Journal of Machine Learning Research 14 (2013) 3011-3040 Submitted 11/12; Revised 9/13; Published 10/13

A Plug-in Approach to Neyman-Pearson Classification

Xin Tong XINT@MARSHALL.USC.EDU
Marshall Business School
University of Southern California
Los Angeles, CA 90089, USA

Editor: John Shawe-Taylor

Abstract

The Neyman-Pearson (NP) paradigm in binary classification treats type I and type II errors with different priorities. It seeks classifiers that minimize type II error, subject to a type I error constraint under a user specified level α. In this paper, plug-in classifiers are developed under the NP paradigm. Based on the fundamental Neyman-Pearson Lemma, we propose two related plug-in classifiers which amount to thresholding respectively the class conditional density ratio and the regression function. These two classifiers handle different sampling schemes. This work focuses on theoretical properties of the proposed classifiers; in particular, we derive oracle inequalities that can be viewed as finite sample versions of risk bounds. NP classification can be used to address anomaly detection problems, where asymmetry in errors is an intrinsic property. As opposed to a common practice in anomaly detection that consists of thresholding normal class density, our approach does not assume a specific form for anomaly distributions. Such consideration is particularly necessary when the anomaly class density is far from uniformly distributed.

Keywords: plug-in approach, Neyman-Pearson paradigm, nonparametric statistics, oracle inequality, anomaly detection

1. Introduction

Classification aims to identify which category a new observation belongs to, on the basis of labeled training data. Applications include disease classification using high-throughput data such as microarrays, SNPs, spam detection and image recognition. This work investigates Neyman-Pearson paradigm in classification with a plug-in approach.

1.1 Neyman-Pearson Paradigm

The Neyman-Pearson (NP) paradigm extends the objective of classical binary classification in that, while the latter focuses on minimizing classification error that is a weighted sum of type I and type II errors, the former minimizes type II error subject to an upper bound α on type I error, where the threshold level α is chosen by the user. The NP paradigm is appropriate in many applications where it is necessary to bring down one kind of error at the expense of the other. One example is medical diagnosis: failing to detect a malignant tumor leads to loss of a life, while flagging a benign one only induces some unnecessary medical cost. As healthy living and longer life expectancy cannot be compensated by any amount of money, it is desirable to control the false negative rate of any medical diagnosis, perhaps with some sacrifice in the false positive rate.

A few commonly used notations in classification literature are set up to facilitate our discussion. Let (X,Y) be a random couple where $X \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of covariates, and where $Y \in \{0,1\}$ is a label that indicates to which class X belongs. A *classifier* h is a mapping $h : \mathcal{X} \to \{0,1\}$ that returns the predicted class given X. An error occurs when $h(X) \neq Y$. It is therefore natural to define the classification loss by $\mathbb{1}(h(X) \neq Y)$, where $\mathbb{1}(\cdot)$ denotes the indicator function. The expectation of the classification loss with respect to the joint distribution of (X,Y) is called *classification risk (error)* and is defined by

$$R(h) = \mathbb{P}\left(h(X) \neq Y\right).$$

The risk function can be expressed as a convex combination of type I and II errors:

$$R(h) = \mathbb{P}(Y=0)R_0(h) + \mathbb{P}(Y=1)R_1(h), \tag{1}$$

where

$$R_0(h) = \mathbb{P}\left(h(X) \neq Y | Y=0\right) \text{denotes the type I error,}$$
$$R_1(h) = \mathbb{P}\left(h(X) \neq Y | Y=1\right) \text{denotes the type II error.}$$

Also recall that the regression function of Y on X is defined by

$$\eta(x) = \mathbb{E}[Y|X=x] = \mathbb{P}(Y=1|X=x).$$

Let $h^*(x) = \mathbb{1}(\eta(x) \geq 1/2)$. The oracle classifier h^* is named the Bayes classifier, and it achieves the minimum risk among all possible candidate classifiers. The risk of h^*, $R^* = R(h^*)$ is called the Bayes risk. A certain classifier $\hat{h}$ in classical binary classification paradigm is good if the excess risk $R(\hat{h}) - R^*$ is small on average or with high probability.

