Persistence of plug-in rule in classification of high dimensional multivariate binary data

Junyong Park[a,*], Jayanta K. Ghosh[a,b]

[a]Department of Mathematics and Statistics, University of Maryland, Baltimore Country, MD 21250, USA
[b]Division of Theoretical Statistics and Mathematics, Indian Statistical Institute, India

Available online 30 March 2007

Abstract

In this paper, we consider the classification problem when the predictors are multivariate binary random variables. Variables are modeled as independent, but not necessarily identical, Bernoulli. A triangular array for parameters, $(p_{11}^{(n)}, \ldots, p_{1d}^{(n)}, p_{21}^{(n)}, \ldots, p_{2d}^{(n)})$, is assumed to allow parameters to change and the number of the variables, d, to increase for adopting more flexible models as the sample size, n, increases. Our results are obtained under moderate assumptions on the triangular array of the probability vectors. We use maximum likelihood estimators for the parameters and plug them into the Bayes classifier. This is a plug-in classifier, a sort of objective Bayes rule. It is shown in Wilbur et al. [2002. Variable selection in high-dimensional multivariate binary data with application to the analysis of microbial DNA fingerprints. Biometrics 58, 378–386] via simulations that the plug-in rule classifies quite well even when the assumption of independence is violated. The main interest in this paper is in the complex case of $d/n^{\nu} \to c$ for some $\nu > 0$ and $c > 0$ for which very little is known. Using linearity of the plug-in rule, we show its persistence, a generalization of the notion of consistency, when the variance of the plug-in rule or a quantity measuring signal to noise ratio is divergent; otherwise we show there exists an example of non-persistence of the plug-in rule. In case of non-persistence, we introduce the notion of sparsity and overcome non-persistence by selecting a subset of the variables. This shows why a variable selection procedure may be effective especially for contemporary practical problems with high dimensional data [Wilbur et al., 2002. Variable selection in high-dimensional multivariate binary data with application to the analysis of microbial DNA fingerprints. Biometrics 58, 378–386].

Keywords: Persistence; Triangular array; High dimensional multivariate binary data; Plug-in rule; Sparsity

1. Introduction

In classical asymptotics, the model and the dimension of the parameter space are held fixed while the sample size n tends to infinity. If the true value of the parameter is also held fixed, one may wish to know if an estimate is consistent, i.e., converges to the true value in some sense. However, in many cases it may be more realistic to assume the model becomes more flexible and so increasingly complex as the sample size increases. To realize such a situation one may consider a triangular array of d-dimensional random vectors $(X_{i1}, \ldots, X_{id})$, $i = 1, \ldots, n$ with d depending on n. In this situation, the true value would also depend on n. Some of the most significant results in the context of linear regression or exponential families are Huber (1973), Portnoy (1984, 1985) and Greenshtein and Ritov (2004). Sufficient conditions for zero misclassification probability in Bayesian discrimination with infinitely many normal or Bernoulli random variables is reported in Dawid and Fang (1992) and Fang and Dawid (1993). High dimensional discrimination is also considered in Ge and Simpson (1998), but it is considered that the Bayes error converges to zero. Recently, Bickel and Levina (2004) consider a situation where the dimension d is bigger than the sample size n and investigate the case that the Bayes error does not tends to zero. Greenshtein and Ritov (2004) consider larger number of variables than the sample size, but showed the order of best subset of variables is at most $o(n/\log n)$.

Independence is a strong assumption but there are both practical and theoretical reasons for making this assumption. According to our microbiologist colleagues, assuming independence is a realistic assumption in microbial fingerprint analysis (Wilbur et al., 2002). Moreover, simulations in Wilbur et al. (2002) show that even under dependence the linear classifier obtained under independence classifies well. Further experimental support comes from bacterial taxonomy (Gyllenberg and Koski, 2001) and medical diagnosis. Choosing independence may also be thought of as an application of parsimony in model selection when we sometimes prefer a simple but false model to the true complex model.

In problems of this kind, the Bayes rule is a linear classifier if the true probability model is known. In Devroye et al. (1996), it is shown that the empirical risk minimizer among linear rules attains the Bayes error asymptotically if $d = o(n/\log n)$.

There are two main issues in this paper:

(1) we investigate asymptotic behavior of the plug-in rule, i.e., its persistence or lack of it in classification with multivariate binary variables,
(2) we justify in asymptotic sense why variable selection procedure is effective in classification of high dimensional multivariate binary data under sparsity condition.

To describe in detail different possible scenarios for the plug-in rule, we introduce moderate conditions on the parameter space and study different sets of sufficient conditions for persistence. In Section 2, we introduce basic notations and define persistence. In Section 3, we introduce our moderate condition on the parameter space and discuss its motivation. In Section 4, we introduce a sparsity condition under which we show how selection of a good subset may overcome non-persistence shown in Section 3. The proofs of main results depend on several lemmas, which are somewhat delicate because one has to compare sums of different functions of parameters of Bernoulli, p_{1i} and p_{2i} (e.g., Lemmas 3.2, 3.3 and A.3 in the Appendix). This is done through the introduction of relative orders of magnitude of partitions, which need to be somewhat different in different contexts. The partitions act like sieves.

2. Multivariate binary data and notations

In Wilbur et al. (2002), the number of variables are $d = 84$ and the sample size is $n = 89$. The fact that d and n are of the same magnitude is typical of many contemporary problems. In some cases, d exceeds n. Since d increases with n, these problems are high dimensional. All such problems are difficult but the cases where $d \geqslant n$ or at least of the same order of magnitude are the most difficult. In some studies as in Wilbur et al. (2002), new methodologies are required. In others, as in the present asymptotic study, we need a new formulation of optimality, namely, a notion of persistence introduced next.

Multivariate binary data are common in several applications arising in agriculture, social sciences and medical diagnosis. In bacterial taxonomy, Gyllenberg and Koski (2001) discuss identification of new bacteria on the basis of many tests, each of which results in a binary output of yes or no. Hand (1981) discusses a similar example for identifying people likely to be suffering from non-psychotic psychiatric illnesses on the basis of binary responses to the General Health Questionnaire (GHQ).

We consider below an example from agricultural microbiology in Wilbur et al. (2002). Plots are placed under four treatments formed by combining rotation (present or absent) and tilling (present or absent). The crop grown was corn. As expected, the four treatments are well-separated by crop yields. To get some microbiological insight about the treatments, soil samples were taken from each of the plots and subjected to DNA analysis. Based on the analysis, for each plot one knows which of d bacteria are present and which are absent in each soil sample. The important question was whether using the full set of d binary variables or a subset, one can provide a good classification of plot into four classes, thus identifying the treatment corresponding to a plot. In this analysis, the yields of the crop is ignored and the primary objective is correspondence between the treatments and values of the selected binary variables. The set of variables which classify with least error identifies the important bacteria.

Since the number of variables and hence the number of parameters grows with n, we consider a triangular array of parameters and data. The data are binary and modeled as Bernoulli random variables. The study in Wilbur et al. (2002) is based on assumption of independence, which was considered reasonable by the associated microbiologists. In many practical studies, the independence assumption is a reasonable approximation, especially in high dimensions. See this connection in Hand and Yu (2001).

Since we consider triangular arrays, for each class $j = 1, 2$, the parameters of Bernoulli variables are denoted as $\theta_j^{(n)} = (p_{j1}^{(n)}, p_{j2}^{(n)}, \ldots, p_{jd}^{(n)})$, $j = 1$ and $d = d_n$ depends on n, e.g.,

$$\begin{aligned}
&\theta_1^{(1)} = (p_{11}^{(1)}),\\
&\theta_1^{(2)} = (p_{11}^{(2)}, p_{12}^{(2)}),\\
&\vdots\\
&\theta_1^{(n)} = (p_{11}^{(n)}, p_{12}^{(n)}, \ldots, p_{1d}^{(n)}).
\end{aligned}$$

Remark. For notational convenience, we omit (n) in the parameters, i.e., we use p_{ji} instead of $p_{ji}^{(n)}$ and $d = d_n$.

We assume uniform prior for classes, $P(j = 1) = P(j = 2) = \frac{1}{2}$, which is the usual choice for classification problems unless other information is available. Below $j = 2$ and -1 denote the same class. Then, the Bayes classifier is

$$g_0(X) \equiv \begin{cases} 1 & \text{if } \prod_{i=1}^d p_{1i}^{X_i}(1-p_{1i})^{1-X_i} > \prod_{i=1}^d p_{2i}^{X_i}(1-p_{2i})^{1-X_i},\\ -1 & \text{otherwise,}\end{cases}$$

where $X = (X_1, \ldots, X_d)$. Taking logarithms, the Bayes classifier becomes linear. In other words,

$$g_0(X) \equiv \begin{cases} 1 & \text{if } \delta_d(X) > 0,\\ -1 & \text{otherwise,}\end{cases}$$

where $\delta_d(X) = \sum_{i=1}^d (c_i X_i + c_{i0})$ and

$$c_i = \log\left(\frac{p_{1i}}{p_{2i}}\frac{1-p_{2i}}{1-p_{1i}}\right), \quad c_{i0} = \log\left(\frac{1-p_{1i}}{1-p_{2i}}\right) \quad 1 \leqslant i \leqslant d.$$

The plug-in rule $g_{\hat{\theta}}$ substitutes the mle $\hat{\theta}$ based on observed data for the unknown θ in $g_{\hat{\theta}}$. The plug-in rule, $g_{\hat{\theta}}(X)$, is $\operatorname{sgn}(\hat{\delta}_d(X))$ where $\hat{\delta}_d(X) = \sum_{i=1}^d (\hat{c}_i X_i + \hat{c}_{i0})$. To avoid the difficulty of having $\log 0$ in $\hat{c}_i$, we use

$$\hat{c}_i = \log\left(\frac{\hat{p}_{1i} + \dfrac{1}{n^2}}{\hat{p}_{2i} + \dfrac{1}{n^2}}\frac{1-\hat{p}_{2i} + \dfrac{1}{n^2}}{1-\hat{p}_{1i} + \dfrac{1}{n^2}}\right), \quad \hat{c}_0 = \sum_{i=1}^d \log\left(\frac{1-\hat{p}_{1i} + \dfrac{1}{n^2}}{1-\hat{p}_{2i} + \dfrac{1}{n^2}}\right), \tag{1}$$

where $\hat{p}_{ji} = \bar{x}_{ji} = (1/n)\sum_{k=1}^n x_{ji}^k$ where x_{ji}^k is kth observation of ith variable in jth class for $1 \leqslant i \leqslant d$, $j = 1, 2$ and $1 \leqslant k \leqslant n$. The performance of $g_{\hat{\theta}}$ in relation to g_0 depends on how well $\hat{\theta}$ estimates θ. The Bayes error is $L_d^* = \frac{1}{2}(P(\delta_d(X) < 0 | j = 1) + P(\delta_d(X) > 0 | j = 2))$ and error of the plug-in rule is $L_{n,d} = \frac{1}{2}(P(\hat{\delta}_d(X) < 0 | j = 1) + P(\hat{\delta}_d(X) > 0 | j = 2))$.

For a triangular array, where the true parameters change with n, an appropriate modification of the notion of consistency is persistence. Following Greenshtein and Ritov (2004), we define persistence as follows.

Definition 2.1 (*Persistence*). If the parameters form a triangular array, a classification rule is said to be persistence if $L_{n,d} - L_d^* \to 0$.

We assume the true class is $j=1$ since the other case can be handled in the same way. So, for example, the Bayes error is $P(\delta_d(X)<0|j=1)$, but we omit $j=1$ for notational convenience, we write $P(\delta_d(X)<0|j=1)$ as $P(\delta_d(X)<0)$. In a similar way, expectation, $E(\cdot)$, is conditional expectation given $j=1$. For example, $E(\delta_d(X)) \equiv E(\delta_d(X)|j=1) = \sum_{i=1}^d K(p_{1i}, p_{2i})$ and $\mathrm{var}(\delta_d(X)) \equiv \mathrm{var}(\delta_d(X)|j=1) = \sum_{i=1}^d c_i^2 p_{1i}(1-p_{1i})$ where $K(p,q)=p\log\frac{p}{q}+(1-p)\log\frac{1-p}{1-q}$ and $c_i = \log\frac{p_{1i}}{p_{2i}}\frac{1-p_{2i}}{1-p_{1i}}$.

3. Persistence and non-persistence of plug-in rule

As mentioned earlier, the performance of $g_{\hat{\theta}}$ depends on how accurately θ is estimated. Under the assumption of the true parameters with fixed dimension, one can estimate the classifier accurately as the sample size increases. So the plug-in rule can achieve Bayes error asymptotically. But, a triangular array with increasing dimension causes two difficulties:

(1) the number of variables increases with the sample size,
(2) the parameter θ changes with the sample size.

Since the number of parameters increases, the accumulation of inaccuracy does not guarantee improvement of the estimated classifier. The problem becomes particularly difficult if the number of variables is almost the same as the sample size, $d \asymp n$ also, as indicated in the previous section, in many practical problems this condition holds. Under this condition, we show that the plug-in rule may be persistent or not persistent depending on other conditions. More generally, we present similar results when $d \asymp n^{\nu}$ for $\nu > 0$.

