A study of logistic classifier: uniform consistency in finite-dimensional linear spaces

Agne Kazakeviciute, Malini Olivo

***Abstract*—Let X be a random variable taking values in a finite dimensional linear space and $Y \in \{0, 1\}$ its associated label. We study the case, where conditional distribution $p(x) = \mathsf{P}(Y = 1 \mid X = x)$ depends on x through some linear form θx. We show that in this case, under a mild assumption on the distribution μ of X, a maximum-likelihood estimator $\hat{p}$, as well as the induced class of logistic classifiers, are uniformly (w.r.t. p) consistent.**

***Index Terms*—Uniform consistency, logistic classifier, finite-dimensional linear spaces.**

I. Introduction

We consider the statistical problem of binary classification, where the goal is to attach every x from a finite-dimensional linear space E to one of the two classes, 0 or 1. Formally, a *classifier* is a Borel function $\delta\colon E \to \{0, 1\}$. The pair (x, y), where $y \in \{0, 1\}$ is the true class of x, is considered as a realization of a random vector (X, Y). We denote the distribution of X by μ and the conditional probability of $Y = 1$, given $X = x$, by $p(x)$. The function p is considered as an element of $\mathcal{L}^1(E, \mu)$, the space of all μ-integrable functions, endowed with the semi-metric

$$d(p_1, p_2) = \int_E |p_1 - p_2| \mathrm{d}\mu.$$

Usually the distribution μ of X and the function p are unknown and are estimated from the training sample $(X_1, Y_1), \dots, (X_n, Y_n)$. Each estimator $\hat{p}$ induces a class of empirical classifiers of the form

$$\hat{\delta}_u(x) = \begin{cases} 1, & \text{if } \hat{p}(x) > u, \\ 0, & \text{otherwise.} \end{cases}$$

where u is the pre-selected threshold. Following [4], we call an estimate $\hat{p}$ *consistent*, if $\hat{p}(X)$ tends in probability to $p_0(X)$, where p_0 is the true conditional probability. We call $\hat{p}$ *uniformly consistent*, if the convergence is uniform in p_0. In our previous work (see [4]) we have proved that the consistency/uniform consistency of $\hat{p}$ yields consistency/uniform consistency of the associated classifier $\hat{\delta}_u$ (see [4] for definitions of consistency and uniform consistency of $\hat{\delta}_u$).

The notion of uniform consistency is important from the practical point of view (it is of no use in knowing that $\mathsf{P}_{p_0}\{d(p_0, \hat{p}) \geq \varepsilon\} < \delta$ for $n \geq n_0$, if that n_0 depends on the unknown distribution of (X, Y)) and may be considered as the first step in analysis of convergence rate of estimators. However, to the best of our knowledge, currently there are no literature that would consider this problem (it should not be confused with the similar notion meaning that $\sup_x |\hat{p}(x) - p_0(x)| \xrightarrow[n\to\infty]{} 0$ in probability or almost surely). While in our previous work (see [4]) we have proved uniform consistency of histogram-type estimators, in this work we are studying the behavior of logistic classifiers. More specifically, we study the case, where p is known to belong to the set

$$P(E) = \{p_\theta \mid \theta \in E'\},$$

where E' is the dual space of E,

$$p_\theta(x) = \phi(\theta x),$$

and ϕ is a known link function (i.e. an increasing homeomorphism between $\mathbb{R} \cup \{-\infty, \infty\}$ and $[0; 1]$). The usual choice of ϕ is the logistic link, i.e.

$$\phi(\theta x) = \frac{1}{1 + \mathrm{e}^{-\theta x}}.$$

Due to technical reasons, we find it more convenient to work with the set $\bar{P}(E)$, the set consisting of all functions p with the following property: there exists $(p_n) \subset P(E)$ such that $p_n(x) \xrightarrow[n\to\infty]{} p(x)$, for all x. For $p, p_0 \in \bar{P}(E)$ denote

$$\begin{aligned}
m_p(x, y) &= \begin{cases} \log p(x), & \text{if } y = 1, \\ \log(1 - p(x)), & \text{if } y = 0, \end{cases} \\
M(p_0, p) &= \mathsf{E}_{p_0} m_p(X, Y), \\
M_n(p) &= \overline{m_p(X, Y)}, \\
m_{p,r}(x, y) &= \sup_{\substack{p' \in \bar{P}(E) \\ d(p', p) < r}} m_{p'}(x, y), \\
M(p_0, p, r) &= \mathsf{E}_{p_0} m_{p,r}(X, Y), \\
M_n(p, r) &= \overline{m_{p,r}(X, Y)},
\end{aligned}$$

where E_{p_0} is the expectation with respect to the true conditional probability p_0 and, for any function $f(X, Y)$, $\overline{f(X, Y)} = \frac{f(X_1, Y_1) + \dots + f(X_n, Y_n)}{n}$. $m_p(x_i, y_i)$ and $M(p_0, p)$ are better known as log-posterior, given its true label, and expected log-posterior, given true posterior distributions of the two populations, respectively. Define

$$\hat{p} = \arg \max_{p \in \bar{P}(E)} M_n(p). \tag{1}$$

The main result of this paper is the following theorem.

Theorem 1. *Let the distribution μ of X satisfy the following condition: for all $p \in \bar{P}(E)$, there exists a set A with*

Agne Kazakeviciute is with the Department of Statistical Science, University College London, London, WC1E 6BT, United Kingdom, and with Singapore Bioimaging Consortium, Agency for Science, Technology and Research, Singapore; e-mail: a.kazakeviciute.12@ucl.ac.uk

Malini Olivo is with Singapore Bioimaging Consortium, Agency for Science, Technology and Research, Singapore, and with the School of Physics, National University of Ireland, Galway, University Road, Galway, Ireland

DOI: 10.5176/2251-3388-3.2.65

$\mu(A) = 1$ *and such that, for all* $x \in A$ *and all* $p' \in \bar{P}(E)$ *with* $d(p', p) = 0$,

$$p'(x) = p(x). \tag{2}$$

Then the estimator $\hat{p}$ *defined as* (1) *is uniformly consistent, that is, for all* $\epsilon > 0$

$$\sup_{p_0 \in \bar{P}(E)} \mathsf{P}_{p_0}\{d(p_0, \hat{p}) \geq \epsilon\} \xrightarrow[n\to\infty]{} 0.$$

The work is organized as follows. In Section II we prove that $\bar{P}(E)$ is a compact set which is a crucial assumption for our main result. In Section III we establish some properties of functions $M(p_0, p)$ and $M(p_0, p, r)$. We follow with the Section IV, where we prove two uniform versions of the Weak Law of Large Numbers. Finally, in Section V we prove Theorem 1 and in Section VI we discuss how restrictive is the condition on μ.

