(1.1^3 - 1))$. Again, we used Monte Carlo method to calculate (7). Figure 1 illustrates the simulated coefficients as well as the difference between the true and the estimated conditional probabilities. The x axis in plots (a)-(c) in Figure 1 represents the coefficient number j which stops after the kth value is generated. The y axis in plots (a)-(c) in Figure 1 represents the values of C_{ij}. As we can see from plots (a)-(c) the C_{ij} are distributed normally with mean 0 and their variance decreases as j increases. Plots (d)-(f) in Figure 1 shows the

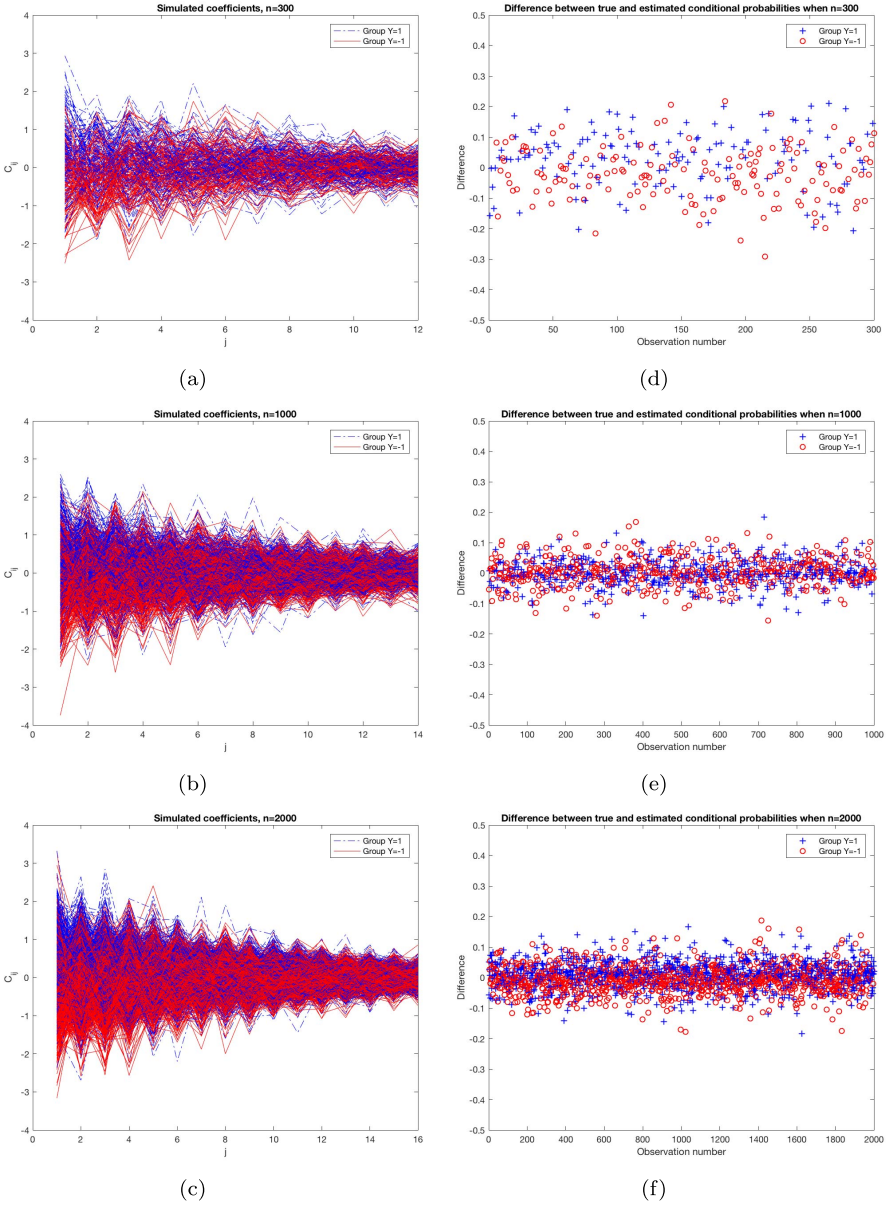

FIG 1. *Illustration of simulated data for Example 1. (a)-(c) Simulated coefficients C_{ij} for $n = 300, 1000$ and 2000, respectively. (d)-(f) Difference $(p_0 - \hat{p})$ between the true conditional probability p_0 and the estimated conditional probability $\hat{p}$, evaluated for the generated observations.*

difference $p_0 - \hat{p}$ between the true and the estimated conditional probabilities p_0 and $\hat{p}$, respectively, as functions of x. The x axis represents the observation

number i and the y-axis shows the value of $p_0 - \hat{p}$ at the $x = x_i, i = 1, \ldots, n$. We can see that the differences between the true and the estimated conditional probabilities are distributed more or less normally around zero and that the variance of them decreases as n increases suggesting that the average difference between the two probabilities tends to zero. This is further confirmed by $\hat{d}(\hat{p}, p_0)$ values in Table 1 which contains numerical results, averaged over 100 independent runs. As we can see from Table 1, the assumption $n\tau_k^4 \to \infty$ holds and $\hat{d}(\hat{p}, p_0) \to 0$, as expected.

TABLE 1
Numerical results for Example 1, averaged over 100 *independent runs*

n	300	500	1000	1500	2000
k	12	13	14	15	16
$n\tau_k^4$	3.1	3.5	4.8	4.9	4.9
$\hat{d}(\hat{p}, p_0)$ (± sd)	0.1 (± 0.017)	0.08 (±0.01)	0.06 (±0.01)	0.05 (±0.01)	0.05 (±0.01)
MCR (%, ± sd)	25.82 (± 2.8)	26.39 (±2.01)	26.44 (±1.24)	26.65 (±1.15)	26.65 (±0.98)
Bayes (%, ± sd)	24.34 (± 0.15)	24.34 (± 0.15)	24.34 (± 0.15)	24.34 (± 0.15)	24.34 (± 0.15)

Example 2. We considered the same settings as for Example 1, except that now we took $c = 6$, so that $n\tau_k^4 \to 0$ and even $n\tau_k^2 \to 0$. Figure 2 illustrates the simulated data as well as the difference between the true and the estimated conditional probabilities, while numerical results, averaged over 100 independent runs, are displayed in Table 2.

TABLE 2
Numerical results for Example 2, averaged over 100 *independent runs*

n	300	500	1000	1500	2000
k	35	38	42	44	46
$n\tau_k^2$	0.4	0.4	0.3	0.3	0.3
$n\tau_k^4$	$4 * 10^{-4}$	$2.5 * 10^{-4}$	$1.1 * 10^{-4}$	$7.8 * 10^{-5}$	$4.8 * 10^{-5}$
$\hat{d}(\hat{p}, p_0)$ (± sd)	0.26 (± 0.028)	0.24 (± 0.024)	0.2 (± 0.023)	0.19 (±0.022)	0.18 (±0.024)
MCR (%, ± sd)	22.92 (± 2.72)	23.97 (± 1.88)	25.4 (± 1.41)	25.6 (± 1.25)	25.95 (± 0.98)
Bayes (%, ± sd)	24.36 (± 0.14)	24.36 (± 0.14)	24.36 (± 0.14)	24.36 (± 0.14)	24.36 (± 0.14)

As we can see from Table 2, the assumption $n\tau_k^4 \to \infty$ (and even weaker assumption $n\tau_k^2 \to \infty$) is violated but $\hat{d}(\hat{p}, p_0) \to 0$, regardless. This suggests that the assumption $n\tau_k^4 \to \infty$ might be not needed to establish the consistency of logistic estimate and could be relaxed in future investigations.

4. Discussion

As we noted in the previous Section, the assumption $n\tau_{k_n}^4 \to \infty$ does not seem to be necessary for our main result to hold. It is interesting that the analogous assumption (M3) in Müller & Stadtmüller (2005) translates into $n\tau_{k_n}^2/k_n^2 \to \infty$. However, our simulation study shows (see Example 2) that even the assumption $n\tau_{k_n}^2 \to \infty$ is not necessary. At the moment either what the true asymptotic