A second difficulty occurs when parameters, p_{ji}, are close to 0. From an empirical point of view, estimation of such p_{ji} leads to difficulties since sampling fluctuations can easily inflate the differences between p_{1i} and p_{2i}. In particular, the coefficient of variation $\frac{\sqrt{p_{ji}(1-p_{ji})}}{p_{ji}}$ can be large if p_{ji} is small. If p_{ji} is close to one, the same problem appears for $1-X_{ji}$. In our experience, data tend to give less importance to such variables.

Also, since the plug-in rule, $\hat{\delta}_d(X)$, includes $\hat{c}_i$ and $\hat{c}_{i0}$, we want to consider the first order or the second order approximation at the neighborhood of the true p_{ji}. In other words, using this approximation, we would like to claim that $\hat{c}_i - c_i = (\frac{\hat{p}_{1i}-p_{1i}}{p_{1i}(1-p_{1i})} - \frac{\hat{p}_{2i}-p_{2i}}{p_{2i}(1-p_{2i})})(1+\mathrm{o_p}(1))$ and $\hat{c}_i - c_i \to 0$ in probability as $n \to \infty$. But, if those parameters converge to 0 or 1 very fast, Taylor expansion may be useless. For example, suppose $X_{ji}^k \sim \mathrm{Bernoulli}(p^{(n)})$, $1 \leqslant k \leqslant n$ and $p^{(n)} = \frac{1}{n}$, then

$$
\begin{aligned}
P\left(\frac{|\hat{p}_{ji} - \frac{1}{n}|}{\frac{1}{n}\left(1-\frac{1}{n}\right)} > \varepsilon\right) &> P\left(\hat{p}_{ji} > \frac{1}{n}\left(1-\frac{1}{n}\right)(1+\varepsilon)\right) \\
&= 1 - P\left(\sum_{k=1}^{n} X_{ji}^k \leqslant \left(1-\frac{1}{n}\right)(1+\varepsilon)\right) \\
&= 1 - \binom{n}{0}\left(\frac{n-1}{n}\right)^n - \binom{n}{1}\left(\frac{1}{n}\right)\left(\frac{n-1}{n}\right)^{n-1} \\
&\to 1 - \frac{2}{\mathrm{e}} > 0.
\end{aligned}
$$

To avoid this, we restrict the range of p_{ji} such that $n^{-\beta} < p_{ji} < 1 - n^{-\beta}$ for $0<\beta<1$. Then, it can be shown that $\hat{c}_i - c_i$ converges to 0 in probability through $\frac{\hat{p}_{ji}-p_{ji}}{p_{ji}(1-p_{ji})}$ for $j=1,2$.

Let $A_d = A_d^{(1)} \cap A_d^{(2)}$ and $A_d^{(j)} = \bigcap_{i=1}^d A_{ji}$ for $j=1,2$ where $A_{ji} = \{|\frac{\hat{p}_{ji}-p_{ji}}{p_{ji}(1-p_{ji})}| \leqslant n^{-\varepsilon^*}\}$ for $1-\beta-2\varepsilon^* > 0$. On each $A_{1i} \cap A_{2i}$, we may consider Taylor expansion for $\hat{c}_i$ and $\hat{c}_{i0}$. It is enough to consider our expansions on A_d since $P(A_d^c)$ is negligible even when d increases. The following lemma clarifies this idea.

Lemma 3.1. *If $d = n^{v}$ for some $v > 0$ and $0 < \beta < 1$, $n^k P(A_d^c) \to 0$ for all integer $k \geqslant 0$.*

Proof. From the definition of A_d, $P(A_d^c) = P(\bigcup_{i=1}^{d}(A_{1i}^c \cup A_{2id}^c)) \leqslant \sum_{i=1}^{d}(P(A_{1i}^c) + P(A_{2i}^c))$ where $A_{ji}^c = \{|\frac{\hat{p}_{ji} - p_{ji}}{p_{ji}(1-p_{ji})}| > n^{-\varepsilon^*}\} = \{|\hat{p}_{ji} - p_{ji}| > p_{ji}(1-p_{ji})n^{-\varepsilon^*}\}$. We only need to show the case $j = 1$. Let $X_{1i}^k \sim$ Bernoulli(p_{1i}) for $1 \leqslant k \leqslant n$. Then, for $\varepsilon = \frac{1}{n^{\varepsilon^*}}$, using Lemma A.1 in the Appendix,

$$\begin{aligned} P\left(|\hat{p}_{1i} - p_{1i}| > \frac{p_{1i}(1-p_{1i})}{n^{\varepsilon^*}}\right) &\leqslant P\left(\left|\sum_{k=1}^{n}(X_{1i}^k - p_{1i})\right| > np_{1i}(1-p_{1i})\frac{1}{n^{\varepsilon^*}}\right) \\ &\leqslant 2\exp\left\{-3np_{1i}(1-p_{1i})\frac{1}{8n^{2\varepsilon^*}}\right\} \\ &\leqslant 2\exp\left\{-\frac{3}{8}\frac{nn^{-\beta}}{n^{2\varepsilon^*}}\right\} \\ &= 2\exp\left\{-\frac{3}{8}n^{1-\beta-2\varepsilon^*}\right\}. \end{aligned}$$

In the same way, we deal with $j = 2$. Hence, $n^k P(A_d^c) \leqslant \sum_{i=1}^{d} n^k(P(A_{1i}^c) + P(A_{2i}^c)) \leqslant 2 \cdot n^k \cdot d \cdot 2\exp\{-\frac{3}{8}n^{1-\beta-\varepsilon^*}\} \to 0$. □

Based on the above discussion, we set the baseline conditions and call them *Condition A*.

Condition A. (1) $\frac{d}{n^{v}} \to c$ for some $v > 0$ and $0 < c < \infty$.
(2) $n^{-\beta} < p_{ji} < 1 - n^{-\beta}$ for $0 < \beta < 1$.

Under *Condition A*, we will show that if $E\delta_d(X)/\sqrt{\operatorname{var}(\delta_d(X))}$ diverges, then the plug-in rule is persistent. But, if $E\delta_d(X)/\sqrt{\operatorname{var}(\delta_d(X))}$ is bounded, then persistence or non-persistence of plug-in rule depends on the behavior of $\operatorname{var}(\delta_d(X))$. The criterion $E\delta_d(X)/\sqrt{\operatorname{var}(\delta_d(X))}$ acts somewhat like a signal to noise ratio in the context of classification. See the discussion in Section 3.1.

The following lemma shows the critical role played by the unboundedness of $\operatorname{var}(\delta_d(X))/n^{v-1}$. The phenomenon of persistence and non-persistence are determined to some extent by the following lemma. Additionally, this lemma is used in the proof of Lemma 3.3.

In several lemmas, starting with Lemma 3.2, we need the following sets. For a given $\varepsilon > 0$ and $l, m = 1, 2, \ldots, [\frac{\beta}{\varepsilon}], \frac{\beta}{\varepsilon}$,

$$\begin{aligned} B_{1l} &= \{i | n^{-\beta+(l-1)\varepsilon} \leqslant p_{1i}(1-p_{1i}) < n^{-\beta+l\varepsilon}\}, \\ B_{2m} &= \{i | n^{-\beta+(m-1)\varepsilon} \leqslant p_{2i}(1-p_{2i}) < n^{-\beta+m\varepsilon}\}, \\ D &= \{i | |c_i| \leqslant c^*\}, \end{aligned}$$

where $c^* > 0$ is arbitrary. The sets B_{1l} and B_{2m} forms a suitable finite partition of $p_{1i}(1-p_{1i})$ and $p_{2i}(1-p_{2i})$ such that in each $B_{1l} \cap B_{2m}$ the order of magnitude of $\frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})}$ is estimated accurately enough. This fact is used repeatedly in the proof of lemmas. In some cases, e.g., in the proof of Lemma A.3 in the Appendix, B_{1l} and B_{2m} are a similar partition but in terms of p_{1i} and p_{2i}. One may think of such sets as sieves.

Lemma 3.2. *Under Condition A*,

(1) *If* $\operatorname{var}(\delta_d(X))/n^{v-1} \to \infty$, *then* $\frac{1}{n\operatorname{var}(\delta_d(X))}\sum_{i=1}^{d}\frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} = o(1)$.
(2) *If* $\operatorname{var}(\delta_d(X))/n^{v-1}$ *is bounded, then* $\frac{1}{n}\sum_{i=1}^{d}\frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} = cn^{v-1}(1 + o(1))$.

Proof. (1) On D, $|c_i| \leqslant c^*$ implies $\mathrm{e}^{-c^*} < \frac{p_{1i}}{1-p_{1i}} \frac{1-p_{2i}}{p_{2i}} < \mathrm{e}^{c^*}$. From $\frac{p_{1i}}{1-p_{1i}} \frac{1-p_{2i}}{p_{2i}} < \mathrm{e}^{c^*}$, $\frac{p_{1i}}{p_{2i}} \leqslant \frac{\mathrm{e}^{c^*}}{1+p_{2i}(\mathrm{e}^{c^*}-1)} \leqslant \mathrm{e}^{c^*}$ since $1 + p_{2i}(\mathrm{e}^{c^*} - 1) > 1$. In the same way, $\frac{1-p_{1i}}{1-p_{2i}}$. From $\mathrm{e}^{-c^*} < \frac{p_{1i}}{1-p_{1i}} \frac{1-p_{2i}}{p_{2i}}$, we derive $\frac{1-p_{1i}}{1-p_{2i}} < \mathrm{e}^{c^*}$. Therefore, $\frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} \leqslant \mathrm{e}^{2c^*}$.

On $D^c = \bigcup_{l,m} (D^c \cap B_{1l} \cap B_{2m})$, we consider each $D^c \cap B_{1l} \cap B_{2m}$ for $l, m = 1, \ldots, [\frac{\beta}{\varepsilon}]$, $\frac{\beta}{\varepsilon}$ and denote the cardinality of the sets by $|D^c \cap B_{1l} \cap B_{2m}|$. Then,

$$\frac{1}{\mathrm{var}(\delta_d(X))} \leqslant \frac{1}{(c^*)^2 n^{-\beta+(l-1)\varepsilon} |D^c \cap B_{1l} \cap B_{2m}|}, \tag{2}$$

$$\sum_{i \in D^c \cap B_{1l} \cap B_{2m}} \frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} \leqslant \sum_{i \in D^c \cap B_{1l} \cap B_{2m}} \frac{n^{-\beta+l\varepsilon}}{n^{-\beta+(m-1)\varepsilon}} \tag{3}$$

$$= |D^c \cap B_{1l} \cap B_{2m}| n^{(l-m+1)\varepsilon}. \tag{4}$$

From these facts,

$$\begin{aligned}
\frac{1}{n\,\mathrm{var}(\delta_d(X))} \sum_{i=1}^{d} \frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} &= \frac{1}{n\,\mathrm{var}(\delta_d(X))} \left(\sum_{i \in D} \frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} + \sum_{l,m} \sum_{i \in D^c \cap B_{1l} \cap B_{2m}} \frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} \right) \\
&\leqslant \frac{\mathrm{e}^{2c^*} d}{n\,\mathrm{var}(\delta_d(X))} + \frac{1}{n} \sum_{l,m} \frac{|D^c \cap B_{1l} \cap B_{2m}| n^{(l-m+1)\varepsilon}}{(c^*)^2 n^{-\beta+(l-1)\varepsilon} |D^c \cap B_{1l} \cap B_{2m}|} \quad \text{by (2) and (4)} \\
&\leqslant \frac{\mathrm{e}^{2c^*} d}{n\,\mathrm{var}(\delta_d(X))} + \sum_{l,m} \frac{1}{(c^*)^2} \frac{1}{n^{1-\beta+(m-2)\varepsilon}} \\
&\leqslant \frac{\mathrm{e}^{2c^*} c n^{\nu-1}(1+\mathrm{o}(1))}{\mathrm{var}(\delta_d(X))} + \left(\left[\frac{\beta}{\varepsilon} \right] + 1 \right)^2 \frac{1}{(c^*)^2} \frac{1}{n^{1-\beta-\varepsilon}} \\
&\to 0.
\end{aligned}$$

(2) We will show that if $\mathrm{var}(\delta_d(X))/n^{\nu-1}$ is bounded, then $|D^c| = \mathrm{o}(d)$ and $|D| = d + \mathrm{o}(d)$. If $|D^c| = \mathrm{O}(d)$, then $\mathrm{var}(\delta_d(X))/n^{\nu-1} = \sum_{i \in D^c} c_i^2 p_{1i}(1-p_{1i})/n^{\nu-1} \geqslant (c^*)^2 n^{-\beta+(l-1)\varepsilon} |D|/n^{\nu-1} \asymp n^{1-\beta+(l-1)\varepsilon} \to \infty$. This is a contradiction. So $|D^c| = \mathrm{o}(d)$ and $|D| = d + \mathrm{o}(d)$.