In contrast to the classical paradigm, the NP classification seeks a minimizer ϕ^* that solves

$$\min_{R_0(\phi) \leq \alpha} R_1(\phi),$$

where a small α (e.g., 5%) reflects very conservative attitude towards type I error.

The NP paradigm is irrelevant if we can achieve very small type I and type II errors simultaneously. This is often impossible as expected, and we will demonstrate this point with a stylized example. Note that for most joint distributions on (X,Y), the Bayes error R^* is well above zero. Suppose in a tumor detection application, $R^* = 10\%$. Clearly by (1), it is not feasible to have both type I error R_0 and type II error R_1 be smaller than 10%. Since we insist on lowering the false negative rate as our priority, with a desirable false negative rate much lower than 10%, we have to sacrifice some false positive rate.

Moreover, even if a classifier $\hat{\phi}$ achieves a small risk, there is no guarantee on attaining desirable type I or type II errors. Take another stylized example in medical diagnosis. Suppose that type I error equals 0.5, that is, with 50% of the chances, detector $\hat{\phi}$ fails to find the malignant tumor, and that type II error equals 0.01. Also assume the chance that a tumor is malignant is only 0.001. Then the risk of $\hat{\phi}$ is approximately 1%. This is low, but $\hat{\phi}$ is by no means a good detector, because it misses a malignant tumor with half of the chances!

Empirical risk minimization (ERM), a common approach to classification, has been studied in the NP classification literature. Cannon et al. (2002) initiated the theoretical treatment of the

NP classification paradigm and an early empirical study can be found in Casasent and Chen (2003). Several results for traditional statistical learning such as PAC bounds or oracle inequalities have been studied in Scott (2005) and Scott and Nowak (2005) in the same framework as the one laid down by Cannon et al. (2002). Scott (2007) proposed performance measures for NP classification that weights type I and type II error in sensible ways. More recently, Blanchard et al. (2010) developed a general solution to semi-supervised novelty detection by reducing it NP classification, and Han et al. (2008) transposed several earlier results to NP classification with convex loss. There is a commonality in this line of literature: a relaxed empirical type I error constraint is used in the optimization program, and as a result, type I errors of the classifiers can only be shown to satisfy a relaxed upper bound. Take the framework set up by Cannon et al. (2002) for example: for some $\varepsilon_0 > 0$ and let $\mathcal{H}$ be a set of classifiers with finite VC dimension. They proposed the program

$$\min_{\phi \in \mathcal{H}, \hat{R}_0(\phi) \leq \alpha + \varepsilon_0/2} \hat{R}_1(\phi),$$

where $\hat{R}_0$ and $\hat{R}_1$ denote empirical type I and type II errors respectively. It is shown that solution to the above program $\hat{\phi}$ satisfies simultaneously with high probability, the type II error $R_1(\hat{\phi})$ is bounded from above by $R_1(\phi^*) + \varepsilon_1$, for some $\varepsilon_1 > 0$, and the type I error $R_0(\hat{\phi})$ is bounded from above by $\alpha + \varepsilon_0$.

However, following the original spirit of NP classification, a good classifier $\hat{\phi}$ should respect the chosen significance level α, rather than some relaxation of α, that is, we should be able to i). satisfy the type I error constraint $R_0(\hat{\phi}) \leq \alpha$ with high probability, while ii). establishing an explicit diminishing rate for the excess type II error $R_1(\hat{\phi}) - R_1(\phi^*)$. The simultaneous achievements of i). and ii). can be thought of as counterpart of oracle inequality in classical binary classification, and we believe they are a desirable formulation of theoretical properties of good classifiers in NP classification. Considering this point, Rigollet and Tong (2011) propose a computationally feasible classifier $\tilde{h}^\tau$, such that φ-type I error of $\tilde{h}^\tau$ is bounded from above by α with high probability and the excess φ-type II error of $\tilde{h}^\tau$ converges to 0 with explicit rates, where φ-type I error and φ-type II error are standard convex relaxations of type I and type II errors respectively. Most related to the current context, they also proved a negative result. Loosely speaking, it is shown by counter examples that under the original type I/II criteria, if one adopts ERM approaches (convexification or not), one cannot guarantee diminishing excess type II error if one insists type I error of the proposed classifier be bounded from above by α with high probability. Interested readers are referred to Section 4.4 of that paper.