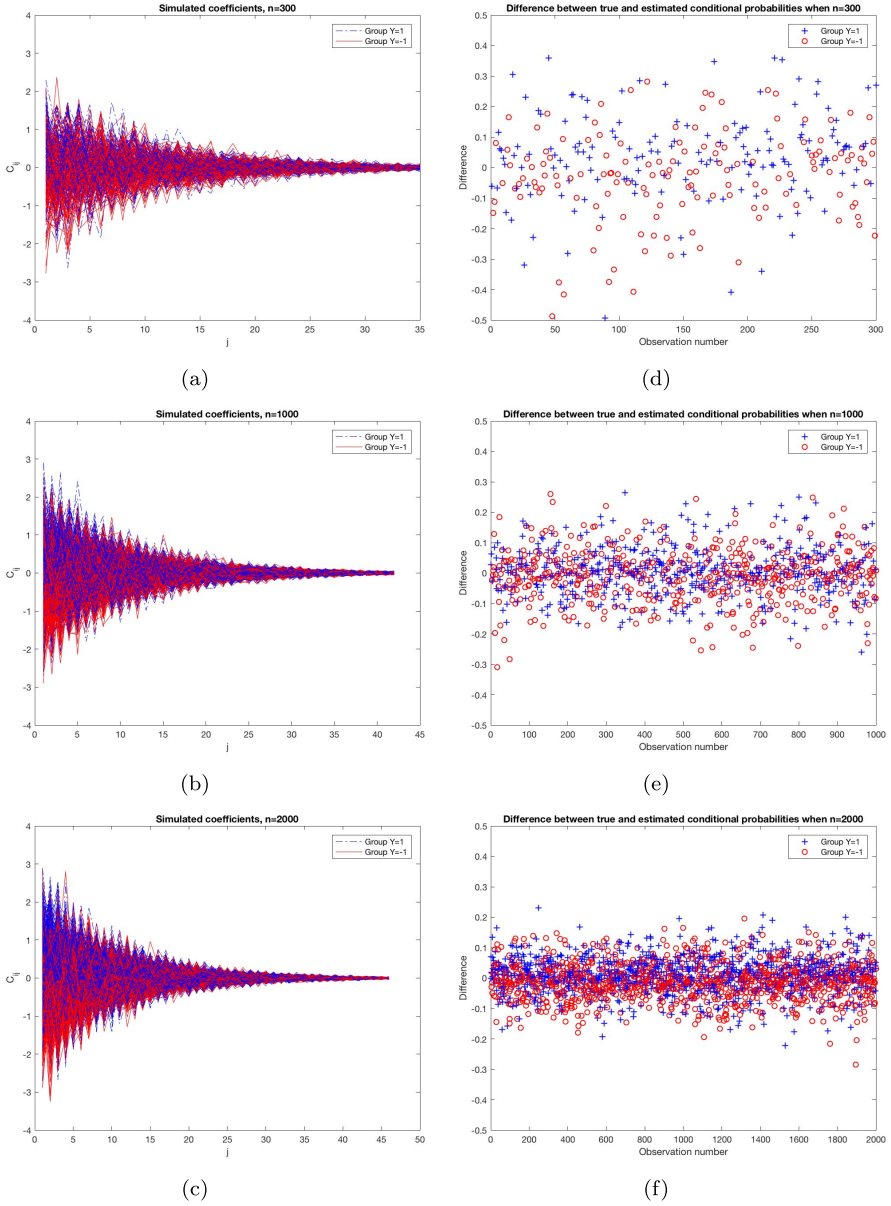

FIG 2. *Illustration of simulated data for Example 2. (a)-(c) Simulated coefficients C_{ij} for $n = 300, 1000$ and 2000, respectively. (d)-(f) Difference $(p_0 - \hat{p})$ between the true conditional probability p_0 and the estimated conditional probability $\hat{p}$, evaluated for the generated observations.*

lower bound for τ_{k_n} is, or how Theorem 1 can be proved under an assumption weaker than $n\tau_{k_n}^2 \to \infty$, is not clear.

5. Proofs

5.1. Facts from probability theory

Further in this Section, $\to_p$ and $\to_d$ denote convergence in probability and convergence in distribution, respectively, while $\to$ is used for the usual convergence in $\mathbb{R}$, or convergence in norm in E. For convenience of reference we recall some well-known facts about convergence and uniform integrability of random variables.

Proposition 1 (Continuous mapping theorem, see Kallenberg (2001), Theorem 3.7)**.** *Let U_n and U be random elements of some metric space S, $\mathsf{P}(U \in C) = 1$, T another metric space, and f_n, f measurable functions from S to T. If $u_n \to u \in C$ implies $f_n(u_n) \to f(u)$, then $U_n \to_d U$ implies $f_n(U_n) \to_d f(U)$.*

Proposition 2 (Subsequence criterion, see Kallenberg (2001), Lemma 3.2)**.** *Let U_n and U be random elements of some metric space S. Then $U_n \to_p U$ if and only if each subsequence of (U_n) has a further subsequence which converges in probability to U.*

Proposition 3 (see Kallenberg (2001), Lemma 3.10)**.** *If (Z_n) is a uniformly integrable sequence of random variables, then $\sup_n \mathsf{E}|Z_n| < \infty$ and $\mathsf{P}(W_n) \to 0$ implies $\mathsf{E}Z_n 1_{W_n} \to 0$.*

Proposition 4 (see Kallenberg (2001), Lemma 3.11)**.** *If (Z_n) is a uniformly integrable sequence of random variables, then $Z_n \to_d Z$ implies $\mathsf{E}Z_n \to \mathsf{E}Z$.*

Proposition 5 (Weak convergence version of Fatou's lemma, see Kallenberg (2001), Lemma 3.11)**.** *If (Z_n) is a sequence of positive random variables, then $Z_n \to_d Z$ implies $\varliminf_{n\to\infty} \mathsf{E}Z_n \geq \mathsf{E}Z$.*

5.2. The function $M(\theta)$

We begin by establishing some properties of the function $M(\theta)$. Recall that θ_0 denotes the "true" value of parameter θ.

Proposition 6. *1. If $\mathsf{E}\|X\| < \infty$, then, for all θ,*

$$0 < M(\theta_0) \leq M(\theta) < \infty.$$

2. If $\mathsf{E}\|X\| < \infty$, then $\theta_n \to \theta$ implies $M(\theta_n) \to M(\theta)$.
3. If $M(\theta_n) \to M(\theta_0)$, then $\langle \theta_n, X\rangle \to_p \langle \theta_0, X\rangle$.

Proof. 1. Inequality $M(\theta) > 0$ is implied by the fact that $m_\theta(x, y) > 0$ for all x and y. Because log function is increasing,

$$\begin{aligned} M(\theta) = \mathsf{E}\log(1 + \mathrm{e}^{-Y\langle\theta,X\rangle}) &\leq \mathsf{E}\log(1 + \mathrm{e}^{\|\theta\|\|X\|}) \\ &\leq \mathsf{E}\log(2\mathrm{e}^{\|\theta\|\|X\|}) = \log 2 + \|\theta\|\mathsf{E}\|X\| < \infty. \end{aligned}$$

Finally, convexity of the function $-\log$ yields

$$\begin{aligned} M(\theta) - M(\theta_0) &= -\mathsf{E} \log \frac{1+\mathrm{e}^{-Y\langle\theta_0,X\rangle}}{1+\mathrm{e}^{-Y\langle\theta,X\rangle}} \geq -\log \mathsf{E} \frac{1+\mathrm{e}^{-Y\langle\theta_0,X\rangle}}{1+\mathrm{e}^{-Y\langle\theta,X\rangle}} \\ &= -\log \mathsf{E} \left(\frac{1+\mathrm{e}^{\langle\theta_0,X\rangle}}{1+\mathrm{e}^{\langle\theta,X\rangle}} (1-p_{\theta_0}(X)) + \frac{1+\mathrm{e}^{-\langle\theta_0,X\rangle}}{1+\mathrm{e}^{-\langle\theta,X\rangle}} p_{\theta_0}(X) \right) \\ &= -\log \mathsf{E} \left(\frac{1}{1+\mathrm{e}^{\langle\theta,X\rangle}} + \frac{1}{1+\mathrm{e}^{-\langle\theta,X\rangle}} \right) = -\log 1 = 0. \end{aligned}$$

2. The statement follows from the dominated convergence theorem, because $\theta_n \to \theta$ implies that

$$m_{\theta_n}(X,Y) \to m_\theta(X,Y)$$

and

$$m_{\theta_n}(X,Y) \leq \log(1+\mathrm{e}^{\|\theta_n\|\|X\|}) \leq \log 2 + \|\theta_n\|\|X\| \leq \log 2 + c\|X\|$$

with $c = \sup_n \|\theta_n\| < \infty$.

3. Let $M(\theta_n) \to M(\theta_0)$. By Proposition 2, we have to prove that any subsequence $(\langle\theta_{n_k}, X\rangle)$ contains a further subsequence that tends in probability to $\langle\theta_0, X\rangle$. Note that $M(\theta_{n_k}) \to M(\theta_0)$, therefore, for ease of notation, we omit the index k.

The sequence of random vectors $(\langle\theta_n, X\rangle, \langle\theta_0, X\rangle)$ is tight in the space $\bar{\mathbb{R}} \times \mathbb{R}$. Indeed, if $K \subset \mathbb{R}$ is a compact interval such that $\mathsf{P}(\langle\theta_0, X\rangle \in K) \geq 1-\epsilon$ (and we can always find such K), then the set $\bar{\mathbb{R}} \times K$ is also compact and for all n

$$\mathsf{P}((\langle\theta_n, X\rangle, \langle\theta_0, X\rangle) \in \bar{\mathbb{R}} \times K) = \mathsf{P}(\langle\theta_0, X\rangle \in K) \geq 1-\epsilon.$$

By the Prokhorov's theorem (see Kallenberg (2001), Theorem 14.3), there exists a subsequence $(\langle\theta_{n_k}, X\rangle, \langle\theta_0, X\rangle)$, which converges in distribution in the space $\bar{\mathbb{R}} \times \mathbb{R}$ to some random vector (U_1, U_2).

By Proposition 5,

$$\begin{aligned} &\mathsf{E} \left(\frac{\log(1+\mathrm{e}^{U_1})}{1+\mathrm{e}^{U_2}} + \frac{\log(1+\mathrm{e}^{-U_1})}{1+\mathrm{e}^{-U_2}} \right) \\ &\leq \varliminf_{k\to\infty} \mathsf{E} \left(\frac{\log(1+\mathrm{e}^{\langle\theta_{n_k},X\rangle})}{1+\mathrm{e}^{\langle\theta_0,X\rangle}} + \frac{\log(1+\mathrm{e}^{-\langle\theta_{n_k},X\rangle})}{1+\mathrm{e}^{-\langle\theta_0,X\rangle}} \right) = \varliminf_{k\to\infty} M(\theta_{n_k}) = M(\theta_0). \end{aligned}$$

Obviously, U_2 is distributed identically to $\langle\theta_0, X\rangle$. Hence

$$\begin{aligned} M(\theta_0) &= \mathsf{E} \left(\frac{\log(1+\mathrm{e}^{\langle\theta_0,X\rangle})}{1+\mathrm{e}^{\langle\theta_0,X\rangle}} + \frac{\log(1+\mathrm{e}^{-\langle\theta_0,X\rangle})}{1+\mathrm{e}^{-\langle\theta_0,X\rangle}} \right) \\ &= \mathsf{E} \left(\frac{\log(1+\mathrm{e}^{U_2})}{1+\mathrm{e}^{U_2}} + \frac{\log(1+\mathrm{e}^{-U_2})}{1+\mathrm{e}^{-U_2}} \right) \end{aligned}$$

and therefore

$$\mathsf{E}\left(\frac{\log(1+\mathrm{e}^{U_1})}{1+\mathrm{e}^{U_2}}+\frac{\log(1+\mathrm{e}^{-U_1})}{1+\mathrm{e}^{-U_2}}\right)\le \mathsf{E}\left(\frac{\log(1+\mathrm{e}^{U_2})}{1+\mathrm{e}^{U_2}}+\frac{\log(1+\mathrm{e}^{-U_2})}{1+\mathrm{e}^{-U_2}}\right).$$