On D^c,

$$\begin{aligned}
\frac{1}{n} \sum_{i \in D^c} \frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} &= \frac{1}{n} \sum_{l,m} \sum_{i \in D^c \cap B_{1l} \cap B_{2m}} \frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} \\
&= \frac{1}{n} \sum_{l,m} \sum_{i \in D^c \cap B_{1l} \cap B_{2m}} \frac{c_i^2 p_{1i}(1-p_{1i})}{c_i^2 p_{2i}(1-p_{2i})} \\
&\leqslant \frac{\sum_{i=1}^{d} c_i^2 p_{1i}(1-p_{1i})}{(c^*)^2 n^{1-\beta}} \\
&\leqslant \frac{\mathrm{var}(\delta_d(X))}{(c^*)^2 n^{1-\beta}} \\
&= \mathrm{o}(n^{\nu-1}).
\end{aligned}$$

The last equation is due to the fact that $\mathrm{var}(\delta_d(X))/n^{1-\beta} = \mathrm{O}(n^{\nu-1}/n^{1-\beta}) = \mathrm{o}(n^{\nu-1})$. On D, for a sufficiently small c^*, $\mathrm{e}^{-2c^*} \leqslant \frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} \leqslant \mathrm{e}^{2c^*}$. Let $\mathrm{e}^{-2c^*} = 1 - c_1^*$ and $\mathrm{e}^{2c^*} = 1 + c_2^*$, then

$$\frac{1}{n}(d + \mathrm{o}(d))(1 - c_1^*) \leqslant \frac{1}{n} \sum_{i \in D} \frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} \leqslant \frac{1}{n}(d + \mathrm{o}(d))(1 + c_2^*).$$

For arbitrary small $c^* > 0$, $c_1^*(>0)$ and $c_2^*(>0)$ are also small. So, $\frac{1}{n}\sum_{i\in D}\frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} = cn^{\nu-1}(1+\mathrm{o}(1))$.

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{d}\frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} &= \frac{1}{n}\sum_{i\in D}\frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} + \frac{1}{n}\sum_{i\in D^c}\frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})}\\ &= cn^{\nu-1}(1+\mathrm{o}(1)). \qquad \square\end{aligned}$$

In (2) in Lemma 3.2, the boundedness of $\mathrm{var}(\delta_d(X))/n^{\nu-1}$ implies that most of $c_i's$ converge to 0 which means p_{1i} and p_{2i} are close. Most of $\frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})}$ converge to one, so $\frac{1}{n}\sum_{i=1}^{d}\frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} \approx \frac{d}{n} \approx cn^{\nu-1}$.

By using Lemma 3.2, we have the following result which plays an important role in this paper as explained in the paragraph below.

Lemma 3.3. *Under Condition A*,

(1) *If* $\mathrm{var}(\delta_d(X))/n^{\nu-1} \to \infty$, *then* $\mathrm{var}(\hat{\delta}_d(X)) = \mathrm{var}(\delta_d(X))(1+\mathrm{o}(1))$.
(2) *If* $\mathrm{var}(\delta_d(X))/n^{\nu-1}$ *is bounded, then* $\mathrm{var}(\hat{\delta}_d(X)) = \mathrm{var}(\delta_d(X)) + \mathrm{o}(\mathrm{var}(\delta_d(X))) + 2cn^{\nu-1} + \mathrm{o}(n^{\nu-1})$.

Proof. (1) Let $\hat{\delta}_i(X) = \hat{c}_i X_i + \hat{c}_{i0}$. Decompose $\hat{\delta}_i(X)$; $\hat{\delta}_i(X) - E\hat{\delta}_i(X) = \hat{c}_i X_i + \hat{c}_{i0} - E(\hat{\delta}_i(X)) = c_i(X_i - p_{1i}) + (\hat{c}_i - c_i)(X_i - p_{1i}) + \hat{c}_i p_{1i} + \hat{c}_{i0} - E(\hat{\delta}_i(X)) \equiv I_{1i} + I_{2i} + I_{3i}$, where $I_{1i} = c_i(X_i - p_{1i})$, $I_{2i} = (\hat{c}_i - c_i)(X_i - p_{1i})$ and $I_{3i} = \hat{c}_i p_{1i} + \hat{c}_{i0} - E(\hat{\delta}_i(X))$. We will show that $\frac{\mathrm{var}(\hat{\delta}_d(X))}{\mathrm{var}(\delta_d(X))} = 1 + \mathrm{o}(1)$.

$$\begin{aligned}\frac{\mathrm{var}(\hat{\delta}_d(X))}{\mathrm{var}(\delta_d(X))} &= \frac{1}{\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}(\mathrm{var}(I_{1i}) + \mathrm{var}(I_{2i}) + \mathrm{var}(I_{3i}))\\ &\quad + \frac{1}{\mathrm{var}(\delta_d(X))}\sum_{l\neq m}(\mathrm{cov}\,(I_{1l}, I_{2m}) + \mathrm{cov}(I_{2l}, I_{3m}) + \mathrm{cov}(I_{3l}, I_{1m}))\\ &= 1 + \frac{1}{\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}\mathrm{var}(I_{2i})\\ &\quad + \frac{1}{\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}\mathrm{var}(I_{3i}) + \frac{1}{\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}\mathrm{cov}(I_{1i}, I_{2i}),\end{aligned}$$

since for $l \neq m$, $\mathrm{cov}(I_{2l}, I_{3m}) = \mathrm{cov}(I_{1l}, I_{3m}) = 0$ and $\sum_{i=1}^{d}\mathrm{var}(I_{1i}) = \mathrm{var}(\delta_d(X))$. It is sufficient to show that the last three terms are $\mathrm{o}(1)$.

By Lemma A.2, $\mathrm{var}(I_{2i}) = E[(\hat{c}_i - c_i)^2(X_i - p_{1i})^2] = E(\hat{c}_i - c_i)^2 E(X_i - p_{1i})^2 = (\frac{1}{n} + \frac{p_{1i}(1-p_{1i})}{np_{2i}(1-p_{2i})})(1+\mathrm{o}(1))$. Using this, we derive

$$\begin{aligned}\frac{1}{\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}\mathrm{var}(I_{2i}) &= \frac{1}{\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}\left(\frac{1}{n} + \frac{1}{n}\frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})}\right)(1+\mathrm{o}(1))\\ &= \frac{d(1+\mathrm{o}(1))}{n\,\mathrm{var}(\delta_d(X))} + \frac{1}{n\,\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}\frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})}(1+\mathrm{o}(1)) = \mathrm{o}(1),\end{aligned}$$

since $\frac{d}{\mathrm{var}(\delta_d(X))n} = \mathrm{o}(1)$ and $\frac{1}{n\,\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}\frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} = \mathrm{o}(1)$ by Lemma 3.2.

We will show $\frac{1}{\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}\mathrm{var}(I_{3i})=\mathrm{o}(1)$. Since, $E(\hat{\delta}_i(X)-\delta_i(X))I(A_d)=\frac{(2p_{1i}-1)(p_{2i}-p_{1i})}{np_{2i}(1-p_{2i})}(1+\mathrm{o}(1))$ $I_{3i}=\hat{c}_i p_{1i}+\hat{c}_{i0}-E\hat{\delta}_i(X)=\hat{c}_i p_{1i}+\hat{c}_{i0}-E\delta_i(X)+\frac{(2p_{1i}-1)(p_{2i}-p_{1i})}{np_{2i}(1-p_{2i})}(1+\mathrm{o}(n^{-\varepsilon^*}))$

$$\begin{aligned}
\frac{1}{\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}\mathrm{var}(I_{3i})I(A_d) &= \frac{1}{\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}E(\hat{c}_i p_{1i}+\hat{c}_{i0}-E\delta_i(X)+E\delta_i(X)-E\hat{\delta}_i(X)I(A_d))^2\\
&\leqslant \frac{2}{\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}E(\hat{c}_i p_{1i}+\hat{c}_{i0}-E\delta_i(X))^2\\
&\quad+\frac{2}{\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}(E(\delta_i(X)-E\delta_i(X))I(A_d))^2\\
&\leqslant \frac{2}{\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}E(\hat{c}_i p_{1i}+\hat{c}_{i0}-c_i p_{1i}-c_{i0})^2\\
&\quad+\frac{2}{\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}\frac{|p_{2i}-p_{1i}|^2}{n^2(p_{2i}(1-p_{2i}))^2}+\mathrm{o}(1)\\
&= \mathrm{o}(1).
\end{aligned}$$

The first term converges to 0 by Lemma A.4 and the second term, $\frac{2}{n^2\,\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}\frac{|p_{2i}-p_{1i}|^2}{(p_{1i}(1-p_{1i}))^2}\leqslant\frac{2}{n\,\mathrm{var}(\delta_d(X))n^{1-\beta}}\sum_{i=1}^{d}\frac{|p_{2i}-p_{1i}|}{(p_{1i}(1-p_{1i}))}\to 0$ by Lemma A.3.

Using Lemma A.3,

$$\frac{1}{\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}\mathrm{cov}(I_{1i},I_{2i})=\frac{1}{n\,\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}\left(c_i-\frac{c_i p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})}\right)(1+\mathrm{o}(1))=\mathrm{o}(1).$$

(2) In the proof of (1) in this Lemma, the only difference is, by Lemma 3.2, $\sum_{i=1}^{d}EI_{2i}=\sum_{i=1}^{d}(\frac{1}{n}+\frac{p_{1i}(1-p_{1i})}{np_{2i}(1-p_{2i})})(1+\mathrm{o}(1))=2cn^{\nu-1}(1+\mathrm{o}(1))$. Except this term, all the other terms are the same as those in (1) in this lemma with the help of Lemma A.3. Therefore, $\mathrm{var}(\hat{\delta}_d(X))=\mathrm{var}(\delta_d(X))+2cn^{\nu-1}(1+\mathrm{o}(1))$. □

Lemma 3.3 shows how the behavior of $\mathrm{var}(\hat{\delta}_d(X))$ depends on that of $\mathrm{var}(\delta_d(X))$. When $\frac{\mathrm{var}(\delta_d(X))}{n^{\nu-1}}\to\infty$, $\mathrm{var}(\hat{\delta}_d(X))$ increases at the same rate as $\mathrm{var}(\delta_d(X))$, however, when $\frac{\mathrm{var}(\delta_d(X))}{n^{\nu-1}}$ is bounded, i.e., $\mathrm{var}(\delta_d(X))=\mathrm{O}(n^{\nu-1})$, then $\mathrm{var}(\hat{\delta}_d(X))$ is larger than $\mathrm{var}(\delta_d(X))$ by $2cn^{\nu-1}$ which is a non-negligible term. For convenience, these two results can be combined to

$$\mathrm{var}(\hat{\delta}_d(X))=\mathrm{var}(\delta_d(X))+\mathrm{o}(\mathrm{var}(\delta_d(X)))+2cn^{\nu-1}+\mathrm{o}(n^{\nu-1}). \tag{5}$$

Based on Lemma 3.3, we show the persistence of the plug-in rule when $\frac{\mathrm{var}(\delta_d(X))}{n^{\nu-1}}$ diverges under *Condition A*; otherwise, there exist cases of non-persistence of the plug-in rule. For both results, we also need other conditions.

3.1. *Persistence of the plug-in rule*

As we mentioned earlier, without loss of generality, let the true class be $j=1$. So the Bayes error is $P(\delta_d(X)<0)\equiv P(\delta_d(X)<0|j=1)$. Let $\frac{\delta_d(X)-E\delta_d(X)}{\sqrt{\mathrm{var}(\delta_d(X))}}\equiv N_d$ and $\frac{E\delta_d(X)}{\sqrt{\mathrm{var}(\delta_d(X))}}\equiv r_d$. In the same way, for the plug-in rule, $\frac{\hat{\delta}_d(X)-E\hat{\delta}_d(X)}{\sqrt{\mathrm{var}(\hat{\delta}_d(X))}}\equiv$

$\tilde{N}_d$ and $\frac{E\hat{\delta}_d(X)}{\sqrt{\mathrm{var}(\hat{\delta}_d(X))}} \equiv \tilde{r}_d$. $\delta_d(X)$ is a measure of what is called the margin in the literature of classification. In view of this, we refer to r_d (or $\tilde{r}_d$)as S.N.R. (signal to noise ratio) of $\delta_d(X)$ (or $\hat{\delta}_d(X)$). First, we investigate the behavior of Bayes rule with the known parameters under *Condition A*. When r_d diverges, then $\delta_d(X)$ discriminates two classes perfectly, i.e., misclassification error converges to 0 as d increases. By Chebyshev's inequality, this can be easily shown in the following way:

$$P(\delta_d(X) < 0) = P((N_d < -r_d) \leqslant \frac{1}{r_d^2} \to 0,$$

as $d \to \infty$. In the same way, if S.N.R. of the $\hat{\delta}_d(X)$, $\tilde{r}_d$, diverges, then $P(\hat{\delta}_d(x) < 0)$ converges to 0. By showing that the divergence of r_d implies that of $\tilde{r}_d$, we show that $P(\hat{\delta}_d(x) < 0)$ converges to 0. The following theorem shows plug-in rule, $\hat{\delta}_d(X)$, is persistent when $r_d \to \infty$ with some additional condition and *Condition A*.