In this work, we will fulfill the original NP paradigm spirit with the plug-in approach. Theoretical properties of the classifiers under the NP paradigm will be derived. To the best of our knowledge, our paper is the first to do so. It looks as if from a theoretical point of view, a plug-in approach is more suitable than ERM for the NP paradigm. However, such a comparison is not fair because the two approaches are based on different sets of assumptions. For the ERM approach, the main assumption is on the complexity of candidate classifiers, leaving the class conditional distributions unrestricted. While with the plug-in approach, we put restrictions on the joint distributions.

A related framework that also addresses asymmetry in errors is the cost-sensitive learning, which assigns different costs as weights of type I and type II errors (see, e.g., Elkan 2001, Zadrozny et al. 2003). This approach has many practical values, but when it is hard to assign costs to errors, or in applications such as medical diagnosis, where it is morally inappropriate to do the usual cost and benefit analysis, the NP paradigm is a natural choice.

1.2 Plug-in Approach Based on the Fundamental Neyman-Pearson Lemma

NP classification is closely related to the NP approach to statistical hypothesis testing. The punch line is that the fundamental Neyman-Pearson lemma itself suggests a direct plug-in classifier. The interested reader is referred to Lehmann and Romano (2005) for a comprehensive treatment of hypothesis testing. Here we only review the central knowledge that brings up this connection.

Hypothesis testing bears strong resemblance with binary classification if we assume the following model. Let P^- and P^+ be two *known* probability distributions on $\mathcal{X} \subset \mathbb{R}^d$. Let $\pi \in (0,1)$ and assume that Y is a random variable defined by

$$Y = \begin{cases} 1 & \text{with probability } \pi, \\ 0 & \text{with probability } 1-\pi. \end{cases}$$

Assume further that the conditional distribution of X given Y is denoted by P^{2Y-1}. Given such a model, the goal of statistical hypothesis testing is to determine whether X was generated from P^- or from P^+. To that end, we construct a randomized test $\phi : \mathcal{X} \to [0,1]$ and the conclusion of the test based on ϕ is that X is generated from P^+ with probability $\phi(X)$ and from P^- with probability $1-\phi(X)$. Note that randomness here comes from an exogenous randomization process such as flipping a biased coin. Two kinds of errors arise: type I error occurs when rejecting P^- when it is true, and type II error occurs when not rejecting P^- when it is false. The Neyman-Pearson paradigm in hypothesis testing amounts to choosing ϕ that solves the following constrained optimization problem

$$\begin{array}{ll} \text{maximize} & \mathbb{E}[\phi(X)|Y=1], \\ \text{subject to} & \mathbb{E}[\phi(X)|Y=0] \le \alpha, \end{array}$$

where $\alpha \in (0,1)$ is the significance level of the test. In other words, we specify a significance level α on type I error, and minimize type II error. We call a solution to this constrained optimization problem *a most powerful test* of level α. The Neyman-Pearson Lemma gives mild sufficient conditions for the existence of such a test.

Theorem 1 (Neyman-Pearson Lemma) *Let P^- and P^+ be probability distributions possessing densities p_0 and p_1 respectively with respect to some measure μ. Let $f_{C_\alpha}(x) = \mathbb{I}(L(x) \ge C_\alpha)$, where $L(x) = p_1(x)/p_0(x)$ and C_α is such that $P^-(L(X) > C_\alpha) \le \alpha$ and $P^-(L(X) \ge C_\alpha) \ge \alpha$. Then,*

- *f_{C_α} is a level $\alpha = \mathbb{E}[f_{C_\alpha}(X)|Y=0]$ most powerful test.*
- *For a given level α, the most powerful test of level α is defined by*

$$\phi(X) = \begin{cases} 1 & \text{if} \quad L(X) > C_\alpha \\ 0 & \text{if} \quad L(X) < C_\alpha \\ \frac{\alpha - P^-(L(X) > C_\alpha)}{P^-(L(X) = C_\alpha)} & \text{if} \quad L(X) = C_\alpha. \end{cases}$$