Let V be a random variable gaining values -1 and 1 with (conditional w.r.t. (U_1,U_2)) probabilities $\frac{1}{1+\mathrm{e}^{U_2}}$ and $\frac{1}{1+\mathrm{e}^{-U_2}}$. Then the above inequality can be re-written as

$$\mathsf{E}\log(1+\mathrm{e}^{-VU_1})\le \mathsf{E}\log(1+\mathrm{e}^{-VU_2}).$$

This yields

$$\begin{aligned}0\le \mathsf{E}\log\frac{1+\mathrm{e}^{-VU_2}}{1+\mathrm{e}^{-VU_1}}&\le \log \mathsf{E}\frac{1+\mathrm{e}^{-VU_2}}{1+\mathrm{e}^{-VU_1}}\\&=\log \mathsf{E}\left(\frac{1}{1+\mathrm{e}^{U_1}}+\frac{1}{1+\mathrm{e}^{-U_1}}\right)=\log 1=0.\end{aligned}$$

Therefore, both inequality signs can be replaced by equalities. However, Jensen's inequality becomes equality if and only if the variable that is being integrated almost surely is a constant. In this case that constant is 0, that is, almost surely

$$\log\frac{1+\mathrm{e}^{-VU_2}}{1+\mathrm{e}^{-VU_1}}=0$$

and $U_1=U_2$.

Hence $(\langle\theta_{n_k},X\rangle,\langle\theta_0,X\rangle)\to_d (U_2,U_2)$ and therefore $\langle\theta_{n_k},X\rangle-\langle\theta_0,X\rangle\to_d U_2-U_2=0$. When the limit random variable is 0 (or a constant), convergence in distribution is equivalent to convergence in probability (Kallenberg (2001), Lemma 3.7). Therefore, $\langle\theta_{n_k},X\rangle-\langle\theta_0,X\rangle\to_p 0$ and $\langle\theta_{n_k},X\rangle\to_p\langle\theta_0,X\rangle$. □

For any $f\in C^r(E_k)$ we assume that its rth derivative at the point $\theta\in E_k$ is a symmetric r-linear form on E_k defined by

$$f^{(r)}(\theta)(d\theta_1,\dots,d\theta_r)=D_{d\theta_r}\cdots D_{d\theta_1}f(\theta),$$

where $D_{d\theta}$ stands for the directional derivative along $d\theta\in E_k$. Its norm is defined by

$$\|f^{(r)}(\theta)\|=\sup_{\|d\theta_1\|\le 1,\dots,\|d\theta_r\|\le 1}|f^{(r)}(\theta)(d\theta_1,\dots,d\theta_r)|.$$

The function $d\theta\mapsto f^{(r)}(\theta)(d\theta,\dots,d\theta)$ is called the rth differential of f and is denoted by $d^rf(\theta)$. For example, $d^2f(\theta)$ is a quadratic form associated with the bilinear form $f''(\theta)$.

For any $x\in E$ and $y\in\{-1,1\}$, function $\theta\mapsto m_\theta(x,y)$ is infinitely differentiable on E_k and

$$m'_\theta(x,y)d\theta=\frac{\mathrm{e}^{-y\langle\theta,x\rangle}}{1+\mathrm{e}^{-y\langle\theta,x\rangle}}(-y\langle d\theta,x\rangle),$$

$$m''_\theta(x,y)(d\theta_1,d\theta_2) = \frac{\mathrm{e}^{-y\langle\theta,x\rangle}}{(1+\mathrm{e}^{-y\langle\theta,x\rangle})^2}\langle d\theta_1,x\rangle\langle d\theta_2,x\rangle,$$
$$m'''_\theta(x,y)(d\theta_1,d\theta_2,d\theta_3) = \frac{\mathrm{e}^{-y\langle\theta,x\rangle}-\mathrm{e}^{-2y\langle\theta,x\rangle}}{(1+\mathrm{e}^{-y\langle\theta,x\rangle})^3}\langle d\theta_1,x\rangle\langle d\theta_2,x\rangle(-y\langle d\theta_3,x\rangle).$$

It is obvious that

$$|m'_\theta(X,Y)d\theta| \le \|d\theta\|\|X\|,$$
$$|m''_\theta(X,Y)(d\theta_1,d\theta_2)| \le |\langle d\theta_1,X\rangle||\langle d\theta_2,X\rangle| \le \|d\theta_1\|\|d\theta_2\|\|X\|^2,$$
$$|m'''_\theta(X,Y)(d\theta_1,d\theta_2,d\theta_3)| \le \|d\theta_1\|\|d\theta_2\|\|d\theta_3\|\|X\|^3.$$

Therefore,

$$\|m'_\theta(X,Y)\| \le \|X\|, \quad \|m''_\theta(X,Y)\| \le \|X\|^2, \quad \|m'''_\theta(X,Y)\| \le \|X\|^3,$$

moreover, $\|X\|, \|X\|^2, \|X\|^3$ are integrable, if $\mathsf{E}\|X\|^3 < \infty$. Hence $M(\theta)$, as a function on E_k, belongs to $C^3(E_k)$, and

$$\mathrm{d}M(\theta) = -\mathsf{E}\frac{\mathrm{e}^{-Y\langle\theta,X\rangle}}{1+\mathrm{e}^{-Y\langle\theta,X\rangle}}Y\langle d\theta,X\rangle,$$
$$\mathrm{d}^2M(\theta) = \mathsf{E}\frac{\mathrm{e}^{-Y\langle\theta,X\rangle}}{(1+\mathrm{e}^{-Y\langle\theta,X\rangle})^2}\langle d\theta,X\rangle^2,$$
$$\mathrm{d}^3M(\theta) = -\mathsf{E}\frac{\mathrm{e}^{-Y\langle\theta,X\rangle}-\mathrm{e}^{-2Y\langle\theta,X\rangle}}{(1+\mathrm{e}^{-Y\langle\theta,X\rangle})^3}Y\langle d\theta,X\rangle^3.$$

If the distribution of X is of full rank, then, for any $d\theta \neq 0$, almost surely $\langle d\theta,X\rangle^2 > 0$ and therefore $\mathrm{d}^2M(\theta) > 0$. Hence, for all θ, $\mathrm{d}^2M(\theta)$ is a positive definite quadratic form. According to Bertsekas et. al. (2003), $M(\theta)$ is strictly convex on E_k.

Proposition 7. *If assumptions (FR) and (M) hold, then, for any $k \ge 1$, the function $M(\theta)$ has a unique minimum point in the space E_k. Furthermore, if θ_k is that point, then $M(\theta_k) \to M(\theta_0)$, as $k \to \infty$.*

Proof. Step 1: we will prove that sets $A_q = \{\theta \in E_k \mid M(\theta) \le q\}$ are bounded.

Suppose the contrary. Then there exists some set A_q that is not bounded. Find a sequence $(\theta_m) \subset E_k$ such that $M(\theta_m) \le q$ for all m, and $\|\theta_m\| \to \infty, \theta_m/\|\theta_m\| \to a$, as $m \to \infty$. Because $\|a\| = 1$ and the distribution of X is of full rank, either $\langle a,X\rangle < 0$ or $\langle a,X\rangle > 0$ with a positive probability. Since $0 < p_{\theta_0} < 1$,

$$0 < \mathsf{P}(Y\langle a,X\rangle < 0) \le \mathsf{P}(\lim_{m\to\infty} m_{\theta_m}(X,Y) = \infty)$$

and so $\mathsf{E}\ \underline{\lim}_{m\to\infty}\ m_{\theta_m}(X,Y) = \infty$. On the other hand, by Fatou's lemma,

$$\mathsf{E}\ \underline{\lim}_{m\to\infty}\ m_{\theta_m}(X,Y) \le \underline{\lim}_{m\to\infty}\ M(\theta_m) \le q.$$

A contradiction.

Step 2: the end of the proof.

The existence of θ_k follows from Proposition 2.1.1 of Bertsekas et. al. (2003). Since $M(\theta)$ is strictly convex, the minimum point is unique.

If $\theta_0^{(k)}$ is the projection of θ_0 in the space E_k, then $M(\theta_0) \leq M(\theta_k) \leq M(\theta_0^{(k)})$. From $\theta_0^{(k)} \to \theta_0$ we get that $M(\theta_0^{(k)}) \to M(\theta_0)$. Therefore, also $M(\theta_k) \to M(\theta_0)$. □

We are now ready to establish the consistency criterion. The following Proposition provides the consistency conditions for the estimate of the type $\hat{p} = p_{\hat{\theta}_n}$, where $\hat{\theta}_n$ is any estimate of θ. If $\hat{\theta}_n$ is defined by (1)-(2), we get the consistency criterion for the logistic estimate.

Proposition 8. *1. If $M(\hat{\theta}_n) \to_p M(\theta_0)$, then the estimate $p_{\hat{\theta}_n}$ is consistent.*

2. Suppose assumptions (FR) and (M) hold, and θ_k is the minimum of the function M in the space E_k. If $k_n \to \infty$ and $M(\hat{\theta}_n) - M(\theta_{k_n}) \to_p 0$, then the estimate $p_{\hat{\theta}_n}$ is consistent.

Proof. 1. By Proposition 6, $M(\theta_n) \to M(\theta_0)$ implies $\langle \theta_n, X\rangle \to_p \langle \theta_0, X\rangle$. Then $p_{\theta_n}(X) \to_p p_{\theta_0}(X)$ and, by Proposition 4, $\mathsf{E}|p_{\theta_n}(X) - p_{\theta_0}(X)| \to 0$.