Theorem 3.1. *Under Condition A, if* $\mathrm{var}(\delta_d(X)) \geqslant Ln^{v-1}$ *for some constant* L *and* $r_d \to \infty$, *then* $P(\hat{\delta}_d(X) < 0) \to 0$.

To prove Theorem 3.1, we need the following lemma.

Lemma 3.4. *Under Condition A*, $\frac{1}{nE\delta_d(X)} \sum_{i=1}^{d} \frac{|p_{1i} - p_{2i}|}{p_{2i}(1-p_{2i})} = \mathrm{o}(1)$ *and* $E\hat{\delta}_d(X) = E\delta_d(X)(1 + \mathrm{o}(1))$.

Proof. See the Appendix.

Proof of Theorem 3.1. As we showed for Bayes rule $\delta_d(X)$, we need to show that $\tilde{r}_d \to \infty$. By Eq. (5), $\mathrm{var}(\hat{\delta}_d(X)) = \mathrm{var}(\delta_d(X)) + 2cn^{v-1}(1 + \mathrm{o}(1))$. And it can be shown that $E\hat{\delta}_d(X) = E\delta_d(X) + \frac{1}{n}\sum_{i=1}^{d} \frac{(2p_{1i}-1)(p_{2i}-p_{1i})}{p_{2i}(1-p_{2i})}(1 + \mathrm{o}(1))$. Therefore, by Lemma 3.4,

$$\tilde{r}_d = \frac{E\hat{\delta}_d(X)}{\sqrt{\mathrm{var}(\hat{\delta}_d(X))}} = \frac{E\delta_d(X)(1 + \mathrm{o}(1))}{\sqrt{\mathrm{var}(\delta_d(X)) + 2cn^{v-1}}}$$

$$= \frac{r_d(1 + \mathrm{o}(1))}{\sqrt{1 + \dfrac{2cn^{v-1}}{\mathrm{var}(\delta_d(X))}}}.$$

Since $\mathrm{var}(\delta_d(X)) \geqslant Ln^{v-1}$, we have $\frac{2c}{L} \geqslant \frac{2cn^{v-1}}{\mathrm{var}(\delta_d(X))}$ and

$$\tilde{r}_d \geqslant \frac{r_d(1 + \mathrm{o}(1))}{\sqrt{1 + \dfrac{2cn^{v-1}}{\mathrm{var}(\delta_d(X))}}} \geqslant \frac{r_d(1 + \mathrm{o}(1))}{\sqrt{1 + \dfrac{2c}{L}}} \to \infty.$$

So $P(\hat{\delta}_d(X) < 0) \to 0$. □

Example 1. Suppose $d = n$ (i.e., $v = 1$ and $c = 1$) and all of the p_{ji} converges to 0 and p_{1i} and p_{2i} are getting closer. In other words, we take $p_{ji} \sim n^{-\beta}$ and $\frac{p_{1i}}{p_{2i}} - 1 = n^{-\gamma}(1 + \mathrm{o}(1))$ for β and γ such that $\beta + 2\gamma > 1$. Then $|c_i| = |\frac{p_{1i}}{p_{2i}} - 1|(1 + \mathrm{o}(1)) = n^{-\gamma}(1 + \mathrm{o}(1))$ so $\mathrm{var}(\delta_d(X)) = \sum_{i=1}^{d} c_i^2 p_{1i}(1 - p_{1i}) \sim d \cdot n^{-2\gamma} n^{-\beta} \sim n^{1-2\gamma-\beta} \to \infty$ and $r_d \sim n^{(1-2\gamma-\beta)/2} \to \infty$ since $1 - \beta - 2\gamma > 0$ which satisfies the conditions in Theorem 3.1. Therefore, $\hat{\delta}_d(X)$ is persistent.

In Theorem 3.1, we showed that both $P(\delta_d(X) < 0)$ and $P(\hat{\delta}_d(X) < 0)$ converge to 0 by using the divergence of r_d and $\tilde{r}_d$. But, when r_d is bounded, it is not guaranteed that $P(\delta_d(X) < 0)$ converges to 0. We need to consider the difference of two probabilities and investigate convergence of difference of them, $P(\hat{\delta}_d(X)) - P(\delta_d(X))$. The difference of two probabilities is $P(\hat{\delta}_d(X) < 0) - P(\delta_d(X) < 0) = P(\tilde{N}_d < -\tilde{r}_d) - P(N_d < -r_d)$. For the plug-in rule to be persistent, we expect that $\tilde{N}_d - N_d \to 0$ in probability and $\tilde{r}_d - r_d \to 0$. The following lemma provides an answer.

Lemma 3.5. *Under Condition A, if* $\operatorname{var}(\delta_d(X))/n^{v-1} \to \infty$ *and* $0 < L < r_d \leqslant M < \infty$ *for some L and M, then*

(1) $\tilde{r}_d - r_d \to 0$,
(2) $\tilde{N}_d - N_d \to 0$ *in probability.*

Proof. See the Appendix.

Even if Lemma 3.5 is satisfied, $P(\delta_d(X) < 0) - P(\hat{\delta}_d(X) < 0)$ may not converge to 0 when r_d (or $\tilde{r}_d$) is a discontinuity point of N_d (or $\tilde{N}_d$). It is not necessarily true that N_d and $\tilde{N}_d$ have limit distribution, but it can be claimed that there exists a subsequence that converges to a random variable by Helley selection Theorem. See Billingsley (1995). To claim the persistence of $\hat{\delta}_d(X)$, we want r_d to converge to a continuity point of limit of N_d. Suppose r_d converges to some random variable, say Y, and r_d converges to y, then misclassification error rate converges to $P(Y < y)$. Then, if y is a continuity point of Y, $P(\delta_d(X) < 0) \to P(Y < y)$ since $\tilde{N}_d$ weakly converges to Y by Lemma 3.5. Using this, we may claim the persistence of the plug-in rule under restricted situation in the following way:

Under the conditions in Lemma 3.5, if N_d converges weakly to some random variable, Y, and r_d converges to some continuity point of Y, then plug-in rule is persistent.

The above claim can be proved in the following way. Let $F_{n,d}$ and F_d be distribution function of $\tilde{N}_d$ and N_d, respectively. Since N_d converges weakly to Y, $\tilde{N}_d$ converges weakly to Y by Lemma 3.5. In the same way, $\tilde{r}_d$ converges to y. With these facts, $P(\hat{\delta}_d(X) < 0) - P(\delta_d(X) < 0) = F_{n,d}(-\tilde{r}_d) - F(-y) + F(-y) - F_d(-r_d) \to 0$.

Unfortunately, we have not been able to construct any example where all the conditions of this theorem hold.

3.2. Non-persistence of the plug-in rule

In Theorem 3.1, we showed the plug-in rule is persistent under some conditions. In this section, we discuss some cases of lack of persistence of the plug-in rule. In Lemma 3.4, $E\hat{\delta}_d(X) = E\delta_d(X)(1 + o(1))$, but as we mentioned in Lemma 3.3, if $\operatorname{var}(\delta_d(X)) = O(n^{v-1})$, then $\operatorname{var}(\hat{\delta}_d(X))$ has a bias in the sense $\operatorname{var}(\hat{\delta}_d(X)) = \operatorname{var}(\delta_d(X)) + o(\operatorname{var}(\delta_d(X))) + 2cn^{v-1} + o(n^{v-1})$. The biased term in $\operatorname{var}(\hat{\delta}_d(X))$ makes r_d and $\tilde{r}_d$ quite different, which leads to lack of persistence of the plug-in rule. To be more precise, we consider a case where N_d has a limiting normal distribution and show error rates of Bayes rule and the plug-in rule may be different.

Lemma 3.6. *Under Condition A with* $v \geqslant 1$, *if*

$$\sum_{i=1}^{d} |c_i|^3 p_{1i}(1 - p_{1i})/(\operatorname{var}(\delta_d(X)))^{3/2} \to 0,$$

then

(1) $N_d \to N(0, 1)$,
(2) $\tilde{N}_d \to N(0, 1)$.

Proof. See the Appendix.

With Lemma 3.6, we derive non-persistence of the plug-in rule.

Theorem 3.2. *Under the conditions in Lemma* 3.6,

(1) (i) $r_d \to \infty$, (ii) $\mathrm{var}(\delta_d(X)) = \mathrm{o}(n^{\nu-1})$ *and* (iii) $E\delta_d(X) = \mathrm{O}(n^{(\nu-1)/2})$ *or*
(2) *if* $0 < L \leqslant r_d \leqslant M < \infty$ *for some constant* L *and* M *and* $\mathrm{var}(\delta_d(X)) = \mathrm{O}(n^{\nu-1})$,

then the plug-in rule is not persistent.

Proof. By Lemma 3.6, $P(\delta_d(X) < 0) - \Phi(-r_d) = \mathrm{o}(1)$ and, by using $\mathrm{var}(\hat{\delta}_d(X)) = \mathrm{var}(\delta_d(X)) + 2cn^{\nu-1} + \mathrm{o}(n^{\nu-1})$ and $E\hat{\delta}_d(X) = (E\delta_d(X))(1+\mathrm{o}(1))$ in Lemma 3.6, $P(\hat{\delta}_d(X) < 0) - \Phi(-\tilde{r}_d) = \mathrm{o}(1)$ where $\tilde{r}_d = \frac{E\delta_d(X)}{\sqrt{\mathrm{var}(\delta_d(X)) + 2cn^{\nu-1}}}(1+\mathrm{o}(1))$. Based on these, we prove the theorem for two different cases.

(1) Since $r_d \to \infty$ and N_d has asymptotic normal distribution, $P(\delta_d(X) < 0) = \Phi(-r_d)(1 + \mathrm{o}(1)) = \mathrm{o}(1)$. Since $\mathrm{var}(\delta_d(X)) = \mathrm{o}(n^{\nu-1})$,

$$\tilde{r}_d = E\delta_d(X)/\sqrt{2cn^{\nu-1}}(1 + \mathrm{o}(1)) = \mathrm{O}(1),$$

which means $\liminf_n \Phi(-\tilde{r}_d) > 0$. Therefore, $P(\hat{\delta}_d(X) < 0) - P(\delta_d(X) < 0) \nrightarrow 0$, i.e., the plug-in rule is not persistent.

(2) Since $0 < L \leqslant r_d \leqslant M < \infty$ for some constant L and M and $\mathrm{var}(\delta_d(X)) = \mathrm{O}(n^{\nu-1})$, $\liminf(r_d - \tilde{r}_d) = \liminf(r_d - \frac{r_d(1+\mathrm{o}(1))}{\sqrt{1+2cn^{\nu-1}/\mathrm{var}(\delta_d(X))}}) > 0$. This implies $\Phi(-\tilde{r}_d) - \Phi(-r_d) \nrightarrow 0$. □

These two cases of non-persistence depend on the behavior of $\mathrm{var}(\delta_d(X))$. In $\mathrm{var}(\hat{\delta}_d(X)) = \mathrm{var}(\delta_d(X)) + 2cn^{\nu-1}(1 + \mathrm{o}(1))$, the term $2cn^{\nu-1}$ is due to estimating $d = 2cn^{\nu}$ parameters. In (1) of Theorem 3.2, if $\mathrm{var}(\delta_d(X)) = \mathrm{o}(n^{\nu-1})$, the variability due to estimating parameters dominates $\mathrm{var}(\delta_d(X))$. With $E\delta_d(X) = \mathrm{O}(n^{(\nu-1)/2})$ in (1), this makes $\tilde{r}_d$ bounded while $r_d \to \infty$. In (2), with the same argument, $\mathrm{var}(\delta_d(X)) = \mathrm{O}(n^{\nu-1})$ makes $2cn^{\nu-1}$ non-negligible term and this cause non-persistence of the plug-in rule. These cases of non-persistence occur since there are too many parameters to be estimated and due to this, the variability of the plug-in rule, $\mathrm{var}(\hat{\delta}_d(X))$, is significantly larger than the original variability, $\mathrm{var}(\delta_d(X))$, while $E\hat{\delta}_d(X) \sim E\delta_d(X)$ by Lemma 3.4.

Example 2 (*Case of (1) in Theorem 3.2*). Suppose $d = n^2$ (i.e., $\nu = 1$ and $c = 1$) and when $1 \leqslant i \leqslant [\sqrt{n}]$, $|p_{1i} - p_{2i}| \geqslant \varepsilon$ for some $\varepsilon > 0$; otherwise $|p_{1i} - p_{2i}| = \mathrm{O}(n^{-3})$. Then $\sum_{i=1}^{d} |c_i|^2 p_{1i}(1 - p_{1i})/(\mathrm{var}(\delta_d(X)))^{3/2} \to 0$ and $\mathrm{var}(\delta_d(X)) \sim \mathrm{O}(\sqrt{n}) \leqslant \mathrm{o}(n^{2-1})$ and $E\delta_d(X) \asymp \sqrt{n} \leqslant \mathrm{O}(n^{(2-1)/1})$. Therefore, $r_d = \sqrt{n}/n^{1/4} \to \infty$. But $\tilde{r}_d = \mathrm{O}(1)$ shows the plug-in rule is not persistent.