Notice that in the learning framework, ϕ cannot be computed since it requires knowledge of the distributions P^- and P^+. Nevertheless, the Neyman-Pearson Lemma motivates a plug-in classifier. Concretely, although we do not know p_1 and p_0, we can find the kernel density estimators $\hat{p}_1$ and $\hat{p}_0$ based on data. Then if we can also detect the approximately right threshold level $\hat{C}_\alpha$, the plug-in approach leads to a classifier $\mathbb{I}(\hat{p}_1(x)/\hat{p}_0(x) \ge \hat{C}_\alpha)$. We expect that this simple classifier would have good type I/II performance bounds, and this intuition will be verified in the following sections. It

is worthy to note that our plug-in approach to NP classification leads to problems related to density level set estimation (see Rigollet and Vert 2009 and reference therein), where the task is to estimate $\{x : p(x) > \lambda\}$, for some level $\lambda > 0$. Density level set estimation has applications in anomaly detection and unsupervised or semi-supervised classification. Plug-in methods for density level set estimation, as opposed to direct methods, do not involve complex optimization procedure, and only amounts to thresholding the density estimate at proper level. The challenges in our setting different from Rigollet and Vert (2009) are two folds. First, the threshold level in our current setup needs to be estimated, and secondly, we deal with density ratios rather than densities. Plug-in methods in classical binary classification have been also studied in the literature. Earlier works seemed to give rise to pessimism of plug-in approach to classification. For example, under certain assumptions, Yang (1999) showed plug-in estimators cannot achieve classification error faster than $O(1/\sqrt{n})$. But direct methods can achieve fast rates up to $O(1/n)$ under *margin assumption* (Mammen and Tsybakov, 1999; Tsybakov, 2004; Tsybakov and van de Geer, 2005; Tarigan and van de Geer, 2006). However Audibert and Tsybakov (2007) combined a smoothness condition condition on regression function with the margin assumption, and showed that plug-in classifiers $\mathbb{I}(\hat{\eta}_n \geq 1/2)$ based on local polynomial estimators can achieve rates faster than $O(1/n)$. We will borrow the smoothness condition on the regression function and margin assumption from Audibert and Tsybakov (2007). However in that paper again, the threshold level is not estimated, so new techniques are called for.

1.3 Application to Anomaly Detection

NP classification is a useful framework to address anomaly detection problems. In anomaly detection, the goal is to discover patterns that are different from usual outcomes or behaviors. An unusual behavior is named an *anomaly*. A variety of problems, such as credit card fraud detection, insider trading detection and system malfunctioning diagnosis, fall into this category. There are many approaches to anomaly detection; some serving a specific purpose while others are more generic. Modeling techniques include classification, clustering, nearest neighbors, statistical and spectrum, etc. A recent comprehensive review of anomaly detection literature is provided by Chandola et al. (2009). Earlier review papers include Agyemang et al. (2006), Hodge and Austin (2004), Markou and Singh (2003a), Markou and Singh (2003b), Patcha and Park (2007), etc.

When we have training data from the normal class, a common approach to anomaly detection is to estimate the normal class density p_0 and try to threshold at a proper level, but this is inappropriate if the anomaly class is far from uniformly distributed. Indeed, to decide whether a certain point is an anomaly, one should consider how likely it is for this point to be normal as opposed to abnormal. The likelihood ratio p_1/p_0 or the regression function η are good to formalize such a concern. Our main results in NP classification will be adapted for anomaly detection applications, where the normal sample size n is much bigger than the anomaly sample size m.

The rest of the paper is organized as follows. In Section 2, we introduce a few notations and definitions. In Section 3, oracle inequalities for a direct plug-in classifier are derived based on the density ratio p_1/p_0. Section 4 investigates another related plug-in classifier, which targets on the regression function η. Finally, proofs of two important technical lemmas are relegated to the Appendix.

2. Notations and Definitions

Following Audibert and Tsybakov (2007), some notations are introduced. For any multi-index $s = (s_1, \ldots, s_d) \in \mathbb{N}^d$ and any $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, define $|s| = \sum_{i=1}^d s_i$, $s! = s_1! \cdots s_d!$, $x^s = x_1^{s_1} \cdots x_d^{s_d}$ and $\|x\| = (x_1^2 + \cdots + x_d^2)^{1/2}$. Let D^s be the differential operator $D^s = \frac{\partial^{s_1+\cdots+s_d}}{\partial x_1^{s_1} \ldots \partial x_d^{s_d}}$.