II. Properties of $\bar{P}(E)$

We denote the elements of E' by θ, each θ is a linear functional on E and the value of θ at x is denoted by θx.

Lemma 1. *Each sequence* $(\theta_n) \subset E'$ *contains a subsequence* (θ_{n_k}) *such that, for all* $x \in E$, $\theta_{n_k} x$ *has a limit, finite or infinite.*

Proof: If $\dim E = 0$, then $E = \{0\}$. If $(\theta_n) \subset E'$, then $\theta_n 0 = 0$, for all n, because each linear functional maps 0 to 0. Therefore, $\theta_n x \xrightarrow[n\to\infty]{} 0$, for all $x \in E$.

Now suppose that $\dim E \geq 1$ and fix any sequence $(\theta_n) \subset E'$. There are two options:

1) (θ_n) is bounded, i.e. $||\theta_n|| \leq c$ with some $c \in \mathbb{R}$. It is well known that every bounded set in a finite-dimensional linear space is precompact. Therefore there exists a subsequence (θ_{n_k}) which tends to some $\theta \in E'$. Obviously, then $\theta_{n_k} x \xrightarrow[k\to\infty]{} \theta x$, for all $x \in E$.
2) (θ_n) is not bounded. Define $a_n = \frac{\theta_n}{\|\theta_n\|}$ and note that $\|a_n\| = 1$ for all n, i.e. (a_n) is a bounded sequence. Find $n_k \xrightarrow[k\to\infty]{} \infty$ such that $\|\theta_{n_k}\| \xrightarrow[k\to\infty]{} \infty$ and $a_{n_k} \xrightarrow[k\to\infty]{} a \in E'$. Note that $\|a\| = 1$ and therefore the set $H = \{x \mid ax = 0\}$ is the hypersubspace of E: $\dim H = \dim E - 1 < \dim E$. Let θ'_n denote the restriction of θ_n on H. Then $\theta'_n \in H'$. By induction assumption there exists a subsequence $(\theta'_{n_{k(l)}})$ of (θ'_{n_k}) such that $\theta_{n_{k(l)}} x = \theta'_{n_{k(l)}} x$ has a limit, finite or infinite, for all $x \in H$. If $x \notin H$, then either $ax > 0$ and then $\theta_{n_{k(l)}} x = \frac{\theta_{n_{k(l)}} x}{\|\theta_{n_{k(l)}}\|} \|\theta_{n_{k(l)}}\| \xrightarrow[l\to\infty]{} \infty$, or $ax < 0$ and $\theta_{n_{k(l)}} x = \frac{\theta_{n_{k(l)}} x}{\|\theta_{n_{k(l)}}\|} \|\theta_{n_{k(l)}}\| \xrightarrow[l\to\infty]{} -\infty$.

■

Lemma 2. $\bar{P}(E)$ *is a compact set in* $\mathcal{L}^1(E, \mu)$.

Proof: Let $(q_n) \subset \bar{P}(E)$. We need to find its convergent (in $\mathcal{L}^1(E, \mu)$) subsequence whose limit is some $q \in \bar{P}(E)$. For all n, there exists a sequence $(p_{nm} \mid m \geq 1) \subset P(E)$ such that $p_{nm}(x) \xrightarrow[m\to\infty]{} q_n(x)$, for all x. Obviously, then $p_{nm} \xrightarrow[m\to\infty]{} q_n$ in $\mathcal{L}^1(E, \mu)$. Find m_n such that $d(p_{nm_n}, q_n) < 1/n$ and denote $p_n = p_{nm_n}$. By Lemma 1, there exists a sequence $n_k \xrightarrow[k\to\infty]{} \infty$ and a function q such that $p_{n_k}(x) \xrightarrow[k\to\infty]{} q(x)$, for all x. Then, by the definition of $\bar{P}(E)$, $q \in \bar{P}(E)$. Moreover, $p_{n_k} \xrightarrow[k\to\infty]{} q$ in $\mathcal{L}^1(E, \mu)$. Therefore,

$$d(q_{n_k}, q) \leq d(q_{n_k}, p_{n_k}) + d(p_{n_k}, q) < \frac{1}{n_k} + d(p_{n_k}, q) \xrightarrow[k\to\infty]{} 0,$$

that is, $q_{n_k} \xrightarrow[k\to\infty]{} q$. ■

Let $S(E')$ be the unit sphere in E' and for all $a \in S(E')$

$$\ker a = \{x \in E \mid ax = 0\}.$$

Lemma 1 suggests that $p \in \bar{P}(E) \backslash P(E)$ has the following form:

$$p(x) = \begin{cases} 0, & \text{if } a_1 x < 0, \\ q_1(x), & \text{if } x \in E_1, \\ 1, & \text{if } a_1 x > 0, \end{cases} \tag{3}$$

where $a_1 \in S(E')$, $E_1 = \ker a_1$ and $q_1 \in \bar{P}(E_1)$. Then we have two situations: function q_1 belongs to $P(E_1)$, or it is of form (3) form, but with some $a_2 \in S(E'_1)$, $E_2 = \ker a_2$ and $q_2 \in \bar{P}(E_2)$ instead of a_1, E_1, q_1. We can repeat this reasoning several times but eventually we will reach a point, where $\dim E_k$ equals 0. Then $E_k = \{0\}$ and $\bar{P}(E_k) = P(E_k)$ contains only one function which maps 0 (the only point of E_k) to $\varphi(0)$ (e.g. to $\frac{1}{2}$, if φ is the standard logistic function).