Let now $M(\hat{\theta}_n) \to_p M(\theta_0)$. We have to prove that $\mathsf{E}|p_{\hat{\theta}_n}(X) - p_{\theta_0}(X)| \to 0$. It is enough to prove that any subsequence $\mathsf{E}|p_{\hat{\theta}_{n_s}}(X) - p_{\theta_0}(X)|$ has a further subsequence that tends to 0. Moreover, it is well-known that any sequence that converges in probability has a subsequence that converges almost everywhere. Therefore, it is enough to prove that, if almost surely $M(\hat{\theta}_{n_s}) \to M(\theta_0)$, then $\mathsf{E}|p_{\hat{\theta}_{n_s}}(X) - p_{\theta_0}(X)| \to 0$.

However, if almost surely $M(\hat{\theta}_{n_s}) \to M(\theta_0)$, then from the first paragraph of this proof we get that almost surely

$$\mathsf{E}^*|p_{\hat{\theta}_{n_s}}(X) - p_{\theta_0}(X)| \to 0,$$

where E^* denotes the conditional mean w.r.t. sequence $((X_i, Y_i) \mid i \geq 1)$. It is enough to use the dominated convergence theorem.

2. The second statement follows from the first one and from Proposition 7. □

5.3. The function $M_n(\theta)$

Now suppose that k and n are fixed and consider $M_n(\theta)$, as a function on E_k. For all θ, $d\theta \in E_k, x \in E$ and $y \in \{-1, 1\}$,

$$m''_\theta(x, y)(d\theta, d\theta) = \frac{\mathrm{e}^{-y\langle\theta,x\rangle}}{(1 + \mathrm{e}^{-y\langle\theta,x\rangle})^2}\langle d\theta, x\rangle^2 \geq 0.$$

Therefore, the function $\theta \mapsto m_\theta(x, y)$ is convex in E_k. Then also the function $M_n(\theta)$ is convex. We first give conditions for its strict convexity.

Note that if $\theta \in E_k$, then $\langle \theta, X_i\rangle = \langle \theta, X_i^{(k)}\rangle$, where $X_i^{(k)}$ denotes the projection of vector X_i in the space E_k.

Proposition 9. *If $n \geq k$ and $X_1^{(k)}, \dots, X_k^{(k)}$ are linearly independent, then function $M_n(\theta)$ is strictly convex on E_k. If assumption (FR) holds, the probability of such event is* 1.

Proof. The function $M_n(\theta)$ is strictly convex if its second differential $\mathrm{d}^2 M_n(\theta)$ is a positive definite quadratic form. Since

$$\mathrm{d}^2 M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{e}^{-Y_i \langle \theta, X_i \rangle}}{(1 + \mathrm{e}^{-Y_i \langle \theta, X_i \rangle})^2} \langle d\theta, X_i^{(k)} \rangle^2,$$

and all summands in the right-hand side are nonnegative, $\mathrm{d}^2 M_n(\theta) = 0$ implies that $d\theta$ is perpendicular to all $X_i^{(k)}$. If $n \geq k$ and $X_1^{(k)}, \dots, X_k^{(k)}$ are linearly independent, then $d\theta = 0$.

The second statement follows from Theorem 1 in Kazakeviciute & Olivo (2017). □

Recall some notions from Kazakeviciute & Olivo (2017). Let $(x_1, y_1), \dots, (x_n, y_n)$ be n vectors from $E_k \times \{-1, 1\}$, called *sample points*, and $a \neq 0$ be another vector from E_k. We say that the vector a *separates sample points* if, for all i,

$$y_i \langle a, x_i \rangle \geq 0.$$

We say that sample points are *separable*, if there exists some $a \neq 0$ that separates them. Note that this definition is equivalent to the definition of quasi-complete separation, given by Albert & Anderson (1984). Next, the statement "the sample $(X_1, Y_1), \dots, (X_n, Y_n)$ is k-separable" defines some event, the set of all elementary events ω such that sample points

$$(X_1^{(k)}(\omega), Y_1(\omega)), \dots, (X_n^{(k)}(\omega), Y_n(\omega)) \tag{8}$$

are separable.

Proposition 10. *If the sample $(X_1, Y_1), \dots, (X_n, Y_n)$ is not k-separable then, for any $q > 0$, the (random) set $A_q = \{\theta \in E_k \mid M_n(\theta) \leq q\}$ is bounded.*

Proof. Fix any ω such that the set $A_q(\omega)$ is not bounded and denote $x_i = X_i^{(k)}(\omega)$, $y_i = Y_i(\omega)$. Find a sequence $(\theta_m) \subset A_q$ such that $\|\theta_m\| \to \infty$ and $\theta_m / \|\theta_m\| \to a$. Then, for all m and all $i = 1, \dots, n$,

$$\log(1 + \mathrm{e}^{-y_i \langle \theta_m, x_i \rangle}) \leq \sum_{i=1}^{n} \log(1 + \mathrm{e}^{-y_i \langle \theta_m, x_i \rangle}) \leq nq.$$

But

$$-y_i \langle \theta_m, x_i \rangle = -\|\theta_m\| y_i \Big\langle \frac{\theta_m}{\|\theta_m\|}, x_i \Big\rangle \to \infty$$

if $y_i \langle a, x_i \rangle < 0$. Hence $y_i \langle a, x_i \rangle \geq 0$ for all i, that is, a separates sample points (8). □

Now suppose $n \geq k$ and let W_{kn} denote the following event: $X_1^{(k)}, \ldots, X_k^{(k)}$ vectors are linearly independent and the sample is not k-separable. If $\omega \in W_{kn}$ then, by Propositions 9 and 10, the function $M_n(\theta)$ is strictly convex and all its sub-level sets A_q are bounded. As is seen from the proof of Proposition 7, then $M_n(\theta)$ has the unique minimum point, which is, of course $\hat{\theta}_{kn}(\omega)$. If $\omega \notin W_{kn}$, we suppose that $\hat{\theta}_{kn}(\omega) = 0$.

Denote $q_{kn} = \mathsf{P}(W_{kn}^c)$. Then, by Proposition 9 and by Corollary 2.1 in Kazakeviciute & Olivo (2017), $q_{k_n n} \to 0$, provided that assumption (FR) holds and $k_n/n \to 0$.

5.4. *Proof of Theorem 1*

We follow the proof of Theorem 5.42 from van der Vaart (2000).

For $k \geq 1$ and $\theta \in E_k, x \in E, y \in \{-1, 1\}$ let us define

$$\psi_{k,\theta}(x, y) = -\frac{\mathrm{e}^{-y\langle \theta, x\rangle}}{1 + \mathrm{e}^{-y\langle \theta, x\rangle}} y x^{(k)},$$

where $x^{(k)}$ denotes the orthogonal projection of x in the space E_k. It is obvious that the function $\theta \mapsto \psi_{k,\theta}(x, y)$ is the gradient of the restriction of the function $m_\theta(x, y)$ on E_k. Also let us define

$$\Psi_{k,n}(\theta) = \overline{\psi_{k,\theta}(X, Y)}, \quad \text{and} \quad \Psi_k(\theta) = \mathsf{E}\psi_{k,\theta}(X, Y).$$

These functions are the gradients of the functions $M_n(\theta)$ and $M(\theta)$, as functions on E_k, respectively. Therefore, both $\Psi_{k,n}$ and Ψ_k are C^2-smooth functions from E_k to E_k. The derivative $\Psi_k'(\theta)$ is the linear operator from E_k to E_k which maps $d\theta_1 \in E_k$ to a vector $\Psi_k'(\theta)d\theta_1 \in E_k$ such that, for all $d\theta_2 \in E_k$,

$$\langle \Psi_k'(\theta)d\theta_1, d\theta_2\rangle = M''(\theta)(d\theta_1, d\theta_2).$$

Proposition 11. *The function Ψ_k is a diffeomorphism.*

Proof. Suppose $\Psi_k(\theta_1) = \Psi_k(\theta_2)$ and denote $d\theta = \theta_2 - \theta_1$. Then, for some $t \in (0, 1)$,

$$0 = \langle \Psi_k(\theta_2), d\theta\rangle - \langle \Psi_k(\theta_1), d\theta\rangle = M''(\theta_1 + td\theta)(d\theta, d\theta).$$

This yields $d\theta = 0$, that is, $\theta_1 = \theta_2$. Therefore, the function Ψ_k is injective.

Analogously, from $\Psi_k'(\theta)d\theta = 0$ we get that

$$0 = \langle \Psi_k'(\theta)d\theta, d\theta\rangle = M''(\theta)(d\theta, d\theta)$$

and $d\theta = 0$. Therefore, the operator $\Psi_k'(\theta)$ is invertible for all θ.

The statement of the theorem now follows from the inverse function theorem. □

Proposition 11 implies that the set $V = \Psi_k(E_k)$ is open. Moreover, $0 \in V$ because $\Psi_k(\theta_k) = 0$. Let us take some δ_k such that $\bar{U}(0, \delta_k) \subset V$ and denote $U_k = \Psi_k^{-1}(U(0, \delta_k))$. Then U_k is the neighborhood of the point θ_k. Moreover, because Ψ_k is a homeomorphism between E_k and V,

$$\Psi_k(\bar{U}_k) = \overline{\Psi_k(U_k)} = \overline{U(0, \delta_k)} = \bar{U}(0, \delta_k).$$

Denote

$$W_{kn}^{'} = \{\sup_{\theta \in \bar{U}_k} \|\Psi_{k,n}(\theta) - \Psi_k(\theta)\| \leq \delta_k\}.$$

The following reasoning is under the assumption that event $W_{kn} \cap W_{kn}'$ occurred.