Let $\beta > 0$. Denote by $\lfloor \beta \rfloor$ the largest integer strictly less than β. For any $x, x' \in \mathbb{R}^d$ and any $\lfloor \beta \rfloor$ times continuously differentiable real valued function g on $\mathbb{R}^d$, we denote by g_x its Taylor polynomial of degree $\lfloor \beta \rfloor$ at point x:

$$g_x(x') = \sum_{|s| \le \lfloor \beta \rfloor} \frac{(x'-x)^s}{s!} D^s g(x) .$$

For $L > 0$, the $(\beta, L, [-1,1]^d)$-Hölder class of functions, denoted by $\Sigma(\beta, L, [-1,1]^d)$, is the set of functions $g : \mathbb{R}^d \to \mathbb{R}$ that are $\lfloor \beta \rfloor$ times continuously differentiable and satisfy, for any $x, x' \in [-1,1]^d$, the inequality:

$$|g(x') - g_x(x')| \le L\|x - x'\|^\beta .$$

The $(\beta, L, [-1,1]^d)$-Hölder class of density is defined as

$$\mathcal{P}_\Sigma(\beta, L, [-1,1]^d) = \left\{ p : p \ge 0, \int p = 1, p \in \Sigma(\beta, L, [-1,1]^d) \right\} .$$

Denote respectively by $\mathbb{P}$ and $\mathbb{E}$ generic probability distribution and expectation. Also recall that we have denoted by p_0 the density of class 0 and by p_1 that of class 1. For all the theoretical discussions in this paper, the domain of densities p_0 and p_1 is $[-1,1]^d$.

We will use β-valid kernels throughout the paper, which are a multi-dimensional analog of univariate higher order kernels. The definition of β-valid kernels is as follows

Definition 1 *Let K be a real-valued function on $\mathbb{R}^d$ with support $[-1,1]^d$. For fixed $\beta > 0$, the function $K(\cdot)$ is a β-valid kernel if it satisfies $\int K = 1$, $\int |K|^p < \infty$ for any $p \ge 1$, $\int \|t\|^\beta |K(t)| dt < \infty$, and in the case $\lfloor \beta \rfloor \ge 1$, it satisfies $\int t^s K(t) dt = 0$ for any $s = (s_1, \ldots, s_d) \in \mathbb{N}^d$ such that $1 \le s_1 + \ldots + s_d \le \lfloor \beta \rfloor$.*

One example of β-valid kernels is the product kernel whose ingredients are kernels of order β in 1 dimension:

$$\tilde{K}(x) = K(x_1)K(x_2)\cdots K(x_d)\, \mathbb{I}(x \in [-1,1]^d),$$

where K is a 1-dimensional β-valid kernel and is constructed based on Legendre polynomials. We refer interested readers to Section 1.2.2 of Tsybakov (2009). These kernels have been considered in the literature, such as Rigollet and Vert (2009). When the β-valid kernel K is constructed out of Legendre polynomials, it is also Lipschitz and bounded. Therefore, such a kernel satisfies conditions for Lemma 1. For simplicity, we assume that all the β-valid kernels considered in this paper are constructed from Legendre polynomials.

The next low noise condition helps characterize the difficulty of a classification problem.

Definition 2 (Margin Assumption) *A function p satisfies the margin assumption of order $\bar{\gamma}$ with respect to probability distribution P at the level C^* if there exist positive constants C_0 and $\bar{\gamma}$, such that $\forall \delta \ge 0$,*

$$P(|p(X) - C^*| \le \delta) \le C_0 \delta^{\bar{\gamma}}.$$

The above condition for densities was first introduced in Polonik (1995), and its counterpart in the classical binary classification was called margin condition (Mammen and Tsybakov, 1999), from which we borrow the same terminology for discussion. A classification problem is less noisy by requiring most data be further away from the optimal decision boundary. Recall that the set $\{x : \eta(x) = 1/2\}$ is the decision boundary of the Bayes classifier in the classical paradigm, and the margin condition in the classical paradigm is a special case of Definition 2 by taking $p = \eta$ and $C^* = 1/2$.