For example, if $\dim E = 2$, then $\bar{P}(E)$ contains the following three types of functions:

$$\begin{gathered} p_\theta \quad \text{for } \theta \in E', \\ p_{a_1,\theta_1} \quad \text{for } a_1 \in S(E'),\ E_1 = \ker a_1 \text{ and } \theta_1 \in E'_1, \\ p_{a_1,a_2,0} \quad \text{for } a_1 \in S(E'),\ E_1 = \ker a_1,\ a_2 \in S(E'_1), \end{gathered}$$

where

$$p_{a_1,\theta_1}(x) = \begin{cases} 0, & \text{if } a_1 x < 0, \\ \varphi(\theta_1 x), & \text{if } a_1 x = 0, \\ 1, & \text{if } a_1 x > 0 \end{cases}$$

and

$$p_{a_1,a_2,0}(x) = \begin{cases} 0, & \text{if } a_1 x < 0, \\ 0, & \text{if } a_1 x = 0,\ a_2 x < 0, \\ \varphi(0), & \text{if } a_1 x = 0,\ a_2 x = 0, \\ 1, & \text{if } a_1 x = 0,\ a_2 x > 0, \\ 1, & \text{if } a_1 x > 0. \end{cases}$$

III. Properties of M

Lemma 3. $M(p_0, p)$ *is an upper semi-continuous function.*

Proof: By one of equivalent definitions of upper semi-continuity (see p. 602 statement B1 from [6]) we need to prove that

$$(p_{0n}, p_n) \xrightarrow[n\to\infty]{} (p_0, p) \Rightarrow \overline{\lim_{n\to\infty}} M(p_{0n}, p_n) \leq M(p_0, p).$$

Suppose the contrary and find $(p_{0n}, p_n) \xrightarrow[n\to\infty]{} (p_0, p)$ such that

$$\overline{\lim_{n\to\infty}} M(p_{0n}, p_n) > M(p_0, p).$$

Find a subsequence (p_{0n_k}, p_{n_k}) such that $\underline{\lim}_{k\to\infty} M(p_{0n_k}, p_{n_k})$ exists and equals $\underline{\lim}_{n\to\infty} M(p_{0n}, p_n)$. We can suppose that almost everywhere $p_{0n_k}(x) \xrightarrow[n\to\infty]{} p_0(x)$, $p_{n_k}(x) \xrightarrow[n\to\infty]{} p(x)$. Then almost surely

$$\begin{aligned} 0 &\le -m_{p_{n_k}}(X,1)p_{0n_k}(X) - m_{p_{nk}}(X,0)(1-p_{0n_k}(X)) \\ &\xrightarrow[k\to\infty]{} -m_p(X,1)p_0(X) - m_p(X,0)(1-p_0(X)), \end{aligned}$$

which yields, by Fatou's lemma,

$$\begin{aligned} &\underline{\lim}_{n\to\infty} \mathsf{E}_{p_{0n_k}}(-m_{p_{n_k}}(X,Y)) \ge \mathsf{E}_{p_0}(-m_p(X,Y)), \\ &\overline{\lim}_{n\to\infty} \mathsf{E}_{p_{0n_k}} m_{p_{n_k}}(X,Y) \le \mathsf{E}_{p_0} m_p(X,Y), \\ &\lim_{n\to\infty} M(p_{0n_k}, p_{n_k}) \le M(p_0, p). \end{aligned}$$

We got a contradiction. ■

Lemma 4. *The function $p_0 \mapsto M(p_0, p_0)$ is continuous.*

Proof: Let $p_{0n} \xrightarrow[n\to\infty]{} p_0$. Note that the function $p \log p$ is continuous in interval $[0;1]$ (if we define $0 \log 0 = 0$), and therefore bounded. Hence by the dominated convergence theorem, $M(p_{0n}, p_{0n}) \xrightarrow[n\to\infty]{} M(p_0, p_0)$. ■

Lemma 5. *$M(p_0, p) < M(p_0, p_0)$, if $d(p_0, p) > 0$.*

Proof: We use Lemma 5.35 in [7]. Denote the counting measure on $\{0, 1\}$ by ν. If $d(p, p') > 0$, then conditional distributions of Y w.r.t. X corresponding to p and p' are different which implies that distributions of (X, Y) differ as well. For any function f

$$\begin{aligned} &\mathsf{E}_p f(X,Y) \\ &= \mathsf{E}\left[f(X,1)p(X) + f(X,0)(1-p(X))\right] \\ &= \int_E \mu(dx) \int_{\{0,1\}} f(x,y) q_p(x,y) \nu(dy), \end{aligned}$$

where

$$q_p(x,y) = \begin{cases} p(x), & \text{if } y = 1, \\ 1 - p(x), & \text{if } y = 0, \end{cases}$$

that is, $q_p(x, y)$ is the density of (X, Y) with respect to $\mu \times \nu$. By Lemma 5.35 in [7], p_0 is the unique maximum of $\mathsf{E}_{p_0} \log \frac{q_p(X,Y)}{q_{p_0}(X,Y)}$. It is enough to note that $m_p(x, y) = \log q_p(x, y)$ which is why

$$\begin{aligned} \mathsf{E}_{p_0} \log \frac{q_p(X,Y)}{q_{p_0}(X,Y)} &= \mathsf{E}_{p_0} \log q_p(X,Y) - \mathsf{E}_{p_0} \log q_{p_0}(X,Y) \\ &= \mathsf{E}_{p_0} m_p(X,Y) - \mathsf{E}_{p_0} m_{p_0}(X,Y) \\ &= M(p_0, p) - M(p_0, p_0). \end{aligned}$$

Therefore p_0 is the unique maximum of $M(p_0, p)$. ■

Lemma 6. *For all p, r and (x, y)*

$$m_{p,r}(x,y) = \sup_{\substack{d(p',p)<r \\ p' \in P(E)}} m_{p'}(x,y).$$

Proof: Fix p, r and x. It is enough to prove that

$$\begin{aligned} \sup_{\substack{d(p',p)<r \\ p' \in \bar{P}(E)}} p'(x) &= \sup_{\substack{d(p',p)<r \\ p' \in P(E)}} p'(x), \\ \inf_{\substack{d(p',p)<r \\ p' \in \bar{P}(E)}} p'(x) &= \inf_{\substack{d(p',p)<r \\ p' \in P(E)}} p'(x). \end{aligned} \tag{4}$$