Example 3 (*Case of (2) in Theorem 3.2*). Suppose $c = 1$, i.e., $d/n \to 1$. Let $p_{1i} = n^{-\beta}$ and p_{2i} such that $\frac{p_{1i}}{p_{2i}} = 1 + n^{-\gamma}$ where $\beta + 2\gamma = 1$ and $\gamma > 0$. Then $|c_i| \sim n^{-\gamma}$ and $\mathrm{var}(\delta_d(X)) \sim \sum_{i=1}^{d} n^{-2\gamma} n^{-\beta}(1 - n^{-\beta}) \sim n^{1-2\gamma-\beta} \sim 1$. Since $\sum_{i=1}^{d} |c_i|^3 p_{1i}(1 - p_{1i}) \sim n^{1-3\gamma-\beta} \to 0$, $N_d = \frac{\delta_d(X) - E\delta_d(X)}{\sqrt{\mathrm{var}(\delta_d(X))}}$ has asymptotic normal distribution. Since $\mathrm{var}(\delta_d(X)) \sim n^{1-2\beta-\gamma} \sim 1$ and $E\delta_d(X) \sim n^{1-2\gamma-\beta} \sim 1$, $r_d \sim 1$ and $P(\delta_d(X) < 0) - \Phi(-1) = \mathrm{o}(1)$. But $\tilde{r}_d \sim 1/\sqrt{1 + 2c} = 1/\sqrt{3}$ implies $P(\hat{\delta}_d(X) < 0) - \Phi(-1/\sqrt{3})$. Therefore, the plug-in rule is not persistent.

4. Sparsity condition under linear rules

In many situations, classification rules with suitably selected variables outperform the original classifier especially in high dimensional data. For variable selection to be effective, sparsity condition is needed to ensure that only a small subset of the variables is helpful in classification. In our context, one defines sparsity condition in a rather simple way as follows. Suppose there is a subset of the variables, D with $|D| = \mathrm{o}(d)$. Let $\delta_d(X) = \sum_{i \in D}(c_i X_i + c_{i0}) + \sum_{i \in D^c}(c_i X_i + c_{i0}) \equiv \delta_d^D(X) + \delta_d^{D^c}(X)$. Assume $\mathrm{var}(\delta_d^D(X)) = \mathrm{var}(\delta_d(X))(1 + \mathrm{o}(1))$ and $E\delta_d^D(X) = E\delta_d(X)(1 + \mathrm{o}(1))$. In this situations, the variables in D^c are redundant in the sense that $\mathrm{var}(\delta_d^{D^c}(X))$ and $E\delta_d^{D^c}(X)$ are negligible compared to $\mathrm{var}(\delta_d^D(X))$ and $E\delta_d^D(X)$, respectively. With these, we propose sparsity condition especially under linear classifier.

4.1. Sparsity condition under linear rule

There exists a subset, D, such that $|D| = \mathrm{o}(d)$, $E\delta_d^{D^c}(X)/E\delta_d^D(X) \to 0$ and $\mathrm{var}(\delta_d^{D^c}(X))/\mathrm{var}(\delta_d^D(X)) \to 0$.

This sparsity condition implies the variables in D^c do not affect $\delta_d(X)$ asymptotically. We define some notations such as $r_d^D, r_d^{D^c}, N_d^D, N_d^{D^c}, \tilde{r}_d^D$ and $\tilde{r}_d^{D^c}$ in the same way as done earlier for $D \cup D^c$. Then $r_d = \frac{E\delta_d^D(X)+E\delta_d^{D^c}(X)}{\sqrt{\mathrm{var}(\delta_d^D(X))+\mathrm{var}(\delta^{D^c}(X))}} = \frac{E\delta_d^D(X)}{\sqrt{\mathrm{var}(\delta_d^D(X))}}(1+\mathrm{o}(1)) = r_d^D(1+\mathrm{o}(1))$. We consider below $|D| = \mathrm{o}(n\,\mathrm{var}(\delta_d(X)))$ under conditions of Theorem 3.2.

Theorem 4.1. *Under the conditions in Theorem 3.2, if there exists a subset D such that $|D| = \mathrm{o}(n\,\mathrm{var}(\delta_d(X)))$ satisfying sparsity condition, then $P(\hat{\delta}_d^D(X) < 0) - P(\delta_d(X) < 0) = \mathrm{o}(1)$.*

Proof. By using $\mathrm{var}(\delta_d(X)) = \mathrm{var}(\delta_d^D(X))(1+\mathrm{o}(1))$ from sparsity condition, it can be shown that $\sum_{i\in D}|c_i^3|p_{1i}(1-p_{1i})/(\mathrm{var}(\delta_d^D(X)))^{3/2} \to 0$ which implies N_d^D has asymptotic normal distribution. Therefore, $P(\hat{\delta}_d^D(X) < 0) - \Phi(-\tilde{r}_d^D) = \mathrm{o}(1)$. Since $|D| = \mathrm{o}(n\,\mathrm{var}(\delta_d(X)))$, $\mathrm{var}(\hat{\delta}_d^D(X)) = \mathrm{var}(\delta_d^D(X))(1+\mathrm{o}(1)) + 2c|D|/n(1+\mathrm{o}(1)) = \mathrm{var}(\delta_d(X))(1+\mathrm{o}(1))$. With $E\delta_d^D(X) = E\delta_d(X)(1+\mathrm{o}(1))$ by sparsity condition, $\tilde{r}_d{}^D = E\delta_d(X)/\sqrt{\mathrm{var}(\delta_d^D(X))} = r_d(1+\mathrm{o}(1))$. This proves $P(\hat{\delta}_d^D(X) < 0) - P(\delta_d(X) < 0) = P(\hat{\delta}_d^D(X) < 0) - \Phi(-\tilde{r}_d) + \Phi(-\tilde{r}_d) - P(-r_d) + P(-r_d) - P(\delta_d(X) < 0) = \mathrm{o}(1)$ which shows $\delta_d^D(X)$ as persistent with the subset D is persistent although $\hat{\delta}_d(X)$ as not persistent by Theorem 3.2. □

Example 4. Suppose $d = n$ (i.e., $v = 1$ and $c = 1$). Let $D = \{i : p_{1i} = n^{-\beta},\ p_{1i}/p_{2i} = 1 + n^{-\gamma}\}$ and $D^c = \{i : p_{1i}/p_{2i} = 1 + n^{-2}\}$ with $|D| = [n^{2\gamma+\beta}]$ where $0 < 2\gamma+\beta < 1$ and $[n^{2\gamma+\beta}]$ is the integer part of $n^{2\gamma+\beta}$. Then $E\delta_d^D(X) \sim 1$ and $\mathrm{var}(\delta_d^D) \sim 1$. From this, $|D| \sim n^{2\gamma+\beta} = \mathrm{o}(n\,\mathrm{var}(\delta_d(X))) = \mathrm{o}(n)$. $\delta_d^D(X)$ has asymptotic normality since $\sum_{i\in D}|c_i^3|p_{1i}(1-p_{1i})/(\mathrm{var}(\delta_d(X)))^{3/2} \sim n^{2\gamma+\beta}n^{-3\gamma}n^{-\beta}(1-n^{-\beta}) \sim n^{-\gamma} \to 0$ by Lyapounov condition (see Billingsley, 1995). By Theorem 4.1, $P(\hat{\delta}_d^D(X) < 0) - P(\delta_d(X) < 0) \to 0$, i.e., $\hat{\delta}_d^D(X)$ is persistent while $P(\hat{\delta}_d(X) < 0) - \Phi(-1/\sqrt{3}) = \mathrm{o}(1)$ which implies non-persistence of the plug-in rule.

The above theorem shows that if there are many variables, then using all the variables increases variability and $\mathrm{var}(\hat{\delta}_d(X))$ is significantly larger than $\mathrm{var}(\delta_d(X))$. This causes a bad prediction performance of the plug-in rule. But, under sparsity condition, selection of a good subset of variables achieves a better performance than the plug-in rule with all the variables. Wilbur et al. (2002) proposed two variable selection methods. Their results show that the plug-in rule with selected variables improve performance. In the Appendix all the lemmas are collected. Before stating a lemma, we mention where it is used. The lemmas have been arranged so that no lemmas requires a later lemma in its proof.

Appendix A. Proof of Lemmas

The following Lemma is used in the proof of Lemma 3.1.

Lemma A.1. *When X_i is i.i.d.* Bernoulli(p), *then, for* $0 < \varepsilon < 1$,

$$P\left\{\sum_{i=1}^{n}(X_i - p) > \varepsilon np(1-p)\right\} \leqslant \mathrm{e}^{-3np(1-p)\varepsilon^2/8},$$

$$P\left\{\sum_{i=1}^{n}(X_i - p) < -\varepsilon np(1-p)\right\} \leqslant \mathrm{e}^{-3np(1-p)\varepsilon^2/8}.$$

Proof. Using the fact $|X_i - p| \leqslant 1$ and Bernstein's inequality (see Devroye et al., 1996),

$$P\left\{\sum_{i=1}^{n}(X_i - p) > n\varepsilon p(1-p)\right\} \leqslant \exp\left\{-\frac{np^2(1-p)^2\varepsilon^2}{2(p(1-p)+\varepsilon p(1-p)/3)}\right\}$$

$$\leqslant \exp\left\{-\frac{np(1-p)\varepsilon^2}{2(1+\varepsilon/3)}\right\}$$

$$\leqslant \exp\left\{-\frac{3np(1-p)\varepsilon^2}{8}\right\}.$$

The second inequality can be shown in the similar way. □

In $\hat{\delta}_d(X) = \sum_{i=1}^{d}(\hat{c}_i X_i + c_{i0})$, we approximate $\hat{c}_i - c_i$ by $\hat{p}_{1i} - p_{1i}$ and $\hat{p}_{2i} - p_{2i}$. Using Taylor expansion, on $A_d = \bigcap_{i=1}^{d}\{|\frac{\hat{p}_{ji}-p_{ji}}{p_{ji}(1-p_{ji})}| < n^{-\varepsilon^*}, j = 1, 2\}$, $\hat{c}_i - c_i = \mathbf{T}_i(1 + \mathrm{o}(n^{-\varepsilon^*}))$ or $\hat{c}_i - c_i = \mathbf{T}_i + \mathbf{R}_i(1 + \mathrm{o}(n^{-\varepsilon^*}))$, where

$$\mathbf{T}_i = \frac{\hat{p}_{1i} - p_{1i}}{p_{1i}(1-p_{1i})} - \frac{\hat{p}_{2i} - p_{2i}}{p_{2i}(1-p_{2i})},$$
$$\mathbf{R}_i = \frac{(1-2p_{1i})}{(p_{1i}(1-p_{1i}))^2}(\hat{p}_{1i} - p_{1i})^2 - \frac{(1-2p_{2i})}{(p_{2i}(1-p_{2i}))^2}(\hat{p}_{2i} - p_{2i})^2.$$

The following lemma will be used in Lemma 3.6.

Lemma A.2. *Under Condition A*,

(1) $E\mathbf{T}_i = 0$, $E\mathbf{R}_i I(A_d) = \mathrm{O}(\frac{1}{np_{1i}(1-p_{1i})} + \frac{1}{np_{2i}(1-p_{2i})})(1 + \mathrm{o}(1))$,
(2) $E\mathbf{T}_i^2 = \frac{1}{np_{1i}(1-p_{1i})} + \frac{1}{np_{2i}(1-p_{2i})}$,
(3) $E(\hat{c}_i - c_i)^2 I(A_d) = (\frac{1}{np_{1i}(1-p_{1i})} + \frac{1}{np_{2i}(1-p_{2i})})(1 + \mathrm{o}(1))$.

Proof. The proof follows from Lemma and direct computation of $E\mathbf{K}_i$ where $\mathbf{K}_i = \mathbf{T}_i, \mathbf{R}_i, \mathbf{T}_i^2$. □

This lemma is used in the proof of Lemma 3.3.

Lemma A.3. *Under Condition A, if* $\mathrm{var}(\delta_d(X))/n^{\gamma-1} \to \infty$, *then*

(1) $\frac{1}{n\,\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}|c_i|^k = \mathrm{o}(1)$ *for* $k \geqslant 1$,
(2) $\frac{1}{n\,\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}\frac{c_i^k p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} = \mathrm{o}(1)$ *for* $k \geqslant 1$,
(3) $\frac{1}{n\,\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}\frac{|p_{2i}-p_{1i}|}{p_{2i}(1-p_{2i})} = \mathrm{o}(1)$ *if* $\mathrm{var}(\delta_d(X))/n^{\gamma-1}$ *is bounded, then*
(4) $\frac{1}{n}\sum_{i=1}^{d}|c_i|^k = \mathrm{o}(n^{\gamma-1})$ *for* $k \geqslant 1$,
(5) $\frac{1}{n}\sum_{i=1}^{d}\frac{c_i^k p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} = \mathrm{o}(n^{\gamma-1})$ *for* $k \geqslant 1$,
(6) $\frac{1}{n}\sum_{i=1}^{d}\frac{|p_{2i}-p_{1i}|}{p_{2i}(1-p_{2i})} = \mathrm{o}(n^{\gamma-1})$.