3. Plug-in Based on Ratio of Class Conditional Densities

In this section, we investigate a plug-in classifier motivated by the Neyman-Pearson Lemma based on the density ratio p_1/p_0. Both the p_0 known and the p_0 unknown cases will be discussed. Although assuming precise knowledge on class 0 density is far from realistic, the subtlety of the plug-in approach in the NP paradigm, as opposed to in the classical paradigm, is revealed through the comparison of the two cases. Most importantly, we formulate some *detection condition* to detect the right threshold level in plug-in classifiers under the NP paradigm.

3.1 Class 0 Density p_0 Known

In this subsection, suppose that we know the class 0 density p_0, but have to estimate the class 1 density p_1. It is interesting to note that this setup is essentially a dual of generalized quantile (minimum volume) set estimation problems, where the volume and mass defining measures are interchanged. Denote by $\hat{p}_1$ the kernel density estimator of p_1 based on an i.i.d. class 1 sample $\mathcal{S}_1 = \{X_1^+, \ldots, X_m^+\}$, that is,

$$\hat{p}_1(x_0) = \frac{1}{mh^d} \sum_{i=1}^{m} K\left(\frac{X_i^+ - x_0}{h}\right),$$

where h is the bandwidth. For a given level α, define respectively $\hat{C}_\alpha$ and C_α^* as solutions of

$$P_0\left(\frac{\hat{p}_1(X)}{p_0(X)} \geq \hat{C}_\alpha\right) = \alpha \qquad \text{and} \qquad P_0\left(\frac{p_1(X)}{p_0(X)} \geq C_\alpha^*\right) = \alpha\,.$$

Note that for some α, $\hat{C}_\alpha$ and C_α^* might not exist. In such cases, randomization is needed to achieve the exact level α. For simplicity, we assume that $\hat{C}_\alpha$ and C_α^* exist and are unique. Note that since p_0 is known, the threshold $\hat{C}_\alpha$ is detected precisely for each sample $\mathcal{S}_1$. The Neyman-Pearson Lemma says that under mild regularity conditions, $\phi^*(x) = \mathbb{I}(p_1(x)/p_0(x) \geq C_\alpha^*)$ is the most powerful test of level α. Therefore, we have a plug-in classifier naturally motivated by the Neyman-Pearson Lemma:

$$\hat{\phi}(x) = \mathbb{I}\left(\frac{\hat{p}_1(x)}{p_0(x)} \geq \hat{C}_\alpha\right), \tag{2}$$

where we plug in estimates $\hat{p}_1$ and $\hat{C}_\alpha$ respectively for the class 1 density p_1 and the threshold level C_α^*. We are interested in the theoretical properties of $\hat{\phi}$. In particular, we will establish oracle inequalities regarding the excess type I and type II errors. Note that since $\hat{C}_\alpha$ is constructed to meet the level α exactly, the excess type I error of $\hat{\phi}$ vanishes, that is,

$$R_0(\hat{\phi}) - R_0(\phi^*) = 0\,.$$

We summarize as follows assumptions on class conditional densities that we will reply upon.

Condition 1 *Suppose that the class conditional densities p_0 and p_1 satisfy:*

i) There exists a positive constant $\mu_{\min}$, such that $p_0 \geq \mu_{\min}$,

ii) The class 1 density $p_1 \in \mathcal{P}_\Sigma(\beta, L, [-1,1]^d)$,

iii) The ratio of class conditional densities p_1/p_0 satisfies the margin assumption of order $\bar{\gamma}$ with respect to probability distribution P_0 at the level C^*_α.

Note that part i) in Condition 1 is the same as assuming $p_0 > 0$ on the compact domain $[-1,1]^d$, as long as p_0 is continuous. Part ii) is a global smoothness condition, which is stronger than the local smoothness conditions used in Rigollet and Vert (2009), in which different smoothness conditions for a neighborhood around the interested level λ and for the complement of the neighborhood are formulated. Rigollet and Vert (2009) emphasized on the smoothness property for a neighborhood around level λ, as only this part affects the rate of convergence. However, as $\hat{C}_\alpha$ is not known a priori in our setup, we rely upon a global smoothness condition as opposed to a local one.

The following theorem addresses the excess type II error of $\hat{\phi}$: $R_1(\hat{\phi}) - R_1(\phi^*)$.