We will prove the first equality, the second is proved analogously. Denote the left hand side of (4) by c and the right hand side by c'. Obviously, $c' \le c$ and it is sufficient to prove that $c \le c'$. Fix any $p' \in \bar{P}(E)$ with $d(p', p) < r$. There exists a sequence $(p_n) \subset P(E)$ such that $p_n(x') \xrightarrow[n\to\infty]{} p'(x')$, for all x'. Then $p_n \xrightarrow[n\to\infty]{} p'$ in $\mathcal{L}^1(E, \mu)$. Fix n_0 such that $d(p_n, p') < r - d(p', p)$ for $n \ge n_0$. Then for $n \ge n_0$

$$d(p_n, p) \le d(p_n, p') + d(p', p) < r$$

and therefore

$$c' \ge \sup_n p_n(x) \ge \lim_n p_n(x) = p'(x). \tag{5}$$

Note that (5) holds for every p' with $d(p', p) < r$. Therefore, also $c \le c'$. ■

Lemma 7. *For any $p \in \bar{P}(E)$ and all x,*

$$\begin{aligned} \lim_{r\to 0} \sup_{\substack{p' \in \bar{P}(E) \\ d(p',p)<r}} p'(x) &= \sup_{\substack{p' \in \bar{P}(E) \\ d(p',p)=0}} p'(x). \\ \lim_{r\to 0} \inf_{\substack{p' \in \bar{P}(E) \\ d(p',p)<r}} p'(x) &= \inf_{\substack{p' \in \bar{P}(E) \\ d(p',p)=0}} p'(x). \end{aligned}$$

Proof: We will prove the first equality, the second is proved analogously. Fix $p \in \bar{P}(E)$, x and denote

$$c_r = \sup_{\substack{p' \in \bar{P}(E) \\ d(p',p)<r}} p'(x), \quad c = \sup_{\substack{p' \in \bar{P}(E) \\ d(p',p)=0}} p'(x).$$

Since c_r decreases, if $r \downarrow 0$, the limit $c' = \lim_{r\to 0} c_r$ exists and we need to prove that it equals c.

Obviously, $c_r \ge c$ for all r. Therefore, $c' \ge c$. We will prove the converse inequality. By Lemma 6, there exists $p_n \in P(E)$ such that $d(p_n, p) \xrightarrow[n\to\infty]{} 0$ and $p_n(x) \xrightarrow[n\to\infty]{} c'$. Without loss of generality we can suppose that $p_n(x') \xrightarrow[n\to\infty]{} p'(x')$, for all x', where p' is a fixed function in $\bar{P}(E)$. Then $p_n \xrightarrow[n\to\infty]{} p'$ in $\mathcal{L}^1(E, \mu)$. Therefore, $d(p', p) = 0$ and $c' = p'(x) \le c$. ■

If the condition (2) holds, then it follows from Lemma 7 that $M(p_0, p, r) \xrightarrow[r\to 0]{} M(p_0, p)$.

IV. Uniform Laws of Large Numbers

In this Section, we present two Lemmas that are the uniform versions of the Weak Law of Large Numbers. In Lemma 8, random variables are required to have finite expectations and the proof is analogous to that of the Khinchin's theorem (see [5], Theorem 4.15). In Lemma 9, the expectations of random variables might be infinite but random variables are required to be non-positive. The proof of Lemma 9 is analogous to that of Cramer's theorem (see [1]).

Recall that a family (Z_s) of random variables, where each Z_s is defined in a probability space (Ω_s, P_s), is called *uniformly integrable*, if

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

Lemma 8. *Let (Z_s) be a uniformly integrable family of random variables and $a_s = \mathsf{E}_s Z_s$. Let (Z_{si}) be a sequence of independent copies of Z_s. Then, for all $\epsilon > 0$,*

$$\sup_s \mathsf{P}_s\{\left|\bar{Z}_s - a_s\right| > \epsilon\} \xrightarrow[n\to\infty]{} 0.$$

Proof: Denote $b = \sup_s \mathsf{E}_s |Z_s|$. First we will show that $b < \infty$. Since (Z_s) is uniformly integrable, there exists $c \in [0; \infty)$ such that

$$\sup_s \mathsf{E}_s |Z_s| 1_{\{|Z_s| \geq c\}} \leq 1.$$

Then, for all s,

$$\mathsf{E}_s |Z_s| = \mathsf{E}_s |Z_s| 1_{\{|Z_s| < c\}} + \mathsf{E}_s |Z_s| 1_{\{|Z_s| \geq c\}} \leq c + 1.$$

Now fix ϵ, set $\delta = \frac{\epsilon^3}{8b}$ and define

$$U_{sni} = \begin{cases} Z_{si}, & \text{if } \left|Z_{si}\right| \leq \delta n, \\ 0, & \text{otherwise.} \end{cases}$$

Obviously, $(U_{sni} \mid i \geq 1)$ is a sequence of independent identically distributed random variables. Also

$$\begin{aligned} \mathsf{Var}_s U_{sn1} &\leq \mathsf{E}_s U_{sn1}^2 = \mathsf{E}_s Z_s^2 1_{\{|Z_s| \leq \delta n\}} \\ &\leq \delta n b. \end{aligned}$$

Denote $a_{sn} = \mathsf{E}_s U_{sn1}$. Then $a_{sn} = \mathsf{E}_s Z_s 1_{\{|Z_s| \leq \delta n\}}$. Note that

$$\begin{aligned} \left|a_{sn} - a_s\right| &= \left|\mathsf{E}_s U_{sn1} - \mathsf{E}_s Z_s\right| \\ &= \left|\mathsf{E}_s Z_s 1_{\{|Z_s| > n\delta\}}\right| \\ &\leq \mathsf{E}_s |Z_s| 1_{\{|Z_s| > n\delta\}} \\ &\leq \epsilon/2, \end{aligned}$$

for all $n \geq n_0$, where n_0 does not depend on s. Then, for all $n \geq n_0$,

$$\begin{aligned} \mathsf{P}_s\{\left|\bar{U}_{sn} - a_s\right| \geq \epsilon\} &\leq \mathsf{P}_s\{\left|\bar{U}_{sn} - a_{sn}\right| \geq \epsilon/2\} \\ &\leq \frac{4}{\epsilon^2} \mathsf{Var}_s \bar{U}_{sn} = \frac{4}{n\epsilon^2} \mathsf{Var}_s U_{sn1} \\ &\leq \frac{4\delta}{\epsilon^2} b = \frac{\epsilon}{2}. \end{aligned}$$