Proof. (1) Take $D_n = \{i \,\|c_i| \leqslant n^{-(1-\beta)/3}\}$ and B_{1l}, B_{2m} as in the proof of Lemma 3.2. By using $|D_n| \leqslant d$ and $|c_i| \leqslant \log n$,

$$\frac{1}{n\,\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}|c_i|^k = \frac{1}{n\,\mathrm{var}(\delta_d(X))}\sum_{i\in D_n}|c_i|^k + \sum_{l,m}\sum_{i\in D_n^c\cap B_{1l}\cap B_{2m}}\frac{|c_i|^k}{n\,\mathrm{var}(\delta_d(X))}$$
$$\equiv (\mathrm{I}) + (\mathrm{II}).$$

We first handle the first term (I) as follows:

$$\text{(I)} \leqslant \frac{cn^{v-1}}{n^{(k(1-\beta)/3)}\operatorname{var}(\delta_d(X))} \leqslant \frac{cn^{v-1}}{n^{(k(1-\beta)/2)}L} \to 0.$$

To handle the second term (II), we have to proceed with the following lower bound for $\operatorname{var}(\delta_d(X))$. On $D_n^c \cap B_{1l} \cap B_{2m}$,

$$\begin{aligned}\frac{1}{\operatorname{var}(\delta_d(X))} &\leqslant \frac{1}{\sum_{i\in D_n^c\cap B_{1l}\cap B_{2m}} c_i^2 p_{1i}(1-p_{1i})}\\ &\leqslant \frac{1}{|D_n^c \cap B_{1l} \cap B_{2m}|n^{-\beta+(l-1)\varepsilon}}.\end{aligned}$$

With the above as well as using $n^{-(1-\beta)/3} \leqslant |c_i| \leqslant \log n$, the second term (II),

$$\begin{aligned}\text{(II)} &\leqslant \sum_{l,m} \frac{|D_n^c \cap B_{1l} \cap B_{2m}|(\log n)^k}{n^{-(2(1-\beta)/3)}n^{1-\beta+(l-1)\varepsilon}|D_n^c \cap B_{1l} \cap B_{2m}|}\\ &\leqslant \left(\left[\frac{\beta}{\varepsilon}\right]+1\right)^2 \frac{(\log n)^k}{n^{-(2(1-\beta)/3)}n^{1-\beta}}\\ &\leqslant \left(\left[\frac{\beta}{\varepsilon}\right]+1\right)^2 \frac{(\log n)^k}{n^{(1-\beta)/3}} \to 0.\end{aligned}$$

This proves $\frac{1}{n\operatorname{var}(\delta_d(X))}\sum_{i=1}^d |c_i|^k = \text{(I)} + \text{(II)} \to 0$.

(2) Take $D_n = \{i \,||c_i| \leqslant n^{-(1-\beta-\varepsilon)/3}\}$.

$$\begin{aligned}&\frac{1}{n\operatorname{var}(\delta_d(X))}\sum_{i=1}^d \frac{|c_i|^k p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})}\\ &\leqslant \frac{1}{n\operatorname{var}(\delta_d(X))}\sum_{i\in D_n} \frac{|c_i|^k p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} + \frac{1}{n\operatorname{var}(\delta_d(X))}\sum_{i\in D_n^c} \frac{|c_i|^k p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})}\\ &\equiv \text{(I)} + \text{(II)}.\end{aligned}$$

In the proof of (1) in Lemma 3.2, we showed that if $|c_i| \leqslant c^*$, then $\frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} \leqslant \mathrm{e}^{2c^*}$. Therefore, on D_n, $\frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} \leqslant \exp(2n^{-(1-\beta-\varepsilon)/3})$. The first term (I) converges to 0 since the first term (I) $\leqslant \frac{d\exp(cn^{v-1}n^{-(1-\beta-\varepsilon)/3})}{\operatorname{var}(\delta_d(X))n^{(1-\beta-\varepsilon)/3}} \to 0$.

On D_n^c, using $n^{-(1-\beta-\varepsilon)/3} \leqslant |c_i| \leqslant \log n$, it can be shown in the same way as in the proof of (1) in Lemma 3.2 the second term (II) converges to 0 since

$$\begin{aligned}\text{(II)} &\leqslant \frac{(\log n)^k}{n\operatorname{var}(\delta_d(X))}\sum_{i\in D_n^c} \frac{p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})}\\ &\leqslant \left(\left[\frac{\beta}{\varepsilon}\right]+1\right)^2 \frac{(\log n)^k}{n^{-(2(1-\beta-\varepsilon)/3)}n^{1-\beta-\varepsilon}}\\ &= \left(\left[\frac{\beta}{\varepsilon}\right]+1\right)^2 \frac{(\log n)^k}{n^{(1-\beta-\varepsilon)/3}}\\ &\to 0.\end{aligned}$$

This proves $\frac{1}{n\operatorname{var}(\delta_d(X))}\sum_{i=1}^d \frac{|c_i|^k p_{1i}(1-p_{1i})}{p_{2i}(1-p_{2i})} = \text{(I)} + \text{(II)} \to 0$.

(3)

$$\begin{aligned}
&\frac{1}{n\operatorname{var}(\delta_d(X))}\sum_{i=1}^{d}\frac{|p_{2i}-p_{1i}|}{p_{2i}(1-p_{2i})}\\
&\quad=\frac{1}{n\operatorname{var}(\delta_d(X))}\sum_{p_{2i}<\frac{1}{2}}\frac{|p_{2i}-p_{1i}|}{p_{2i}(1-p_{2i})}+\frac{1}{n\operatorname{var}(\delta_d(X))}\sum_{p_{2i}\geqslant\frac{1}{2}}\frac{|p_{2i}-p_{1i}|}{p_{2i}(1-p_{2i})}\\
&\quad\leqslant\frac{2}{n\operatorname{var}(\delta_d(X))}\sum_{p_{2i}<\frac{1}{2}}\frac{|p_{2i}-p_{1i}|}{p_{2i}}+\frac{2}{n\operatorname{var}(\delta_d(X))}\sum_{p_{2i}\geqslant\frac{1}{2}}\frac{|(1-p_{2i})-(1-p_{1i})|}{1-p_{2i}}.
\end{aligned}$$

Since $\frac{1}{n\operatorname{var}(\delta_d(X))}\sum_{p_{2i}\geqslant\frac{1}{2}}\frac{|(1-p_{2i})-(1-p_{1i})|}{1-p_{2i}}=o(1)$ can be shown in the same way as $\frac{1}{n\operatorname{var}(\delta_d(X))}\sum_{p_{2i}<\frac{1}{2}}\frac{|p_{2i}-p_{1i}|}{p_{2i}}=o(1)$. For $l,m=1,\ldots,[\frac{\beta}{\varepsilon}],\frac{\beta}{\varepsilon}$, denote

$$\begin{aligned}
&B_{1l}=\{i|n^{-\beta+(l-1)\varepsilon}\leqslant p_{1i}<n^{-\beta+l\varepsilon}\},\\
&B_{2m}=\{i|n^{-\beta+(m-1)\varepsilon}\leqslant p_{2i}<n^{-\beta+m\varepsilon}\},\\
&D_n=\{i||c_i|\leqslant n^{-(1-\beta)/3}\}.
\end{aligned}$$

For $i\in D_n$, there exists c_n^* such that $\frac{p_{1i}}{p_{2i}}\leqslant 1+c_n^*$ where $c_n^*\to 0$. For notational convenience, we omit $p_{2i}<\frac{1}{2}$. Then

$$\begin{aligned}
&\frac{1}{n\operatorname{var}(\delta_d(X))}\sum_{i}\frac{|p_{2i}-p_{1i}|}{p_{2i}}\\
&\quad\leqslant\frac{c_n^*|D_n|}{n\operatorname{var}(\delta_d(X))}+\frac{1}{n\operatorname{var}(\delta_d(X))}\sum_{l,m}\sum_{i\in D_n^c\cap B_{1l}\cap B_{2m}}\frac{|p_{2i}-p_{1i}|}{p_{2i}}\\
&\quad\leqslant\frac{c_n^*|D_n|}{n\operatorname{var}(\delta_d(X))}+\sum_{l,m}\sum_{i\in D_n^c\cap B_{1l}\cap B_{2m},\,p_{1i}\geqslant p_{2i}}\frac{|p_{2i}-p_{1i}|}{p_{2i}}+\sum_{l,m}\sum_{i\in D_n^c\cap B_{1l}\cap B_{2m},\,p_{1i}<p_{2i}}\frac{|p_{2i}-p_{1i}|}{p_{2i}}\\
&\quad\equiv(\mathrm{I})+(\mathrm{II})+(\mathrm{III}).
\end{aligned}$$

The first term (I) is $\frac{c_n^*|D_n|}{n\operatorname{var}(\delta_d(X))}=\frac{c_n^*cn^{r-1}}{\operatorname{var}(\delta_d(X))}\leqslant\frac{c_n^*}{L}\to 0$

If $p_{1i}\geqslant p_{2i}$, then $\frac{|p_{1i}-p_{2i}|}{p_{2i}}\leqslant\frac{p_{2i}}{p_{1i}}$. So the second term (II) is

$$\begin{aligned}
(\mathrm{II})&=\frac{1}{n\operatorname{var}(\delta_d(X))}\sum_{l,m}\sum_{i\in D_n^c\cap B_{1l}\cap B_{2m},\,p_{1i}\geqslant p_{2m}}\frac{|p_{2i}-p_{1i}|}{p_{2i}}\\
&\leqslant\frac{1}{n\operatorname{var}(\delta_d(X))}\sum_{l,m}\sum_{i\in D_n^c\cap B_{1l}\cap B_{2m},\,p_{1i}\geqslant p_{2i}}\frac{p_{1i}}{p_{2i}},
\end{aligned}$$

since $\frac{1}{\operatorname{var}(\delta_d(X))}\leqslant\frac{2}{|D_n^c\cap B_{1l}\cap B_{2m}|n^{-(2(1-\beta)/3)}n^{-\beta+(l-1)\varepsilon}}$ for sufficiently large n, the above is

$$\begin{aligned}
&\leqslant\sum_{l,m}\frac{2|D_n^c\cap B_{1l}\cap B_{2m}|}{|D_n^c\cap B_{1l}\cap B_{2m}|n^{(1-\beta)/3+(l-1)\varepsilon}}\frac{n^{-\beta+l\varepsilon}}{n^{-\beta+(m-1)\varepsilon}}\\
&\leqslant\left(\left[\frac{\beta}{\varepsilon}\right]+1\right)^2\frac{2}{n^{(1-\beta)/3-\varepsilon}}\to 0.
\end{aligned}$$

If $p_{1i} < p_{2i}$, then $\frac{|p_{1i}-p_{2i}|}{p_{2i}} \leqslant 1$. The third term (III) is

$$
\begin{aligned}
\text{(III)} &\leqslant \frac{1}{n\,\mathrm{var}(\delta_d(X))} \sum_{l,m} \sum_{i \in D_n^c \cap B_{1l} \cap B_{2m},\, p_{1i} < p_{2i}} 1 \\
&\leqslant \sum_{l,m} \frac{|D_n^c \cap B_{1l} \cap B_{2m}|}{|D_n^c \cap B_{1l} \cap B_{2m}| n^{1-\beta+(l-1)\varepsilon}} \\
&\leqslant \left(\left[\frac{\beta}{\varepsilon} \right] + 1 \right)^2 \frac{1}{n^{1-\beta}} \\
&\to 0.
\end{aligned}
$$

This proves $\frac{1}{n\,\mathrm{var}(\delta_d(X))} \sum_{i=1}^{d} \frac{|p_{1i}-p_{2i}|}{p_{2i}} = \text{(I)} + \text{(II)} + \text{(III)} \to 0$. In a similar way, $\frac{1}{n\,\mathrm{var}(\delta_d(X))} \sum_{i=1}^{d} \frac{|p_{1i}-p_{2i}|}{1-p_{2i}} \to 0$. So $\frac{1}{n\,\mathrm{var}(\delta_d(X))} \sum_{i=1}^{d} \frac{|p_{1i}-p_{2i}|}{p_{2i}(1-p_{2i})} \to 0$. For proof of (4), (5) and (6), if we use the idea of (2) in Lemma 3.2, they can be shown easily. □

The following lemma is used in the proof of Lemmas 3.3, 3.5 and 3.6.

Lemma A.4. *Under Condition A,*

(1) $\frac{1}{\mathrm{var}(\hat{\delta}_d(X))} \sum_{i=1}^{d} E(\hat{c}_i p_{1i} + \hat{c}_{i0} - c_i p_{1i} - c_{i0})^2 = \mathrm{o}(1)$,
(2) *if* $r_d \leqslant M$ *for some constant* M, $\frac{1}{\sqrt{\mathrm{var}(\delta_d(X))}} \sum_{i=1}^{d} E(\hat{c}_i p_{1i} + \hat{c}_{i0} - c_{i0} p_{1i} - c_{i0}) = \mathrm{o}(1)$.