If $z \in \bar{U}(0, \delta_k)$, then $\Psi_k^{-1}(z) \in \bar{U}_k$ and then

$$\|z - \Psi_{k,n}(\Psi_k^{-1}(z))\| = \|\Psi_k(\Psi_k^{-1}(z)) - \Psi_{k,n}(\Psi_k^{-1}(z))\| \leq \delta_k.$$

Therefore $z \mapsto z - \Psi_{k,n}(\Psi_k^{-1}(z))$ is a continuous function from $\bar{U}(0, \delta_k)$ to $\bar{U}(0, \delta_k)$. From the Brouwer's Fixed Point Theorem we get that, for some $z \in \bar{U}(0, \delta_k)$,

$$z = z - \Psi_{k,n}(\Psi_k^{-1}(z)),$$

that is, $\Psi_{k,n}(\Psi_k^{-1}(z)) = 0$. Because the function $M_n(\theta)$ is strictly convex, $\hat{\theta}_{kn}$ is the unique zero of the function $\Psi_{k,n}$. Therefore, $\hat{\theta}_{kn} = \Psi_k^{-1}(z) \in \bar{U}_k$.

Let $d_k = \text{diam}\bar{U}_k$. Then $\|\hat{\theta}_{kn} - \theta_k\| \leq d_k$ and

$$|M(\hat{\theta}_{kn}) - M(\theta_k)| \leq \sup_{\theta} \|\Psi_k(\theta)\| d_k \leq \mathsf{E}\|X\| d_k.$$

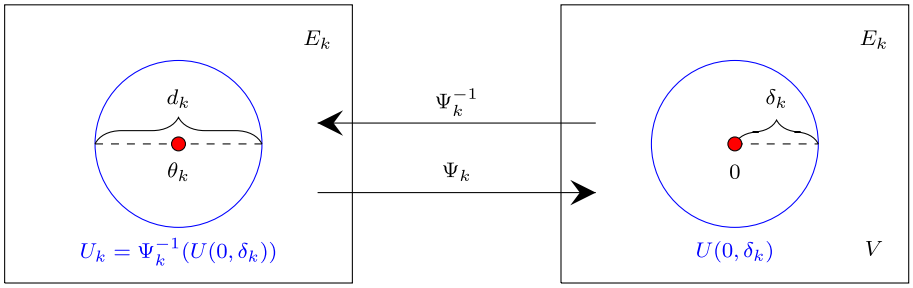

FIG 3. *Conceptual illustration of the ideas from Theorem 5.42 in van der Vaart (2000) that solves the well-known problem in statistics: by Law of Large Numbers, the empirical expectation tends to the true expectation. How to prove that $\hat{\theta}_{kn}$ that minimizes the empirical expectation tends to θ_k that minimizes the true expectation? As van der Vaart suggests, if the distance between the gradients of the empirical and the true expectations are bounded by δ_k, then the distance between $\hat{\theta}_{kn}$ and θ_k is bounded by d_k.*

Therefore, in order to prove Theorem 1 it is enough to choose δ_k in such a way that $d_{k_n} \to 0$ and $\mathsf{P}(W_{k_n,n}^{\prime c}) \to 0$.

We now need to evaluate the diameter d_k. The following Proposition gives the necessary result.

Proposition 12. *Suppose assumptions (FR), (M) and (UI) are satisfied and* $\delta_k = o(\sqrt{\tau_k})$, *as* $k \to \infty$. *Then* $d_k = O(\delta_k/\tau_k)$.

The proof of Proposition 12 is preceded with three lemmas.

Lemma 1. *Let* (Z_n) *be a sequence of positive integrable variables such that the sequence* $(Z_n/\mathsf{E}Z_n)$ *is uniformly integrable. Then, for all* $q < 1$,

$$\varliminf_{n\to\infty} \mathsf{P}(Z_n \geq q\mathsf{E}Z_n) > 0.$$

Proof. Suppose the contrary. Without loss of generality, we can assume that

$$\mathsf{P}(Z_n \geq q\mathsf{E}Z_n) \to 0.$$

From uniform integrability we get that

$$\mathsf{E}\frac{Z_n}{\mathsf{E}Z_n}1_{\{Z_n \geq q\mathsf{E}Z_n\}} \to 0.$$

Therefore, there exists n such that

$$\mathsf{E}Z_n 1_{\{Z_n \geq q\mathsf{E}Z_n\}} < (1-q)\mathsf{E}Z_n.$$

But then

$$\mathsf{E}Z_n = \mathsf{E}Z_n 1_{\{Z_n \geq q\mathsf{E}Z_n\}} + \mathsf{E}Z_n 1_{\{Z_n < q\mathsf{E}Z_n\}} < (1-q)\mathsf{E}Z_n + q\mathsf{E}Z_n = \mathsf{E}Z_n.$$

A contradiction. □

Lemma 2. *Suppose the assumptions (FR), (M) and (UI) hold and* $\delta_k = o(\sqrt{\tau_k})$, *as* $k \to \infty$. *Then there exists* k_0 *such that, for all* $k \geq k_0$ *and all* $d\theta \in E_k$ *with* $\|d\theta\| = 1$,

$$\exists t > 0 \ \langle \Psi_k(\theta_k + td\theta), d\theta\rangle > \delta_k. \tag{9}$$

Proof. Step 1: we prove that if (9) fails, for some $k \geq 1$ and $d\theta \in E_k$ with $\|d\theta\| = 1$, then

$$\mathsf{E}(Y\langle d\theta, X\rangle)^- \leq \delta_k. \tag{10}$$

If (9) fails then, for some $t_m \to \infty$,

$$\delta_k \geq \langle \Psi_k(\theta_k + t_m d\theta), d\theta\rangle = -\mathsf{E}\frac{\mathrm{e}^{-Y\langle\theta_k,X\rangle - t_m Y\langle d\theta,X\rangle}}{1+\mathrm{e}^{-Y\langle\theta_k,X\rangle - t_m Y\langle d\theta,X\rangle}} Y\langle d\theta, X\rangle.$$

Note that

$$\frac{\mathrm{e}^{-Y\langle\theta_k,X\rangle - t_m Y\langle d\theta,X\rangle}}{1+\mathrm{e}^{-Y\langle\theta_k,X\rangle - t_m Y\langle d\theta,X\rangle}} \xrightarrow[m\to\infty]{} \begin{cases} 0, & \text{if } Y\langle d\theta, X\rangle > 0, \\ 1, & \text{if } Y\langle d\theta, X\rangle < 0, \end{cases}$$

Therefore (10) follows by dominated convergence.

Step 2: the end of the proof.

Suppose $\delta_k = o(\sqrt{\tau_k})$, as $k \to \infty$, but the assertion of the Lemma is false. Then there exists a sequence $k_m \to \infty$ and a sequence $(d\theta_m)$ such that, for all $m \geq 1$, $d\theta_m \in E_{k_m}$, $\|d\theta_m\| = 1$ and, by the result of Step 1, $\mathsf{E}(Y\langle d\theta_m, X\rangle)^- \leq \delta_{k_m}$. Hence

$$\frac{\mathsf{E}(Y\langle d\theta_m, X\rangle)^-}{\sqrt{\tau_{k_m}}} \xrightarrow[m\to\infty]{} 0.$$

Then also

$$\frac{\mathsf{E}(Y\langle d\theta_m, X\rangle)^-}{\sqrt{C(d\theta_m, d\theta_m)}} \xrightarrow[m\to\infty]{} 0.$$

But

$$\begin{aligned}
\mathsf{E}(Y\langle d\theta_m, X\rangle)^- &= -\mathsf{E}\langle d\theta_m, X\rangle 1_{\{\langle d\theta_m, X\rangle<0, Y=1\}} + \mathsf{E}\langle d\theta_m, X\rangle 1_{\{\langle d\theta_m, X\rangle>0, Y=-1\}} \\
&= \mathsf{E}|\langle d\theta_m, X\rangle| \left(\frac{1_{\{\langle d\theta_m, X\rangle<0\}}}{1+\mathrm{e}^{-\langle\theta_0, X\rangle}} + \frac{1_{\{\langle d\theta_m, X\rangle>0\}}}{1+\mathrm{e}^{\langle\theta_0, X\rangle}} \right) \\
&\geq \mathsf{E}\frac{|\langle d\theta_m, X\rangle|}{1+\mathrm{e}^{|\langle\theta_0, X\rangle|}} \\
&\geq \frac{\sqrt{C(d\theta_m, d\theta_m)}}{2} \mathsf{E}\frac{1_{\{|\langle d\theta_m, X\rangle| \geq \sqrt{C(d\theta_m, d\theta_m)}/2\}}}{1+\mathrm{e}^{|\langle\theta_0, X\rangle|}},
\end{aligned}$$

therefore

$$\mathsf{E}\frac{1_{\{|\langle d\theta_m, X\rangle| \geq \sqrt{C(d\theta_m, d\theta_m)}/2\}}}{1+\mathrm{e}^{|\langle\theta_0, X\rangle|}} \to 0.$$

This yields

$$\frac{1_{\{|\langle d\theta_m, X\rangle| \geq \sqrt{C(d\theta_m, d\theta_m)}/2\}}}{1+\mathrm{e}^{|\langle\theta_0, X\rangle|}} \to_p 0$$

and then

$$1_{\{|\langle d\theta_m, X\rangle| \geq \sqrt{C(d\theta_m, d\theta_m)}/2\}} \to_p 0,$$

that is,

$$\mathsf{P}(\langle d\theta_m, X\rangle^2 \geq C(d\theta_m, d\theta_m)/4) \to 0.$$

This contradicts Lemma 1. □

If Z is a positive random variable and $\mathsf{E}Z = 1$, we can consider Z as a density, that is, with any random vector U there exists a random vector $\tilde{U}$ such that with any nonnegative or any bounded Borel function f

$$\mathsf{E}f(\tilde{U}) = \mathsf{E}f(U)Z.$$

We need the following property of the transformation $U \mapsto \tilde{U}$.

Lemma 3. *Let (Z_n) be a sequence of positive random variables, $\mathsf{E}Z_n = 1$ for all n, (U_n) be another sequence of random variables and let $\tilde{U}_n$ be a random variable such that with any nonnegative or any bounded Borel function f*

$$\mathsf{E}f(\tilde{U}_n) = \mathsf{E}f(U_n)Z_n.$$

If the sequence (Z_n) is uniformly integrable, then $U_n = O_p(1)$ implies $\tilde{U}_n = O_p(1)$.