Also, by Chebyshev's inequality,

$$\begin{aligned} \mathsf{P}_s\{\bar{Z}_s \neq \bar{U}_{sn}\} &\leq \sum_{i=1}^n \mathsf{P}_s\{Z_{si} \neq U_{sni}\} = n \mathsf{P}_s\{|Z_s| > n\delta\} \\ &\leq \frac{1}{\delta} \sup_s \mathsf{E}_s |Z_s| 1_{\{|Z_s| > n\delta\}} < \epsilon/2, \end{aligned}$$

for $n \geq n_1$, where n_1 does not depend on s. Therefore, for $n \geq \max(n_0, n_1)$,

$$\begin{aligned} \mathsf{P}_s\{\left|\bar{Z}_s - a_s\right| \geq \epsilon\} &\leq \mathsf{P}_s\{\bar{Z}_s \neq \bar{U}_{sn}\} + \mathsf{P}_s\{\left|\bar{U}_{sn} - a_s\right| \geq \epsilon\} \\ &\leq \epsilon/2 + \epsilon/2 = \epsilon. \end{aligned}$$

■

In the next Section, we use this Lemma with $Z_{p_0} = m_{p_0}(X, Y)$, where p_0 is the true conditional probability. The family (Z_{p_0}) is uniformly integrable because

$$\begin{aligned} &\mathsf{E}_{p_0} |m_{p_0}(X,Y)| 1_{\{|m_{p_0}(X,Y)| \geq c\}} \\ &= \mathsf{E} |\log p_0(X)| 1_{\{|\log p_0(X)| \geq c\}} p_0(X) \\ &+ \mathsf{E} |\log(1 - p_0(X))| 1_{\{|\log(1-p_0(X))| \geq c\}} (1 - p_0(X)) \\ &\leq \epsilon, \end{aligned}$$

for $c \geq c_0$, where c_0 is such that $p|\log p| \leq \epsilon/2$ and $(1 - p)|\log(1-p)| \leq \epsilon/2$ for $p \leq \mathrm{e}^{-c_0}$.

Lemma 9. *Let $m(x, y)$ be a non-positive function, satisfying the following condition: for all $p_0 \in \bar{P}(E)$,*

$$\mathsf{E}_{p_0} m(X, Y) \leq a.$$

Then for all $q > a$ there exists $c > 0$ such that, for all $p_0 \in \bar{P}(E)$ and for all n,

$$\mathsf{P}_{p_0}\{\overline{m(X,Y)} \geq q\} \leq \mathrm{e}^{-cn}.$$

Proof: For $\lambda \geq 0$ denote

$$\begin{aligned} M_{p_0}(\lambda) &= \mathsf{E}_{p_0} \mathrm{e}^{\lambda m(X,Y)}, \\ \Lambda_{p_0}(\lambda) &= \log M_{p_0}(\lambda). \end{aligned}$$

These are known as the moment generating function and the cummulant generating function of $m(X, Y)$, respectively. We will use the following properties of the function $M_{p_0}(\lambda)$ (see [3], Section XIII.2):

1) $M_{p_0}(\lambda)$ is defined and continuous for $\lambda \geq 0$ and $M_{p_0}(0) = 1$.
2) $M_{p_0}(\lambda)$ is differentiable in $(0; \infty)$ and $M'_{p_0}(\lambda) = \mathsf{E}_{p_0} m(X,Y) \mathrm{e}^{\lambda m(X,Y)}$.
3) $M'_{p_0}(\lambda) \xrightarrow[\lambda\to 0]{} \mathsf{E}_{p_0} m(X, Y)$.

From these properties we get the following properties of the function $\Lambda_{p_0}(\lambda)$:

1) $\Lambda_{p_0}(\lambda)$ is defined and continuous for $\lambda \geq 0$ and $\Lambda_{p_0}(0) = 0$.
2) $\Lambda_{p_0}(\lambda)$ is differentiable in $(0; \infty)$ and $\Lambda'_{p_0}(\lambda) = \frac{M'_{p_0}(\lambda)}{M_{p_0}(\lambda)}$.
3) $\Lambda'_{p_0}(\lambda) \xrightarrow[\lambda\to 0]{} \mathsf{E}_{p_0} m(X, Y)$.

Step 1. Fix any $a < q' < q$. First we will prove that

$$\exists \lambda > 0 \; \forall p_0 \; \Lambda_{p_0}(\lambda) < \lambda q'. \tag{6}$$

Suppose the contrary. Then

$$\forall \lambda > 0 \; \exists p_0 \; \Lambda_{p_0}(\lambda) \geq \lambda q'.$$

Take $\lambda = 1/n$ and find p_{0n} such that

$$\Lambda_{p_{0n}}\left(\frac{1}{n}\right) \geq \frac{q'}{n}. \tag{7}$$

Since $p_{0n} \in \bar{P}(E)$ and $\bar{P}(E)$ is compact, without loss of generality, we can assume that $p_{0n}(X) \xrightarrow[n\to\infty]{} p_0(X)$ almost surely. Inequality (7) can be rewritten as

$$\frac{\Lambda_{p_{0n}}\left(\frac{1}{n}\right) - \Lambda_{p_{0n}}(0)}{1/n - 0} \geq q'. \tag{8}$$

By Lagrange's mean value theorem, the left hand side of (8) equals $\Lambda'_{p_{0n}}(\tilde{\lambda}_n)$, where $\tilde{\lambda}_n \in (0;1/n)$. Therefore

$$M'_{p_{0n}}(\tilde{\lambda}_n) = M_{p_{0n}}(\tilde{\lambda}_n)\Lambda_{p_{0n}}(\tilde{\lambda}_n) \geq q' M_{p_{0n}}(\tilde{\lambda}_n) \geq q'',$$

for n large enough, where $a < q'' < q'$, that is,

$$\mathsf{E}_{p_{0n}} m(X,Y) \mathrm{e}^{\tilde{\lambda}_n m(X,Y)} \geq q''. \tag{9}$$