Proof. Let $\eta_i = \hat{c}_i p_{1i} + \hat{c}_{i0} - c_i p_{1i} - c_{i0}$. We use the first order approximation in (1) and the second order approximation in (2). Let $f(x, y) = p_{1i} \log(\frac{x}{1-x} \frac{1-y}{y}) + \log(\frac{1-x}{1-y})$, then $\eta_i = f(\hat{p}_{1i}, \hat{p}_{2i}) - f(p_{1i}, p_{2i})$. Define f_1, f_2, f_{11}, f_{12} and f_{22} as the partial derivatives with respect to x and y corresponding to 1 and 2, respectively. Then $f_1(x, y) = \frac{p_{1i}-x}{x(1-x)}$, $f_2(x, y) = -\frac{p_{2i}-y}{y(1-y)}$, $f_{11}(x, y) = \frac{-x^2+2p_{1i}x-p_{1i}}{(x(1-x))^2}$ and $f_{22}(x, y) = \frac{-y^2+2p_{1i}y-p_{1i}}{(y(1-y))^2}$, $f_{12}(x, y) = 0$.

(1) To show this, we use the first order expansion. $\eta_i = \frac{p_{1i}-\psi_{1i}}{\psi_{1i}(1-\psi_{1i})}(\hat{p}_{1i} - p_{1i}) - \frac{p_{2i}-\psi_{2i}}{\psi_{2i}(1-\psi_{2i})}(\hat{p}_{2i} - p_{2i})$ where, for $j = 1, 2$, ψ_{ji} is the interior point between p_{ji} and $\hat{p}_{ji}$. By Lemma 3.1, on A_d, since $\hat{p}_{ji} = p_{ji}(1 + \mathrm{O}(n^{-\varepsilon^*}))$ and ψ_{ji} is between p_{ji} and $\hat{p}_{ji}$, $\psi_{ji} = p_{ji}(1 + \mathrm{O}(n^{-\varepsilon^*}))$. In the same way, $(1 - \psi_{ji}) = (1 - p_{ji})(1 + \mathrm{O}(n^{-\varepsilon^*}))$. Using these, $\frac{1}{\psi_{ji}(1-\psi_{ji})} = \frac{1}{p_{ji}(1-p_{ji})}(1 + \mathrm{O}(n^{-\varepsilon^*}))$. Since$|f_k(\psi_{1i}, \psi_{2i})| \leqslant n^{-\varepsilon^*}$ for $k = 1, 2$ and $(a + b)^2 \leqslant 2(a^2 + b^2)$,

$$
\begin{aligned}
&\frac{1}{\mathrm{var}(\hat{\delta}_d(X))} \sum_{i=1}^{d} E\eta_i^2 I(A_d) \\
&\quad = \frac{1}{\mathrm{var}(\hat{\delta}_d(X))} \sum_{i=1}^{d} E(f_1(\psi_{1i}, \psi_{2i})(\hat{p}_{1i} - p_{1i}) - f_2(\psi_{1i}, \psi_{2i})(\hat{p}_{2i} - p_{2i}))^2 I(A_d) \\
&\quad \leqslant \frac{2}{\mathrm{var}(\delta_d(X))} \sum_{i=1}^{d} [E(f_1(\psi_{1i}, \psi_{2i})(\hat{p}_{1i} - p_{1i}))^2 I(A_d) + E(f_2(\psi_{1i}, \psi_{2i})(\hat{p}_{2i} - p_{2i}))^2 I(A_d)] \\
&\quad \leqslant \frac{2n^{-2\varepsilon^*}}{\mathrm{var}(\hat{\delta}_d(X))} \sum_{i=1}^{d} \left(\frac{p_{1i}(1-p_{1i})}{n} + \frac{p_{2i}(1-p_{2i})}{n} \right)(1 + \mathrm{o}(1)) \\
&\quad \leqslant \frac{2n^{-2\varepsilon^*}}{\mathrm{var}(\delta_d(X)) + 2cn^{\nu-1}(1 + \mathrm{o}(1))} \frac{d}{n} \leqslant \frac{2n^{-2\varepsilon^*} cn^{\nu-1}}{2cn^{\nu-1}(1 + \mathrm{o}(1))} = \mathrm{o}(1).
\end{aligned}
$$

And, by Lemma 3.1, $\frac{1}{\mathrm{var}(\hat{\delta}_d(X))} \sum_{i=1}^{d} E[\eta_i^2 I(A_d^c)] \leqslant \frac{1}{2cn^{\nu-1}(1+\mathrm{o}(1))} \sum_{i=1}^{d} E[\eta_i^2 I(A_d^c)] = \mathrm{o}(1)$ since η_i^2 is at most of polynomial order of n. So $\frac{1}{\mathrm{var}(\hat{\delta}_d(X))} \sum_{i=1}^{d} E\eta_i^2 = \mathrm{o}(1)$.

To show $\frac{1}{\sqrt{\mathrm{var}(\delta_d(X))}}\sum_{i=1}^{d} E\eta_i \rightarrow 0$, we use the second order Taylor expansion, on A_d, $\eta_i = f_x(p_{1i}, p_{2i})(\hat{p}_{1i} - p_{1i}) + f_y(p_{1i}, p_{2i})(\hat{p}_{2i} - p_{2i}) + \frac{1}{2} f_{11}(\psi_{1i}, \psi_{2i})(\hat{p}_{1i} - p_{1i})^2 + \frac{1}{2} f_{22}(\psi_{1i}, \psi_{2i})(\hat{p}_{2i} - p_{2i})^2$ where ψ_{ji} is between p_{ji} and $\hat{p}_{ji}$ for $j=1, 2$. Using $\psi_{ji} = p_{ji}(1+\mathrm{o}(n^{-\varepsilon^*}))$ and $1-\psi_{ji} = (1-p_{ji})(1+\mathrm{o}(n^{-\varepsilon^*}))$ on A_d and $\frac{1}{\sqrt{\mathrm{var}(\delta_d(X))}} \leqslant \frac{M}{E\delta_d(X)}$,

$$\begin{aligned}
\frac{1}{\sqrt{\mathrm{var}(\delta_d(X))}} \sum_{i=1}^{d} |E\eta_i I(A_d)| &= \frac{1}{n\sqrt{\mathrm{var}(\delta_d(X))}} \sum_{i=1}^{d} \frac{|(2p_{1i}-1)(p_{2i}-p_{1i})|}{p_{2i}(1-p_{2i})}(1+\mathrm{o}(1)) \\
&\leqslant \frac{M}{nE\delta_d(X)} \sum_{i=1}^{d} \frac{|p_{2i}-p_{1i}|}{p_{2i}(1-p_{2i})}(1+\mathrm{o}(1)) \\
&= \mathrm{o}(1) \quad \text{by Lemma 3.4.}
\end{aligned}$$

With $\frac{1}{\sqrt{\mathrm{var}(\delta_d(X))}}\sum_{i=1}^{d} E\eta_i I(A_d^c) = \mathrm{o}(1)$, we proved $\frac{1}{\sqrt{\mathrm{var}(\delta_d(X))}}\sum_{i=1}^{d} E\eta_i = \mathrm{o}(1)$. □

Appendix B. Proof of Lemma 3.4

$$\begin{aligned}
\frac{1}{nE\delta_d(X)} \sum_{i=1}^{d} \frac{|p_{1i}-p_{2i}|}{p_{2i}(1-p_{2i})} &\leqslant \frac{2}{nE\delta_d(X)} \sum_{i:p_{2i}<\frac{1}{2}} \frac{|p_{1i}-p_{2i}|}{p_{2i}} + \frac{2}{nE\delta_d(X)} \sum_{i:p_{2i}\geqslant\frac{1}{2}} \frac{|(1-p_{1i})-(1-p_{2i})|}{1-p_{2i}} \\
&\equiv (\mathrm{I}) + (\mathrm{II}).
\end{aligned}$$

It is enough to show that the first term converges to 0. As in the proof of (3) in Lemma A.3, denote $B_{1l} = \{i : n^{-\beta+(l-1)\varepsilon} \leqslant p_{1i} < n^{-\beta+l\varepsilon}\}$, $B_{2m} = \{i : n^{-\beta+(m-1)\varepsilon} \leqslant p_{2i} < n^{-\beta+m\varepsilon}\}$, $D = \{|\frac{p_{1i}-p_{2i}}{p_{2i}}| \leqslant M_n\}$ where $M_n = n^{-(1/3)(1-\beta-\varepsilon)}$.

On D, it can be shown that $K(p_{1i}, p_{2i}) = \frac{(p_{2i}-p_{1i})^2}{p_{2i}(1-p_{2i})}(1+\mathrm{o}(1))$. Therefore, when $p_{2i} \leqslant \frac{1}{2}$, $\frac{1}{n}\frac{|p_{2i}-p_{1i}|}{p_{2i}} \leqslant \frac{1}{n^{1-\beta}}\frac{|p_{2i}-p_{1i}|^2}{2p_{2i}(1-p_{2i})}$ which means

$$\begin{aligned}
(I) &= \frac{1}{nE\delta_d(X)} \sum_{i\in D} \frac{|p_{2i}-p_{1i}|}{p_{2i}} \\
&\leqslant \frac{1}{\sum_{i\in D} K(p_{1i}, p_{2i})} \frac{1}{n} \sum_{i\in D} \frac{|p_{2i}-p_{1i}|}{p_{2i}} \leqslant \frac{1}{n^{1-\beta}} \rightarrow 0.
\end{aligned}$$

On $D^c \cap B_{1l} \cap B_{2m}$, we use $K(p_{1i}, p_{2i}) \geqslant (\sqrt{p_{1i}} - \sqrt{p_{2i}})^2$ (see in Devroye et al., 1996, p. 131) and $\frac{p_{1i}}{p_{2i}} \geqslant 1 + M$ and $\frac{p_{2i}}{p_{1i}} \geqslant \frac{1}{1-M}$. For $p_{1i} p_{2i}$, $K(p_{1i}, p_{2i}) \geqslant p_{1i}(\sqrt{\frac{p_{2i}}{p_{1i}}} - 1)^2 \geqslant p_{1i}(\sqrt{\frac{1}{1-M_n}} - 1)^2 \geqslant n^{-\beta+(l-1)\varepsilon}(\sqrt{\frac{1}{1-M_n}} - 1)^2 \sim n^{-\beta+(l-1)\varepsilon} n^{-(2/3)(1-\beta-\varepsilon)}$. In the same way, for $p_{2i} > p_{1i}$, $K(p_{1i}, p_{2i}) \geqslant n^{-\beta+(m-1)\varepsilon}(\sqrt{M_n+1} - 1)^2 \sim n^{-\beta+(m-1)\varepsilon} n^{-(2/3)(1-\beta-\varepsilon)}$. Combining these, $K(p_{1i}, p_{2i}) \geqslant n^{-\beta+(\max(l,m)-1)\varepsilon} n^{-(2/3)(1-\beta-\varepsilon)}$for some constant M'.

$$\begin{aligned}
(\mathrm{II}) &= \frac{2}{nE\delta_d(X)} \sum_{i\in D^c\cap B_{1l}\cap B_{2m}} \frac{|p_{1i}-p_{2i}|}{p_{2i}} \\
&\leqslant \frac{2}{n^{-(2/3)(1-\beta-\varepsilon)}|D^c \cap B_{1l} \cap B_{2m}| n^{1-\beta+(\max(l,m)-1)\varepsilon}} \sum_{i\in D^c\cap B_{1l}\cap B_{2m}} \frac{\max(p_{1i}, p_{2i})}{n^{-\beta+(m-1)\varepsilon}} \\
&\leqslant \frac{2|D^c \cap B_{1l} \cap B_{2m}| n^{-\beta+\max(l,m)\varepsilon}}{n^{-(2/3)(1-\beta-\varepsilon)}|D^c \cap B_{1l} \cap B_{2m}| n^{1-\beta+(\max(l,m)-1)\varepsilon} n^{-\beta+(m-1)\varepsilon}} \\
&\leqslant \frac{2}{n^{(1/3)(1-\beta-\varepsilon)}} \rightarrow 0.
\end{aligned}$$

So $\frac{1}{nE\delta_d(X)}\sum_{i=1}^{d}\frac{|p_{1i}-p_{2i}|}{p_{2i}(1-p_{2i})} = \mathrm{o}(1)$. $E\hat{\delta}_d(X) = E\delta_d(X)(1+\mathrm{o}(1))$ follows from $E\hat{\delta}_d(X) = E\delta_d(X) + \frac{1}{n}\sum_{i=1}^{d}\frac{(2p_{1i}-1)(p_{2i}-p_{1i})}{p_{2i}(1-p_{2i})}(1+\mathrm{o}(1))$. □

Appendix C. Proof of Lemma 3.5

(1) When $\mathrm{var}(\delta_d(X))/n^{v-1}\to\infty$, $\mathrm{var}(\hat{\delta}_d(X)) = \mathrm{var}(\delta_d(X))(1+\mathrm{o}(1))$ by Lemma 3.3. Since $r_d \leqslant M$, $\frac{1}{\sqrt{\mathrm{var}(\delta_d(X))}} \leqslant \frac{M}{E\delta_d(X)}$. By using this, $|\tilde{r}_d - r_d| = \frac{|E\hat{\delta}_d(X)-E\delta_d(X)|}{\sqrt{\mathrm{var}(\delta_d(X))}}(1+\mathrm{o}(1)) \leqslant \frac{1}{n\sqrt{\mathrm{var}(\delta_d(X))}}\sum_{i=1}^{d}\frac{|p_{1i}-p_{2i}|}{p_{1i}(1-p_{1i})}(1+\mathrm{o}(1)) \leqslant \frac{M}{nE\delta_d(X)}\sum_{i=1}^{d}\frac{|p_{1i}-p_{2i}|}{p_{1i}(1-p_{1i})}(1+\mathrm{o}(1)) \to 0$ by Lemma 3.4.