Proposition 1 *Let $\hat{\phi}$ be the plug-in classifier defined by (2). Assume that the class conditional densities p_0 and p_1 satisfy the Condition 1 and that the kernel K is β-valid and L'-Lipschitz. Then for any $\delta \in (0,1)$, and any class 1 sample size m is such that $\sqrt{\frac{\log(m/\delta)}{mh^d}} < 1$, where the bandwidth $h = (\frac{\log m}{m})^{1/(2\beta+d)}$, the excess type II error is bounded, with probability $1-\delta$, by*

$$R_1(\hat{\phi}) - R_1(\phi^*) \leq \frac{2^{2+\bar{\gamma}} C_0 C^{1+\bar{\gamma}}}{(\mu_{\min})^{1+\bar{\gamma}}} \left(\frac{\log(m/\delta)}{mh^d} \right)^{\frac{1+\bar{\gamma}}{2}},$$

where the constant C is the same as in Lemma 1 applied to density p_1. In particular, there exists a positive $\bar{C}$, such that for any $m \geq 1/\delta$,

$$R_1(\hat{\phi}) - R_1(\phi^*) \leq \bar{C} \left(\frac{\log m}{m} \right)^{\frac{\beta(1+\bar{\gamma})}{2\beta+d}}.$$

Note that the dependency of the upper bound for the excess type II error on parameters β, L, and L' is incorporated into the constant C, whose explicit formula is given in Lemma 1, which has an important role in the proof. Lemma 1 is a finite sample uniform deviation result on kernel density estimators. Here we digress slightly and remark that theoretical properties of kernel density estimators have been studied intensively in the literature. A result of similar flavor was obtained in Lei et al. (2013). Readers are referred to Wied and Weiβbach (2010) and references therein for a survey on consistency of kernel density estimators. Convergence in distribution for weighted sup norms was derived in Giné et al. (2004). Lepski (2013) studied expected sup-norm loss of multivariate density estimation with an oracle approach. We have the technical Lemma 1 and it proof in the appendix, as none of previous results is tailored to our use. Another phenomenon worth mentioning is that the upper bound does not explicitly depend on the significance level α. This results from the way we formulate the margin assumption. Suppose we were to allow $\bar{\gamma}$ in the margin assumption to depend on α, that is, $\bar{\gamma} = \bar{\gamma}(\alpha)$, or let C_0 depend on α, the upper bound would have explicit dependency on α. Also from the upper bound, we can see that the larger the parameter $\bar{\gamma}$, the sharper the margin assumption, and then the faster the rate of convergence for the

excess type II error. Also, we re-emphasize that the feature dimension d considered in this paper is fixed and does not increase with sample sizes.

Proof

First note that the excess type II error can be represented by

$$R_1(\hat{\phi}) - R_1(\phi^*) = \int_{G^* \triangle \hat{G}} \left| \frac{p_1}{p_0} - C^*_\alpha \right| dP_0 \,,$$

where $G^* = \left\{ \frac{p_1}{p_0} < C^*_\alpha \right\}$ and $\hat{G} = \left\{ \frac{\hat{p}_1}{p_0} < \hat{C}_\alpha \right\}$, and $G^* \triangle \hat{G} = (G^* \cap \hat{G}^c) \cup (G^{*c} \cap \hat{G})$ is the symmetric difference between G^* and $\hat{G}$. Indeed,

$$\begin{aligned}
& \int_{G^* \triangle \hat{G}} \left| \frac{p_1}{p_0} - C^*_\alpha \right| dP_0 \\
= & \int_{G^* \cap \hat{G}^c} \left(C^*_\alpha - \frac{p_1}{p_0} \right) dP_0 + \int_{G^{*c} \cap \hat{G}} \left(\frac{p_1}{p_0} - C^*_\alpha \right) dP_0 \\
= & \int_{G^*} \left(C^*_\alpha - \frac{p_1}{p_0} \right) dP_0 + \int_{\hat{G}} \left(\frac{p_1}{p_0} - C^*_\alpha \right) dP_0 \\
= & C^*_\alpha P_0(G^*) - P_1(G^*) - C^*_\alpha P_0(\hat{G}) + P_1(\hat{G}) \\
= & P_1(\hat{G}) - P_1(G^*) \,.
\end{aligned}$$

Define an event regarding the sample $\mathcal{S}_1$: $\mathcal{E} = \{ \|\hat{p}_1 - p_1\|_\infty < \frac{\delta_1}{2} \mu_{\min} \}$, where $\delta_1 = \frac{2C}{\mu_{\min}} \sqrt{\frac{\log(m/\delta)}{mh^d}}$, and C is the same as in Lemma 1 (with p replaced by p_1). From this point to the end of the proof, we restrict ourselves to the event $\mathcal{E}$.