Note that

$$\mathsf{E}_{p_{0n}} m(X,Y)\mathrm{e}^{\tilde{\lambda}_n m(X,Y)} = \mathsf{E}\mathsf{E}^X_{p_{0n}} m(X,Y)\mathrm{e}^{\tilde{\lambda}_n m(X,Y)},$$

where $\mathsf{E}^X_{p_{0n}}$ denotes the conditional expectation, given X, and

$$\begin{aligned}
&\mathsf{E}^X_{p_{0n}} m(X,Y)\mathrm{e}^{\tilde{\lambda}_n m(X,Y)} \\
&= m(X,0)(1-p_{0n}(X))\mathrm{e}^{\tilde{\lambda}_n m(X,0)} \\
&+ m(X,1)p_{0n}(X)\mathrm{e}^{\tilde{\lambda}_n m(X,1)} \\
&\xrightarrow[n\to\infty]{} m(X,0)(1-p_0(X)) + m(X,1)p_0(X) \\
&= \mathsf{E}^X_{p_0} m(X,Y).
\end{aligned}$$

Then by Fatou's lemma,

$$\begin{aligned}
&\mathsf{E}\varliminf_{n\to\infty} \mathsf{E}^X_{p_{0n}}(-m(X,Y)\mathrm{e}^{\tilde{\lambda}_n m(X,Y)}) \\
&\leq \varliminf_{n\to\infty} \mathsf{E}\mathsf{E}^X_{p_{0n}}(-m(X,Y)\mathrm{e}^{\tilde{\lambda}_n m(X,Y)}), \\
&\mathsf{E}\mathsf{E}^X_{p_0}(-m(X,Y)) \\
&\leq \varliminf_{n\to\infty} \mathsf{E}\mathsf{E}^X_{p_{0n}}(-m(X,Y)\mathrm{e}^{\tilde{\lambda}_n m(X,Y)}), \\
&\varlimsup_{n\to\infty} \mathsf{E}_{p_{0n}} m(X,Y)\mathrm{e}^{\tilde{\lambda}_n m(X,Y)} \\
&\leq \mathsf{E}_{p_0} m(X,Y) \leq a < q'',
\end{aligned}$$

a contradiction to (9).

Step 2. Find $\lambda > 0$ such that (6) holds and set $c = \lambda q - \lambda q'$. Then $c > 0$ and by Chebyshev's inequality,

$$\begin{aligned}
&\mathsf{P}_{p_0}\{\overline{m(X,Y)} \geq q\} \\
&= \mathsf{P}_{p_0}\{m(X_1,Y_1) + \cdots + m(X_n,Y_n) \geq nq\} \\
&= \mathsf{P}_{p_0}\{\mathrm{e}^{\lambda(m(X_1,Y_1)+\cdots+m(X_n,Y_n))} \geq \mathrm{e}^{\lambda nq}\} \\
&\leq \mathrm{e}^{-\lambda nq}\mathsf{E}_{p_0}\mathrm{e}^{\lambda(m(X_1,Y_1)+\cdots+m(X_n,Y_n))} \\
&\leq \left(\mathrm{e}^{-\lambda q}\mathsf{E}_{p_0}\mathrm{e}^{\lambda m(X,Y)}\right)^n \\
&= \left(\mathrm{e}^{-\lambda q + \Lambda_{p_0}(\lambda)}\right)^n \\
&\leq \left(\mathrm{e}^{-\lambda q + \lambda q'}\right)^n \\
&= \mathrm{e}^{-cn}.
\end{aligned}$$

■

In the next Section we will use this Lemma with $m(x,y) = m_{p,r}(x,y)$.

V. Uniform consistency

In this Section we prove our main result, Theorem 1. We precede the proof by two supporting Lemmas.

Lemma 10. *For all ϵ there exists ϵ' such that for all p_0, p with $d(p_0,p) \geq \epsilon$,*

$$M(p_0,p) < M(p_0,p_0) - \epsilon'. \tag{10}$$

Proof: Suppose the contrary. Then there exists ϵ such that

$$\forall \epsilon' \ \exists p_0 \ \exists p \ \left(d(p_0,p) \geq \epsilon, \ M(p_0,p) \geq M(p_0,p_0) - \epsilon'\right).$$

Take $\epsilon' = 1/n$ and find p_{0n} and p_n such that

$$d(p_{0n},p_n) \geq \epsilon, \ M(p_{0n},p_n) \geq M(p_{0n},p_{0n}) - 1/n. \tag{11}$$

Without loss of generality, we can suppose that $p_n \xrightarrow[n\to\infty]{} p$ and $p_{0n} \xrightarrow[n\to\infty]{} p_0$ almost everywhere. By Lemma 4 then $M(p_{0n},p_{0n}) \xrightarrow[n\to\infty]{} M(p_0,p_0)$. By (11), $d(p_0,p) \geq \varepsilon$. Moreover, by Lemma 3, Lemma 8 and (11),

$$M(p_0,p) \geq \varlimsup_{n\to\infty} M(p_{0n},p_n) \geq \lim_{n\to\infty} M(p_{0n},p_{0n}) = M(p_0,p_0).$$

This contradicts to Lemma 5. ■

Lemma 11. *Let the condition of Theorem 1 hold. Take any ϵ and let ϵ' be as in Lemma 10. Then for all p there exists r such that, for all p_0 with $d(p_0,p) \geq \epsilon$,*

$$M(p_0,p,r) \leq M(p_0,p_0) - \epsilon'.$$

Proof: Suppose the contrary. Then

$$\exists p \ \forall r \ \exists p_0 \ \left(d(p_0,p) \geq \epsilon, \ M(p_0,p,r) > M(p_0,p_0) - \epsilon'\right).$$

Find such p, take $r = 1/n$ and find p_{0n} such that

$$d(p_{0n},p) \geq \epsilon, \quad M(p_{0n},p,1/n) > M(p_{0n},p_{0n}) - \epsilon'. \tag{12}$$

Without loss of generality, we can suppose that $p_{0n} \xrightarrow[n\to\infty]{} p_0$. Metric is a continuous function, therefore, $d(p_{0n},p) \xrightarrow[n\to\infty]{} d(p_0,p)$. Also, by Lemma 4, $M(p_{0n},p_{0n}) \xrightarrow[n\to\infty]{} M(p_0,p_0)$. Moreover,

$$\begin{aligned}
M(p_{0n},p,1/n) &= \mathsf{E}_{p_{0n}} m_{p,1/n}(X,Y) \\
&= \mathsf{E}p_{0n}(X) \sup_{d(p',p)<1/n} \log p'(X) \\
&+ \mathsf{E}(1-p_{0n}(X)) \sup_{d(p',p)<1/n} \log(1-p'(X)).
\end{aligned}$$