(2) Since $\mathrm{var}(\hat{\delta}_d(X)) = \mathrm{var}(\delta_d(X))(1+\mathrm{o}(1))$, $\frac{\sum_{i=1}^{d}(\hat{\delta}_i(X)-E\hat{\delta}_i(X)-\delta_i(X)+E\delta_i(X))}{\sqrt{\mathrm{var}(\delta_d(X))}} = \frac{\sum_{i=1}^{d}(\hat{c}_i X_i+\hat{c}_{i0}-c_i X_i-c_{i0})}{\sqrt{\mathrm{var}(\delta_d(X))}} - \frac{E\hat{\delta}_d(X)-E\delta_d(X)}{\sqrt{\mathrm{var}(\delta_d(X))}}$. The second term is o(1) by (1) in this Lemma. We only need to show the first term is o(1). The first term is $\sum_{i=1}^{d}(\hat{c}_i X_i + \hat{c}_{i0} - c_i X_i - c_{i0}) = \sum_{i=1}^{d}(\hat{c}_i - c_i)(X_i - p_{1i}) + \sum_{i=1}^{d}(\hat{c}_i p_{1i} + \hat{c}_i - c_i p_{1i} - c_{i0})$. By Lemma A.4, $\frac{1}{\sqrt{\mathrm{var}(\delta_d(X))}}\sum_{i=1}^{d}(\hat{c}_i p_{1i} + \hat{c}_i - c_i p_{1i} - c_{i0}) = \mathrm{o_p}(1)$. By Lemma A.3, $\frac{1}{\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}E[(\hat{c}_i - c_i)^2(X_i - p_{1i})^2] = \frac{1}{\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}(\frac{1}{n} + \frac{p_{1i}(1-p_{1i})}{np_{2i}(1-p_{2i})})(1+\mathrm{o}(1)) = \mathrm{o}(1)$. $\sum_{i=1}^{d}E[(\hat{c}_i - c_i)(X_i - p_{1i})] = 0$ and $\frac{1}{\mathrm{var}(\delta_d(X))}\sum_{i=1}^{d}E[(\hat{c}_i - c_i)^2(X_i - p_{1i})^2] = \mathrm{o}(1)$ shows

$$\frac{1}{\sqrt{\mathrm{var}(\delta_d(X))}}\sum_{i=1}^{d}(\hat{c}_i - c_i)(X_i - p_{1i}) = \mathrm{o_p}(1).$$

These results prove $\tilde{N}_d - N_d \to 0$ in probability.

Appendix D. Proof of Lemma 3.6

(1) This is a direct consequence of the Lyapounov condition.

(2) We need to show that $\sum_{i=1}^{d}E|\hat{\delta}_i(X) - E\hat{\delta}_i(X)|^3/(\mathrm{var}(\hat{\delta}_d(X)))^{3/2} \to 0$. Let $\hat{\delta}_i(X) - E\hat{\delta}_i(X) = c_i(X_i - p_{1i}) + (\hat{c}_i - c_i)(X_i - p_{1i}) + \hat{c}_i p_{1i} + \hat{c}_{i0} - E\hat{\delta}_i(X) \equiv I_{1i} + I_{2i} + I_{3i}$. By the condition $\sum_{i=1}^{d}|c_i|^3 p_{1i}(1-p_{1i})/(\mathrm{var}(\hat{\delta}_d(X)))^{3/2} \leqslant \sum_{i=1}^{d}|c_i|^3 p_{1i}(1-p_{1i})/(\mathrm{var}(\delta_d(X)))^{3/2} \to 0$.

We will show that $\sum_{i=1}^{d}E|I_{2i}|^3/(\mathrm{var}(\hat{\delta}_d(X)))^{3/2} \to 0$. On A_d, $|\hat{c}_i - c_i| = |\mathbf{T}_i + \mathrm{o}(\mathbf{T}_i)| \leqslant 2n^{-\varepsilon^*}$ and $|\hat{p}_{ji} - p_{ji}| \leqslant n^{-\varepsilon^*}p_{ji}(1-p_{ji})$. Therefore,

$$\begin{aligned}\frac{1}{(\mathrm{var}(\hat{\delta}_d(X)))^{3/2}}\sum_{i=1}^{d}E|I_{2i}|^3 &= \frac{1}{(\mathrm{var}(\hat{\delta}_d(X)))^{3/2}}\sum_{i=1}^{d}E|I_{2i}|^3 I(A_d) + \sum_{i=1}^{d}E|I_{2i}|^3 I(A_d^c)\\ &\leqslant \frac{1}{(\mathrm{var}(\hat{\delta}_d(X)))^{3/2}}\sum_{i=1}^{d}E|\mathbf{T}_i + \mathrm{o}(\mathbf{T}_i)|^3|X_i - p_{1i}|^3 I(A_d) + n^2 P(A_d^c)\\ &\leqslant \frac{8n^{-3\varepsilon^*}}{(\mathrm{var}(\hat{\delta}_d(X)))^{3/2}}\sum_{i=1}^{d}\frac{p_{1i}(1-p_{1i})}{n} + \mathrm{o}(1)\\ &= \frac{8n^{-3\varepsilon^*}}{(\mathrm{var}(\delta_d(X)) + 2cn^{v-1}(1+\mathrm{o}(1)))^{3/2}}cn^{v-1} + \mathrm{o}(1) = \mathrm{o}(1).\end{aligned}$$

The last equality is due to $v \geqslant 1$.

For $\sum_{i=1}^{d} E|I_{3i}|^3/(\mathrm{var}(\hat{\delta}_d(X)))^{3/2} \to 0$, on A_d, $|I_{3i}| = |\hat{c}_i p_{1i} + \hat{c}_{i0} - c_i p_{1i} - c_{i0} + \frac{p_{2i}-p_{1i}}{np_{2i}(1-p_{2i})}| \leqslant |(\hat{c}_i - c_i)p_{1i}| + |\hat{c}_{i0} - c_{i0}| + |\frac{p_{2i}-p_{1i}}{np_{2i}(1-p_{2i})}| \leqslant 6n^{-\varepsilon^*}$. Using $\frac{1}{(\mathrm{var}(\hat{\delta}_d(X)))^{3/2}} \leqslant \frac{1}{\sqrt{2c}} \frac{1}{\mathrm{var}(\hat{\delta}_d(X))}$ and $\frac{1}{(\mathrm{var}(\hat{\delta}_d(X)))^{3/2}} \leqslant \frac{1}{2c^{3/2}}(1+\mathrm{o}(1))$,

$$\begin{aligned}
&\frac{1}{(\mathrm{var}(\hat{\delta}_d(X)))^{3/2}} \sum_{i=1}^{d} E|I_{3i}|^3 \\
&\leqslant \frac{6n^{-\varepsilon^*}}{\sqrt{2c}\,\mathrm{var}(\hat{\delta}_d(X))} \sum_{i=1}^{d} E[|I_{3i}|^2 I(A_d)] + \frac{1}{(\mathrm{var}(\hat{\delta}_d(X)))^{3/2}} \sum_{i=1}^{d} E[|I_{3i}|^3 I(A_d^c)] \\
&\leqslant \frac{6n^{-\varepsilon^*}}{\sqrt{2c}\,\mathrm{var}(\hat{\delta}_d(X))} \sum_{i=1}^{d} (\hat{c}_i p_{1i} + \hat{c}_{i0} - c_i p_{1i} - c_{i0})^2 \\
&\quad + \frac{6n^{-\varepsilon^*}}{(\mathrm{var}(\hat{\delta}_d(X)))^{3/2}} \frac{1}{n^2} \sum_{i=1}^{d} \frac{|p_{2i} - p_{1i}|^2}{(p_{2i}(1-p_{2i}))^2} + \frac{1}{2c^{3/2}} n^2 P(A_d^c) \\
&\equiv (\mathrm{I}) + (\mathrm{II}) + (\mathrm{III}).
\end{aligned}$$

By (2) in Lemma A.4, (I) converges to 0. By (3) in Lemma A.3, when $\mathrm{var}(\delta_d(X))/n^{\nu-1} \to \infty$, using $\frac{1}{(\mathrm{var}(\hat{\delta}_d(X)))^{3/2}} \leqslant \frac{1}{\sqrt{2c}} \frac{1}{\mathrm{var}(\delta_d(X))}$, (II) $\leqslant \frac{1}{\sqrt{2c}\,\mathrm{var}(\delta_d(X))n^2} \sum_{i=1}^{d} \frac{(p_{2i}-p_{1i})^2}{(p_{2i}(1-p_{2i}))^2} \leqslant \frac{1}{\sqrt{2c}\,\mathrm{var}(\delta_d(X))n^{1-\beta}} \frac{1}{n} \sum_{i=1}^{d} \frac{|p_{2i}-p_{1i}|}{p_{2i}(1-p_{2i})} \to 0$; when $\mathrm{var}(\delta_d(X)) = \mathrm{O}(n^{\nu-1})$, using $\frac{1}{(\mathrm{var}(\hat{\delta}_d(X)))^{3/2}} \leqslant \frac{1}{2c^{3/2}}(1+\mathrm{o}(1))$, the second term, (II) $\leqslant \frac{1}{2c^{3/2}n^2} \sum_{i=1}^{d} \frac{(p_{2i}-p_{1i})^2}{(p_{2i}(1-p_{2i}))^2} \leqslant \frac{1}{2c^{3/2}n^{1-\beta}} \frac{1}{n} \sum_{i=1}^{d} \frac{|p_{2i}-p_{1i}|}{p_{2i}(1-p_{2i})} \to 0$. So the second term converges to 0. By Lemma 3.1, the third term, (III), $n^2 P(A_d^c) \to 0$. Therefore, $\frac{1}{(\mathrm{var}(\hat{\delta}_d(X)))^{3/2}} \sum_{i=1}^{d} |\hat{\delta}_i(X) - E\hat{\delta}_d(X)|^3 \to 0$. □

References

Bickel, P.J., Levina, L., 2004. Some theory for Fisher's linear discriminant function, 'naive Bayes' and some alternatives when there are many more variables than observations. Bernoulli 10, 989–1910.

Billingsley, P., 1995. Probability and Measure. Wiley, New York.

Dawid, A.P., Fang, B.Q., 1992. Conjugate Bayes discrimination with infinitely many variables. J. Multivariate Anal. 41, 27–42.

Devroye, L., Gyorfi, L., Lugosi, G., 1996. A Probabilistic Theory of Pattern Recognition. Springer, Berlin.

Fang, B.G., Dawid, A.P., 1993. Asymptotic properties of conjugate bayes discrete discrimination. J. Multivariate Anal. 46, 83–96.

Ge, L., Simpson, D.G., 1998. Correlation and high-dimensional consistency in pattern recognition. J. Amer. Statist. Assoc. 93, 995–1006.

Greenshtein, E., Ritov, R., 2004. Persistence in high-dimensional linear predictor selection and the virtue of over parametrization. Bernoulli 10, 971–988.

Gyllenberg, M., Koski, T., 2001. Probabilistic models for bacterial taxonomy. Internat. Statist. Rev. 69, 799–821.

Hand, D.J., 1981. Discrimination and Classification. Wiley, New York.

Hand, D.J., Yu, K., 2001. Idiot's Bayes—not so stupid after all? Internat. Statist. Rev. 69, 385–398.

Huber, P.J., 1973. Robust regression: asymptotics, conjectures and Monte Carlo. Ann. Statist. 1, 799–821.

Portnoy, S., 1984. Asymptotic behavior of m-estimators of p regression parameters when p^2/n is large. I. Consistency. Ann. Statist. 12, 1289–1309.

Portnoy, S., 1985. Asymptotic behavior of m-estimators of p regression parameters when p^2/n is large. II. Normal approximation. Ann. Statist. 13, 1403–1417.

Wilbur, J.D., Ghosh, J.K., Nakatsu, C.H., Brouder, S.M., Doerge, R.W., 2002. Variable selection in high-dimensional multivariate binary data with application to the analysis of microbial DNA fingerprints. Biometrics 58, 378–386.