Proof. Fix ϵ and find c_1 such that

$$\sup_n \mathsf{E} Z_n 1_{\{Z_n > c_1\}} < \epsilon.$$

Then find c such that

$$\sup_n \mathsf{P}(|U_n| > c) < \epsilon/c_1.$$

Then for all n,

$$\begin{aligned}\mathsf{P}(|\tilde{U}_n| > c) &= \mathsf{E}1_{\{|\tilde{U}_n|>c\}} = \mathsf{E}1_{\{|U_n|>c\}} Z_n \\ &= \mathsf{E}1_{\{|U_n|>c, Z_n \le c_1\}} Z_n + \mathsf{E}1_{\{|U_n|>c, Z_n > c_1\}} Z_n \\ &\le c_1 \mathsf{P}(|U_n| > c) + \mathsf{E}1_{\{Z_n > c_1\}} Z_n < 2\epsilon.\end{aligned}$$

Therefore, $\tilde{U}_n = O_p(1)$. □

Now we are ready to prove Proposition 12.

Proof. Lemma 2 implies that if k is large enough then, for any $d\theta \in E_k$ with $\|d\theta\| = 1$, at least one of the values of the function $f(t) = \langle \Psi_k(\theta_k + td\theta), d\theta\rangle$ is greater than δ_k. The function is continuous, strictly increasing and equal to 0, when $t = 0$. Therefore, there exists unique $t = t_k(d\theta) > 0$ such that $\langle \Psi_k(\theta_k + td\theta), d\theta\rangle = \delta_k$.

Step 1: we will prove that $d_k \le 2\alpha_k$, where

$$\alpha_k = \sup_{\substack{d\theta \in E_k \\ \|d\theta\|=1}} t_k(d\theta).$$

It is enough to prove that $\Psi_k^{-1}(\bar{U}(0,\delta_k)) \subset \bar{U}(\theta_k, \alpha_k)$. Let $\theta \in \Psi_k^{-1}(\bar{U}(0,\delta_k))$, that is $\|\Psi_k(\theta)\| \le \delta_k$. Denote $d\theta = (\theta - \theta_k)/\|\theta - \theta_k\|$. Then

$$\langle \Psi_k(\theta_k + \|\theta - \theta_k\| d\theta), d\theta\rangle = \langle \Psi_k(\theta), d\theta\rangle \le \|\Psi_k(\theta)\| \|d\theta\| \le \delta_k.$$

Therefore, $\|\theta - \theta_k\| \le t_k(d\theta) \le \alpha_k$.

Step 2: transforming the task to a simpler one.

From the result in Step 1 we get that it is enough to prove that $\alpha_k = O(\delta_k/\tau_k)$, that is that $\alpha_k \tau_k/\delta_k = O(1)$. Suppose the contrary, that there exists some subsequence that is unbounded. Then, without loss of generality, we can assume

$$\alpha_k \tau_k/\delta_k \to \infty$$

and we need to get a contradiction.

Let $d\theta_k$ be unit-length vectors from E_k such that $t_k(d\theta_k)/\alpha_k \to 1$. Then

$$\tau_k t_k(d\theta_k)/\delta_k \to \infty$$

and so

$$C(d\theta_k, d\theta_k) t_k(d\theta_k)/\delta_k \to \infty. \tag{11}$$

For short, denote

$$t_k = t_k(d\theta_k), \quad u_k = t_k\sqrt{C(d\theta_k, d\theta_k)}, \quad \beta_k = \frac{\delta_k}{\sqrt{C(d\theta_k, d\theta_k)}}$$

and

$$Z_{1k} = \langle \theta_k, X\rangle, \quad Z_{2k} = \frac{\langle d\theta_k, X\rangle}{\sqrt{C(d\theta_k, d\theta_k)}}.$$

It is obvious that $\beta_k \le \delta_k/\sqrt{\tau_k} \to 0$ and from (11) we get that $u_k/\beta_k \to \infty$. Moreover,

$$\delta_k = \langle \Psi_k(\theta_k + t_k d\theta_k), d\theta_k\rangle = f_k(1) - f_k(0) = \int_0^1 f_k'(t)\mathrm{d}t,$$

where

$$f_k(t) = \langle \Psi_k(\theta_k + tt_k d\theta_k), d\theta_k\rangle$$

and

$$\begin{aligned} f_k'(t) &= t_k M^{''}(\theta_k + tt_k d\theta_k)(d\theta_k, d\theta_k) = t_k \mathsf{E}\frac{\mathrm{e}^{-Y\langle \theta_k + tt_k d\theta_k, X\rangle}}{(1+\mathrm{e}^{-Y\langle \theta_k + tt_k d\theta_k, X\rangle})^2}\langle d\theta_k, X\rangle^2 \\ &= t_k C(d\theta_k, d\theta_k)\mathsf{E}\frac{\mathrm{e}^{-Y(Z_{1k}+tu_kZ_{2k})}}{(1+\mathrm{e}^{-Y(Z_{1k}+tu_kZ_{2k})})^2}Z_{2k}^2. \end{aligned}$$

Therefore,

$$\beta_k \to 0, \quad \beta_k/u_k \to 0, \quad \beta_k = u_k\mathsf{E}\int_0^1 \frac{\mathrm{e}^{-Y(Z_{1k}+tu_kZ_{2k})}}{(1+\mathrm{e}^{-Y(Z_{1k}+tu_kZ_{2k})})^2}\mathrm{d}tZ_{2k}^2$$

and we have to obtain a contradiction.

Step 3: selecting one more subsequence.

Since $\mathsf{E}Z_{2k}^2 = 1$, we can consider Z_{2k}^2 as a density. Then there exist random variables $\tilde{Y}_k, \tilde{Z}_{1k}$ and $\tilde{Z}_{2k}$ such that with any Borel function f

$$\mathsf{E}f(\tilde{Y}_k, \tilde{Z}_{1k}, \tilde{Z}_{2k}) = \mathsf{E}f(Y, Z_{1k}, Z_{2k})Z_{2k}^2.$$

As a separate case,

$$\mathsf{P}(|\tilde{Y}_k| = 1) = \mathsf{E}1_{\{|\tilde{Y}_k|=1\}} = \mathsf{E}1_{\{|Y|=1\}}Z_k^2 = \mathsf{E}Z_k^2 = 1,$$

that is, almost surely $\tilde{Y}_k \in \{-1, 1\}$. Moreover,

$$\beta_k = u_k\mathsf{E}\int_0^1 \frac{\mathrm{e}^{-\tilde{Y}_k(\tilde{Z}_{1k}+tu_k\tilde{Z}_{2k})}}{(1+\mathrm{e}^{-\tilde{Y}_k(\tilde{Z}_{1k}+tu_k\tilde{Z}_{2k})})^2}\mathrm{d}t.$$

Since $Z_{1k} = \langle \theta_k, X\rangle \to_p \langle \theta_0, X\rangle$, we get $Z_{1k} = O_p(1)$. Since the sequence (Z_{2k}^2) is uniformly integrable, $Z_{2k}^2 = O_p(1)$ and then also $Z_{2k} = O_p(1)$. Then from Lemma 3 we get that $\tilde{Y}_k = O_p(1), \tilde{Z}_{1k} = O_p(1)$ and $\tilde{Z}_{2k} = O_p(1)$. This

means that also $(\tilde{Y}_k, \tilde{Z}_{1k}, \tilde{Z}_{2k}) = O_p(1)$. From Prochorov's theorem we get that some subsequence of that sequence converges in distribution. Therefore we can suppose that $u_k \to u$ (where u can be infinite), and $(\tilde{Y}_k, \tilde{Z}_{1k}, \tilde{Z}_{2k}) \to_d (\tilde{Y}, \tilde{Z}_1, \tilde{Z}_2)$.

Step 4: the case, where $u_k \to u < \infty$.

Denote

$$g_u(y, z_1, z_2) = \int_0^1 \frac{\mathrm{e}^{-y(z_1+tuz_2)}}{(1+\mathrm{e}^{-y(z_1+tuz_2)^2})}\mathrm{d}t.$$

If $(y_k, z_{1k}, z_{2k}) \to (y, z_1, z_2)$, then for all t,

$$\frac{\mathrm{e}^{-y_k(z_{1k}+tu_kz_{2k})}}{(1+\mathrm{e}^{-y_k(z_{1k}+tu_kz_{2k})})^2} \to \frac{\mathrm{e}^{-y(z_1+tuz_2)}}{(1+\mathrm{e}^{-y(z_1+tuz_2)})^2}.$$

The sequence on the left is not greater than 1 for all t. Therefore, by the dominated convergence theorem $g_{u_k}(y_k, z_{1k}, z_{2k}) \to g_u(y, z_1, z_2)$. Then, by Proposition 1,

$$g_{u_k}(\tilde{Y}_k, \tilde{Z}_{1k}, \tilde{Z}_{2k}) \to_d g_u(\tilde{Y}, \tilde{Z}_1, \tilde{Z}_2).$$

The sequence of random variables on the left hand side is not greater than 1. Therefore, by the Proposition 4

$$\mathsf{E}g_u(\tilde{Y}, \tilde{Z}_1, \tilde{Z}_2) = \lim_{k\to\infty} \mathsf{E}g_{u_k}(\tilde{Y}_k, \tilde{Z}_{1k}, \tilde{Z}_{2k}) = \lim_{k\to\infty} \frac{\beta_k}{u_k} = 0.$$

We got a contradiction because g_u function is everywhere positive.

Step 5: the case, where $u_k \to \infty$.