By Fatou's lemma and Lemma 7,

$$\begin{aligned}
&\varlimsup_{n\to\infty} \mathsf{E}\big[p_{0n}(X) \sup_{d(p',p)<1/n} \log p'(X) \\
&+ (1-p_{0n}(X)) \sup_{d(p',p)<1/n} \log(1-p'(X))\big] \\
&\leq \mathsf{E}\varlimsup_{n\to\infty}\big[p_{0n}(X) \sup_{d(p',p)<1/n} \log p'(X) \\
&+ (1-p_{0n}(X)) \sup_{d(p',p)<1/n} \log(1-p'(X))\big] \\
&= \mathsf{E}\left[p_0(X)\log p(X) + (1-p_0(X))\log(1-p(X))\right] \\
&= M(p_0,p),
\end{aligned}$$

that is,

$$\varlimsup_{n\to\infty} M(p_{0n},p,1/n) \leq M(p_0,p). \tag{13}$$

If we calculate limits in (12), we get that

$$d(p_0,p) \geq \epsilon, \quad \varlimsup_{n\to\infty} M(p_{0n},p,1/n) \geq M(p_0,p_0) - \epsilon'.$$

But by (13) and Lemma 10,

$$\overline{\lim_{n\to\infty}} M(p_{0n}, p, 1/n) \le M(p_0, p) < M(p_0, p_0) - \epsilon',$$

a contradiction. ∎

Proof of Theorem 1: Fix some $\epsilon > 0$ and $\delta > 0$. Denote

$$B = \{(p_0, p) \in \bar{P}(E) \times \bar{P}(E) \mid d(p_0, p) \ge \epsilon\},$$

By Lemma 10, there exists ϵ' such that for all $(p_0, p) \in B$

$$M(p_0, p) < M(p_0, p_0) - \epsilon'.$$

Then by Lemma 11, for all p there exists $r(p)$ such that for all $(p_0, p) \in B$

$$M(p_0, p, r(p)) \le M(p_0, p_0) - \epsilon'.$$

The set B is compact and is covered by sets $A(p) = \{(p_0, p') \mid d(p', p) < r(p)\}$. These sets are open because metric d is continuous w.r.t. p' and $r(p)$ does not depend on p_0. Therefore, there exist $p_1, \dots, p_k$ such that

$$B \subset A(p_1) \cup \dots \cup A(p_k).$$

Choose n_0 such that, for all $n \ge n_0$ and for $j = 1, \dots, k$,

$$\sup_{p_0 \in \bar{P}(E)} \mathsf{P}_{p_0}\{M_n(p_0) \le M(p_0, p_0) - \epsilon'/2\} \le \frac{\delta}{2k}$$

and

$$\sup_{p_0 \in \bar{P}(E)} \mathsf{P}_{p_0}\{M_n(p_j, r(p_j)) \ge M(p_0, p_0) - \epsilon'/2\} \le \frac{\delta}{2k}.$$

Fix p_0 and $n \ge n_0$ and set $q = M(p_0, p_0) - \varepsilon'/2$. Then

$$\begin{aligned} &\mathsf{P}_{p_0}\{M_n(p_0) \le M_n(p_j, r(p_j))\} \\ &\le \mathsf{P}_{p_0}\{M_n(p_0) \le q\} + \mathsf{P}_{p_0}\{M_n(p_j, r(p_j)) \ge q\} \\ &\le \delta/k. \end{aligned}$$

If $(p_0, \hat{p}) \in A(p_j)$, then

$$M_n(p_0) \le M_n(\hat{p}) \le M_n(p_j, r(p_j)),$$

where the first inequality holds because $\hat{p}$ is chosen to maximize $M_n(p)$ and the second inequality holds because $d(\hat{p}, p_j) < r(p_j)$. Therefore,

$$\begin{aligned} &\mathsf{P}_{p_0}\{d(p_0, \hat{p}) \ge \epsilon\} = \mathsf{P}_{p_0}\{(p_0, \hat{p}) \in B\} \\ &\le \sum_{j=1}^{k} \mathsf{P}_{p_0}\{(p_0, \hat{p}) \in A(p_j)\} \\ &\le \sum_{j=1}^{k} \mathsf{P}_{p_0}\{M_n(p_0) \le M_n(p_j, r(p_j))\} \\ &\le k\delta/k = \delta. \end{aligned}$$

∎

VI. Discussion

We proved that the logistic classifier is uniformly consistent, provided that the distribution μ of X satisfies condition (2). A natural question arises, how restrictive is that condition. We can prove that in the case $E = \mathbb{R}^k$ condition (2) holds, if μ has a density with respect to the Lebesgue's measure. In such a case, $\mu(A) = 0$ with any A of zero Lebesgue's measure.

Since Lebesgue's measure of any hyperplane is 0, there are essentially only two types of functions $p \in \bar{P}(E)$: functions p_θ and functions $p_{a,0}$. If $p = p_\theta$, then there is no other function $p' \in \bar{P}(E)$ equal to p almost everywhere: if $p' = p_{\theta'}$ with $\theta \ne \theta'$,

$$\mu\{p = p'\} = \mu\{x \mid \theta x = \theta' x\} = 0,$$

and if $p' = p_{a,0}$, then $\mu\{p = p'\} = 0$ because $p(x) \in (0;1)$ and $p'(x)$ is equal to 0 or 1 for almost all x.

Now let $p = p_{a,0}$. If p' is equal to p almost everywhere, then p' is equal to $p_{a',0}$ almost everywhere. Denote

$$\begin{aligned} C_+(a') &= \{x \mid ax \ge 0,\ a'x \ge 0\}, \\ C_-(a') &= \{x \mid ax \le 0,\ a'x \le 0\}, \\ C_+ &= \bigcap_{d(p_{a',0}, p) = 0} C_+(a'), \\ C_- &= \bigcap_{d(p_{a',0}, p) = 0} C_-(a') \end{aligned}$$

and

$$C = \bigcap_{d(p_{a',0}, p) = 0} \left(C_+(a') \cup C_-(a')\right).$$

Every intersected set is closed and covers $\operatorname{supp}\mu$. Therefore, $C \supset \operatorname{supp}\mu$ and $\mu(C) = 1$. On the other hand, C is the union of three convex sets: C_+, C_- and $\{x \mid ax = 0\}$. The boundary of a convex set is of Lebesgue's measure 0. Therefore, if C° is the interior of C, then $\mu(C^\circ) = 1$. If $x \in C^\circ$, then with any a', such that $d(p_{a',0}, p) = 0$, one of the following holds: $ax < 0$, $a'x < 0$ or $ax > 0$, $a'x > 0$. In both cases, $p_{a',0}(x) = p_{a,0}(x)$.

The simplest example, where condition (2) does not hold, is the following. Let $E = \mathbb{R}^3$ and μ be the distribution of the random vector $X = (1, \cos T, \sin T)$, where T is distributed uniformly in $[0; 2\pi]$ (see Figure 1). Let $p = p_{a,0}$, where

$$ax = -x_1 \text{ for } x \in \mathbb{R}^3.$$

In other words,

$$p(x) = \begin{cases} 0, & \text{if } x_1 > 0, \\ 1/2, & \text{if } x_1 = 0, \\ 1, & \text{if } x_1 < 0. \end{cases}$$

It is obvious that almost surely $p(X) = 0$.

For all $t \in [0; 2\pi]$ let $p'_t = p_{a_t,0}$, where

$$a_t x = (-x_1 + x_2 \cos t + x_3 \sin t)/\sqrt{2}.$$

In other words,

$$p'_t(x) = \begin{cases} 0, & \text{if } -x_1 + x_2 \cos t + x_3 \sin t < 0, \\ 1/2, & \text{if } -x_1 + x_2 \cos t + x_3 \sin t = 0, \\ 1, & \text{if } -x_1 + x_2 \cos t + x_3 \sin t > 0. \end{cases}$$

Let $x_s = (1, \cos s, \sin s)$ for $s \in [0; 2\pi]$. If $s \neq t$, then $p'_t(x_s) = 0$ because

$$-1 + \cos s \cos t + \sin s \sin t = -1 + \cos(s - t) < 0.$$

Therefore, almost surely $p'_t(X) = p(X)$. On the other hand, $p'_t(x_t) = \frac{1}{2}$ because

$$-1 + \cos^2 t + \sin^2 t = 0.$$

Therefore,

$$p'_t(x_t) = 1/2 \neq p(x_t).$$

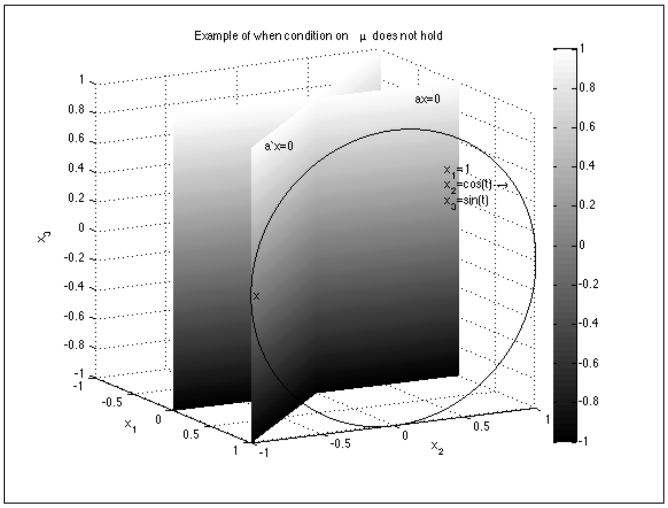

Fig. 1. Example of the situation, where condition on μ does not hold. At the point x in the plot, $p'_t(x) \neq p(x)$.

To sum up, we could not prove uniform consistency of a logistic classifier, without putting any restrictions on the distribution μ of X. Given that in our previous work (see [4]) we have proved the uniform consistency of histogram type classifiers and no restrictions were needed, an open question remains, whether a class of logistic classifiers is not uniformly consistent, in general, or its uniform consistency can be proved in some other way.

ACKNOWLEDGMENT

This work was supported by the Singapore Bioimaging Consortium, Agency for Science, Technology and Research, Singapore and Department of Statistical Science, University College London, United Kingdom.

REFERENCES

[1] A. Dembo, O. Zeitouni, *Large deviations techniques and applications*, Springer, 1998.

[2] R.M. Dudley, *Real analysis and probability*, Cambridge University Press, 2002.

[3] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 2, Wiley, 1971.

[4] A. Kazakeviciute, V. Kazakevicius, M. Olivo, *Conditions for existence of uniformly consistent classifiers*, unpublished.

[5] V.V. Petrov, *Limit theorems of probability theory: sequences of independent random variables*, Clarendon Press, Oxford, 1995.

[6] M.L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley, 2005.

[7] A.W. van der Vaart, *Asymptotic Statistics*, Cambridge University Press, 2000.

Agne Kazakeviciute received a Bsc degree in Statistics from Vilnius University, Lithuania, in 2012 and a MSc degree in Medical Statistics from University College London, United Kingdom, in 2013. She is currently pursuing her PhD in Statistical Science at University College London under the joint UCL-A*STAR research scholarship. She is particularly interested in statistical and machine learning pattern recognition and functional data analysis. On top of this, her research interests include high-dimensional image processing.

Malini Olivo is currently Head of Bio-Optical Imaging Group, Singapore Bioimaging Consortium, A*STAR, Singapore. She holds an adjunct Professorship at the National University of Ireland, Royal College of Surgeons, Ireland and is also a visiting Professor at Harvard Medical school. She obtained a Ph.D. in Bio-Medical Physics in 1990 from University Malaya/University College London, and did her post-doctoral training between 1991-1995 in University College London, UK and both McMaster University and University of Toronto, Canada. She was a Principal Investigator at the Singapore National Cancer Centre and Singhealth from 1996 to 2009, heading biophotonics research for clinical translation in cancer. In 2009, she took a Stokes Professorship Science Foundation Ireland Chair in the National University of Ireland and returned to Singapore in 2012 to head bio-optical imaging in SBIC, A*STAR. She is recognized as a pioneer in the area of clinical applications of photomedicine in optical diagnostics and therapeutics in Singapore. Her current research interests are in nano-biophotonics and its applications in translational medicine. Malini Olivo is well recognized internationally in her field and serves in numerous scientific advisory boards in the area of Photonics in Medicine. She currently serves in the EU research commission in Photonics 21 shaping research in photonics for Life sciences till 2015.