From

$$\mathsf{E}\frac{1}{\tilde{Z}_{2k}^2} = \mathsf{E}\frac{Z_{2k}^2}{Z_{2k}^2} = 1$$

we get that the sequence of random variables $(1/|\tilde{Z}_{2k}|)$ is uniformly integrable. Then by Proposition 3

$$\mathsf{E}\frac{1}{|\tilde{Z}_2|} = \lim_{k\to\infty} \mathsf{E}\frac{1}{|\tilde{Z}_{2k}|} \le \sup_k \mathsf{E}\frac{1}{|\tilde{Z}_{2k}|} < \infty.$$

Therefore almost surely $\tilde{Z}_2 \neq 0$.

For all $u > 0, y \in \{-1, 1\}, z_1 \in \mathbb{R}$ and $z_2 \neq 0$,

$$\begin{aligned} ug_u(y, z_1, z_2) &= u\int_0^1 \frac{\mathrm{e}^{-y(z_1+tuz_2)}}{(1+\mathrm{e}^{-y(z_1+tuz_2)})^2}\mathrm{d}t = \frac{1}{yz_2}\frac{1}{1+\mathrm{e}^{-y(z_1+tuz_2)}}\Big|_0^1 \\ &= \frac{1}{yz_2}\left(\frac{1}{1+\mathrm{e}^{-y(z_1+uz_2)}} - \frac{1}{1+\mathrm{e}^{-yz_1}}\right) \\ &= \frac{\mathrm{e}^{-yz_1} - \mathrm{e}^{-y(z_1+uz_2)}}{yz_2(1+\mathrm{e}^{-y(z_1+uz_2)})(1+\mathrm{e}^{-yz_1})}. \end{aligned}$$

Let $u_k \to \infty$ and $(y_k, z_{1k}, z_{2k}) \to (y, z_1, z_2)$ with $z_2 \neq 0$. Then if $yz_2 < 0$, then

$$u_k g_{u_k}(y_k, z_{1k}, z_{2k}) \to -\frac{1}{yz_2(1+\mathrm{e}^{-yz_1})},$$

and if $yz_2 > 0$, then

$$u_k g_{u_k}(y_k, z_{1k}, z_{2k}) \to \frac{\mathrm{e}^{-yz_1}}{yz_2(1+\mathrm{e}^{-yz_1})}.$$

In other words,

$$u_k g_{u_k}(y_k, z_{1k}, z_{2k}) \to \frac{1}{|yz_2|(1+\mathrm{e}^{-yz_1})} h(y, z_1, z_2) = \frac{1}{|z_2|(1+\mathrm{e}^{-yz_1})} h(y, z_1, z_2),$$

where

$$h(y, z_1, z_2) = \begin{cases} 1, & \text{if } yz_2 < 0, \\ \mathrm{e}^{-yz_1}, & \text{if } yz_2 > 0. \end{cases}$$

By Proposition 1,

$$u_k g_{u_k}(\tilde{Y}_k, \tilde{Z}_{1k}, \tilde{Z}_{2k}) \to_d \frac{1}{|\tilde{Z}_2|(1+\mathrm{e}^{-\tilde{Y}\tilde{Z}_1})} h(\tilde{Y}, \tilde{Z}_1, \tilde{Z}_2).$$

The sequence of random variables on the left hand side is dominated by the sequence $(1/|\tilde{Z}_{2k}|)$ which is uniformly integrable. Therefore by Proposition 4

$$\mathsf{E}\frac{1}{|\tilde{Z}_2|} h(\tilde{Y}, \tilde{Z}_1, \tilde{Z}_2) = \lim_{k\to\infty} u_k \mathsf{E} g_{u_k}(\tilde{Y}_k, \tilde{Z}_{1k}, \tilde{Z}_{2k}) = \lim_{k\to\infty} \beta_k = 0.$$

Again, we got a contradiction because almost surely $\dfrac{1}{|\tilde{Z}_2|} h(\tilde{Y}, \tilde{Z}_1, \tilde{Z}_2) > 0$. □

It remains to estimate the probability $\mathsf{P}(W_{kn}^{'c})$. In order to do this, we have to estimate

$$\sup_{\theta \in \bar{U}_k} \|\Psi_{k,n}(\theta) - \Psi_k(\theta)\|.$$

Fix $\theta \in \bar{U}_k$ and denote $d\theta = \theta - \theta_k$. By using Taylor's expansion we get

$$\begin{aligned} \Psi_{k,n}(\theta) &= \Psi_{k,n}(\theta_k) + \Psi'_{k,n}(\theta_k) d\theta + r_{k,n}(\theta, d\theta), \\ \Psi_k(\theta) &= \Psi'_k(\theta_k) d\theta + r_k(\theta, d\theta), \end{aligned}$$

where

$$\begin{aligned} \|r_{k,n}(\theta, d\theta)\| &\le \sup_{0<t<1} \|\Psi''_{k,n}(\theta_k + t d\theta)\| \|d\theta\|^2 \le \overline{\|X\|^3} d_k^2, \\ \|r_k(\theta, d\theta)\| &\le \sup_{0<t<1} \|\Psi''_k(\theta_k + t d\theta)\| \|d\theta\|^2 \le \mathsf{E}\|X\|^3 d_k^2. \end{aligned}$$

Therefore,

$$\sup_{\theta\in\bar{U}_k}\|\Psi_{k,n}(\theta)-\Psi_k(\theta)\| \leq \|\Psi_{k,n}(\theta_k)\|+d_k\|\Psi'_{k,n}(\theta_k)-\Psi'_k(\theta_k)\|+d_k^2(\overline{\|X\|^3}+\mathsf{E}\|X\|^3)$$

and

$$\begin{aligned}\mathsf{P}(W_{nk}^{'c}) \leq \mathsf{P}(\|\Psi_{k,n}(\theta_k)\|>\delta_k/3)+\mathsf{P}(d_k\|\Psi'_{k,n}(\theta_k)-\Psi'_k(\theta_k)\|>\delta_k/3)\\ +\mathsf{P}(d_k^2(\overline{\|X\|^3}+\mathsf{E}\|X\|^3)>\delta_k/3).\end{aligned} \tag{12}$$

The first term on the right hand of (12) is estimated as follows. Let $(e_1,\dots,e_k)$ be an orthonormal basis of E_k. Then

$$\begin{aligned}\mathsf{E}\|\Psi_{k,n}(\theta_k)\|^2 &= \sum_{j=1}^k \mathsf{E}\langle\Psi_{k,n}(\theta_k),e_j\rangle^2 = \sum_{j=1}^k \mathsf{Var}\langle\Psi_{k,n}(\theta_k),e_j\rangle\\ &= \frac{1}{n}\sum_{j=1}^k \mathsf{Var}\langle\psi_{k,\theta_k}(X,Y),e_j\rangle = \frac{1}{n}\sum_{j=1}^k \mathsf{E}\langle\psi_{k,\theta_k}(X,Y),e_j\rangle^2\\ &= \frac{1}{n}\mathsf{E}\|\psi_{k,\theta_k}(X,Y)\|^2 \leq \frac{1}{n}\mathsf{E}\|X\|^2.\end{aligned}$$

Therefore, the probability that we are interested does not exceed

$$\frac{9\mathsf{E}\|X\|^2}{n\delta_{k_n}^2}.$$

Similarly, we can evaluate the second term of (12). Again, we would like to apply Chebyshev's inequality and get that

$$\mathsf{P}(Z>\delta_k/3d_k) \leq \frac{9d_k^2}{\delta_k^2}\mathsf{E}Z^2,$$

where $Z=\|\Psi'_{k,n}(\theta_k)-\Psi'_k(\theta_k)\|$. However, since $\Psi_{k,n}$ is a vector-valued function, its derivative is a linear operator which makes the exact computation of its norm very complex. To make things simpler, here we can use the Hilbert-Schmidt norm instead, which is known to be greater than usual norm. Therefore,

$$\begin{aligned}&\mathsf{E}\|\Psi'_{k,n}(\theta_k)-\Psi'_k(\theta_k)\|^2\\ &\leq \sum_{j,j'=1}^k \mathsf{E}(\langle\Psi'_{k,n}(\theta_k)e_{j'},e_j\rangle-\langle\Psi'_k(\theta_k)e_{j'},e_j\rangle)^2\\ &= \sum_{j,j'=1}^k \mathsf{Var}\langle\Psi'_{k,n}(\theta_k)e_{j'},e_j\rangle = \frac{1}{n}\sum_{j,j'=1}^k \mathsf{Var}\langle\psi'_{k,\theta_k}(X,Y)e_{j'},e_j\rangle\\ &\leq \frac{1}{n}\sum_{j,j'=1}^k \mathsf{E}\langle\psi'_{k,\theta_k}(X,Y)e_{j'},e_j\rangle^2 = \frac{1}{n}\sum_{j,j'=1}^k \mathsf{E}(m''_{\theta_k}(X,Y)(e_{j'},e_j))^2\end{aligned}$$

$$\leq \frac{1}{n}\sum_{j,j'=1}^{k}\mathsf{E}\langle X,e_j\rangle^2\langle X,e_{j'}\rangle^2 = \frac{1}{n}\mathsf{E}\left(\sum_{j=1}^{k}\langle X,e_j\rangle^2\right)^2 = \frac{1}{n}\mathsf{E}\|X^{(k)}\|^4$$
$$\leq \frac{1}{n}\mathsf{E}\|X\|^4$$

and the second term on the right hand side of (12) does not exceed

$$\frac{9\mathsf{E}\|X\|^4 d_{k_n}^2}{n\delta_{k_n}^2}.$$

The third term of (12) tends to 0, if $d_{k_n}^2/\delta_{k_n}\to 0$.

Therefore, Theorem 1 will be proved, if we can select δ_k such that

$$d_{k_n}\to 0,\quad n\delta_{k_n}^2\to\infty,\quad \frac{d_{k_n}^2}{n\delta_{k_n}^2}\to 0,\quad \frac{d_{k_n}^2}{\delta_{k_n}}\to 0.$$

Note that the third condition is implied by the first and the second ones. If we take $\delta_k = o(\tau_k^2)$, then the first and the fourth conditions are met because then $d_k = O(\delta_k/\tau_k) = o(1)$ and $d_k^2/\delta_k = O(\delta_k/\tau_k^2) = o(1)$. Therefore, it is enough to select $\delta_k = o(\tau_k^2)$ such that $n\delta_{k_n}^2\to\infty$, that is, in such a way that asymptotically

$$n^{-1/2}\prec\delta_{k_n}\prec\tau_{k_n}^2,$$

where $a\prec b$ means that $a = o(b)$. Clearly, we can achieve this, if

$$n^{-1/2}\prec\tau_{k_n}^2,$$

that is, if $n\tau_{k_n}^4\to\infty$ which is exactly the assumption of Theorem 1.

5.5. Proof of Theorem 2

Proof. Define a new Hilbert space $\bar{E} = \mathbb{R}\times E$ with the inner product

$$\langle(\alpha,\theta),(a,x)\rangle = \alpha a + \langle\theta,x\rangle,$$

where $\alpha,a\in\mathbb{R}$ and $\theta,x\in E$, and set $\bar{X} = (1,X)\in\bar{E}$. Take any $\bar{\theta} = (\alpha,\theta)\neq 0$. If $\theta\neq 0$, then $\mathsf{P}(\langle\bar{\theta},\bar{X}\rangle = 0) = 0$ because of (FR').

If $\theta = 0$, then $\alpha\neq 0$ and therefore

$$\mathsf{P}(\langle\bar{\theta},\bar{X}\rangle = 0) = \mathsf{P}(\alpha = 0) = 0.$$

Hence $\bar{X}$ satisfies condition (FR). Moreover, if X satisfies (M), then

$$\mathsf{E}\|\bar{X}\|^4 = \mathsf{E}\langle\bar{X},\bar{X}\rangle^2 = \mathsf{E}\left(1+\langle X,X\rangle\right)^2 = 1+2\mathsf{E}\|X\|^2+\mathsf{E}\|X\|^4<\infty,$$

that is, $\bar{X}$ also satisfies (M). Finally, suppose X satisfies (UI). Fix ϵ and find c_0 such that for all $c>c_0$ and all θ

$$\mathsf{E}\langle\theta,X\rangle^2 1_{\{\langle\theta,X\rangle^2>(c\mathsf{E}\langle\theta,X\rangle^2)/2\}}\leq\epsilon\mathsf{E}\langle\theta,X\rangle^2.$$

Denote $\bar{c}_0 = \max(c_0, 2, 1/\epsilon)$. Take $c > \bar{c}_0$ and any $\bar{\theta} = (\alpha, \theta)$ with norm equal to 1. Then by Chebyshev's inequality

$$\alpha^2 \mathsf{P}((\alpha + \langle \theta, X\rangle)^2 > c(\alpha^2 + \mathsf{E}\langle \theta, X\rangle^2)) \le \alpha^2/c \le \alpha^2\epsilon$$

and

$$\begin{aligned}\mathsf{E}\langle \theta, X\rangle^2 1_{\{(\alpha+\langle\theta,X\rangle)^2 > c(\alpha^2+\mathsf{E}\langle\theta,X\rangle^2)\}} &\le \mathsf{E}\langle \theta, X\rangle^2 1_{\{2\alpha^2+2\langle\theta,X\rangle^2 > c(\alpha^2+\mathsf{E}\langle\theta,X\rangle^2)\}} \\ &= \mathsf{E}\langle \theta, X\rangle^2 1_{\{2\langle\theta,X\rangle^2 > c\mathsf{E}\langle\theta,X\rangle^2 + (c-2)\alpha^2\}} \\ &\le \mathsf{E}\langle \theta, X\rangle^2 1_{\{\langle\theta,X\rangle^2 > c/2\mathsf{E}\langle\theta,X\rangle^2\}} \\ &< \epsilon \mathsf{E}\langle \theta, X\rangle^2.\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathsf{E}\langle \bar{\theta}, \bar{X}\rangle^2 1_{\{\langle\bar{\theta},\bar{X}\rangle^2 > c(\mathsf{E}\langle\bar{\theta},\bar{X}\rangle^2)\}} \\ &= \mathsf{E}(\alpha + \langle \theta, X\rangle)^2 1_{\{(\alpha+\langle\theta,X\rangle)^2 > c(\alpha^2+\mathsf{E}\langle\theta,X\rangle^2)\}} \\ &\le 2\alpha^2 \mathsf{E} 1_{\{(\alpha+\langle\theta,X\rangle)^2 > c(\alpha^2+\mathsf{E}\langle\theta,X\rangle^2)\}} + 2\mathsf{E}\langle \theta, X\rangle^2 1_{\{(\alpha+\langle\theta,X\rangle)^2 > c(\alpha^2+\mathsf{E}\langle\theta,X\rangle^2)\}} \\ &\le 2\epsilon(\alpha^2 + \mathsf{E}\langle \theta, X\rangle^2),\end{aligned}$$

that is, $\bar{X}$ satisfies condition (UI).

Define

$$\bar{C}(\bar{\theta}_1, \bar{\theta}_2) = \mathsf{E}\langle \bar{\theta}_1, \bar{X}\rangle\langle \bar{\theta}_2, \bar{X}\rangle, \quad \bar{\tau}_k = \min_{\substack{\bar{\theta}\in\mathbb{R}\times E_k \\ \|\bar{\theta}\|=1}} \bar{C}(\bar{\theta}, \bar{\theta}).$$

Note that

$$\bar{C}(\bar{\theta}, \bar{\theta}) = \mathsf{E}\langle \bar{\theta}, \bar{X}\rangle^2 = \mathsf{E}(\alpha+\langle\theta, X\rangle)^2 = \alpha^2 + 2\alpha\mathsf{E}\langle\theta, X\rangle + \mathsf{E}\langle\theta, X\rangle^2 = \alpha^2 + \mathsf{E}\langle\theta, X\rangle^2.$$

Since C is a bilinear form, for all $\theta \in E_k$

$$\alpha^2 + C(\theta, \theta) = \alpha^2 + \|\theta\|^2 C(\theta/\|\theta\|, \theta/\|\theta\|) \ge \alpha^2 + \|\theta\|^2 \tau_k.$$

Therefore,

$$\bar{\tau}_k \ge \min_{|\alpha|\le 1} \left(\alpha^2 + (1-\alpha^2)\tau_k\right) = \min(1, \tau_k)$$

and

$$n\bar{\tau}_{k_n}^4 = n\min(1, \tau_{k_n}^4) = \min(n, n\tau_{k_n}^4) \to \infty.$$

Then, by Theorem 1, the corresponding logistic estimate

$$\tilde{\theta}_{kn} = \arg \min_{\bar{\theta}\in\mathbb{R}\times E_k} \bar{M}_n(\bar{\theta}), \tag{13}$$

where

$$\bar{M}_n(\bar{\theta}) = \overline{m_{\bar{\theta}}(\bar{X}, Y)}, \quad m_{\bar{\theta}}(\bar{x}, y) = \log(1 + \mathrm{e}^{-y\langle\bar{\theta},\bar{x}\rangle})$$

is consistent on $\bar{E} = \mathbb{R} \times E$. It remains to note that the logistic estimate (13) is the same as the estimate (6). □

Acknowledgments

This work was supported by the Department of Statistical Science, University College London, United Kingdom, and Singapore Bioimaging Consortium, Agency for Science, Technology and Research, Singapore. The authors would like to thank prof. Vytautas Kazakevicius for guidance and support when working on this problem as well as to the anonymous referee and the anonymous associate editor for the ideas on future work.

References

Albert, A. & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71(1)**, 1–10. MR0738319

Bertsekas, D. P., Nedic, A. & Ozdaglar, A. E. (2003). *Convex Analysis and Optimization.* Athena Scientific. MR2184037

Chen, K., Hu, I. & Ying, Z. (1999). Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *Ann. Statist.* **27(4)**, 1155–1163. MR1740117

Escabias, M., Aguilera, A. M. & Valderrama, M. J. (2007). Functional PLS logit regression model. *Comput. Statist. Data Anal.* **51(10)**, 4891–4902. MR2364547

Fan, J. & Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38(6)**, 3567–3604. MR2766861

Van de Geer, S.A. (2008). High-dimensional generalized linear models and the Lasso. *Ann. Statist.* **38(2)**, 614–645. MR2396809

Kallenberg, O. (2001). *Foundations of Modern Probability*, 2nd edn. Springer. MR1464694

Kazakeviciute, A. & Olivo, M. (2016). A study of logistic classifier: uniform consistency in finite-dimensional linear spaces. *Journal of Mathematics, Statistics and Operations Research* **3(2)**, 1–7.

Kazakeviciute, A. & Olivo, M. (2017). Point separation in logistic regression on Hilbert space-valued variables. *Statist. Probab. Lett.*, **128**, 84-88. MR3656380

Kazakeviciute, A., Kazakevicius, V. & Olivo, M. (2017). Conditions for existence of uniformly consistent classifiers. *IEEE Trans. Inform. Theory*, **63(6)**, 3425-3432. MR3658533

Liang, H. & Du, P. (2012). Maximum likelihood estimation in logistic regression models with a diverging number of covariates. *Electron. J. Stat.* **6**, 1838–1846. MR2988466

Müller, H. G. & Stadtmüller, S. (2005). Generalized Functional Linear Models. *Ann. Statist.* **32(2)**, 774–805. MR2163159

Ramsay, J. O. & Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies.* Springer. MR1910407

Ramsay, J. O. & Silverman, B. W. (2005). *Functional data analysis.* Springer. MR2168993

VAN DER VAART, A.W. (2000). *Asymptotic Statistics.* Cambridge University Press. MR1652247

VAN RYZIN, J. (1966). Bayes risk consistency of classification procedures using density estimation. *Sankhya: The Indian Journal of Statistics, Series A.*, **28(2/3)**, 261–270. MR0210264

WANG, L. (2011). GEE analysis of clustered binary data with diverging number of covariates. *Ann. Statist.* **39(1)**, 389–417. MR2797851