Since $G^{*c} \cap \hat{G}$ and $G^* \cap \hat{G}^c$ are disjoint, we can handle the two parts separately. Decompose

$$G^{*c} \cap \hat{G} = \left\{ \frac{p_1}{p_0} \geq C^*_\alpha, \frac{\hat{p}_1}{p_0} < \hat{C}_\alpha \right\} = A_1 \cup A_2 \,,$$

where

$$A_1 = \left\{ C^*_\alpha + \delta_1 \geq \frac{p_1}{p_0} \geq C^*_\alpha, \frac{\hat{p}_1}{p_0} < \hat{C}_\alpha \right\} ,$$

and

$$A_2 = \left\{ \frac{p_1}{p_0} > C^*_\alpha + \delta_1, \frac{\hat{p}_1}{p_0} < \hat{C}_\alpha \right\} .$$

Then,

$$\int_{A_1} \left(\frac{p_1}{p_0} - C^*_\alpha \right) dP_0 \leq \delta_1 P_0(A_1) \leq C_0 (\delta_1)^{1+\bar{\gamma}} .$$

We can control the distance between $\hat{C}_\alpha$ and C^*_α. Indeed,

$$\begin{aligned}
\alpha &= P_0 \left(\frac{\hat{p}_1(X)}{p_0(X)} \geq \hat{C}_\alpha \right) = P_0 \left(\frac{p_1(X)}{p_0(X)} \geq C^*_\alpha \right) \geq P_0 \left(\frac{\hat{p}_1(X)}{p_0(X)} \geq C^*_\alpha + \frac{|p_1(X) - \hat{p}_1(X)|}{p_0(X)} \right) \\
&\geq P_0 \left(\frac{\hat{p}_1(X)}{p_0(X)} \geq C^*_\alpha + \frac{\delta_1 \mu_{\min}}{2\mu_{\min}} \right) = P_0 \left(\frac{\hat{p}_1(X)}{p_0(X)} \geq C^*_\alpha + \frac{\delta_1}{2} \right) .
\end{aligned}$$

This implies that $\hat{C}_\alpha < C^*_\alpha + \frac{\delta_1}{2}$. Therefore,

$$A_2 \subset A_3 := \left\{ \frac{p_1}{p_0} \geq C^*_\alpha + \delta_1, \frac{\hat{p}_1}{p_0} < C^*_\alpha + \delta_1/2 \right\}.$$

It is clear that on the event $\mathcal{E}$, $P_0(A_3) = 0$. Therefore,

$$\int_{G^{*c} \cap \hat{G}} \left(\frac{p_1}{p_0} - C^*_\alpha \right) dP_0 \leq C_0 (\delta_1)^{1+\bar{\gamma}}.$$

Similarly, it can be shown that $\int_{G^* \cap \hat{G}^c} \left(C^*_\alpha - \frac{p_1}{p_0} \right) dP_0 \leq C_0 (\delta_1)^{1+\bar{\gamma}}$. Therefore,

$$\int_{G^* \triangle \hat{G}} \left| \frac{p_1}{p_0} - C^*_\alpha \right| dP_0 \leq 2C_0 (\delta_1)^{1+\bar{\gamma}}.$$

Finally, Lemma 1 implies $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$. This completes the proof.

■

Although it is reasonable to assume that the class 0 density p_0 can be approximated well, assuming p_0 known exactly is not realistic for most applications. Next, we consider the unknown p_0 case.

3.2 Class 0 Density p_0 Unknown

Assume that both the class 0 density p_0 and the class 1 density p_1 are unknown. Knowledge on class conditional densities is passed to us through samples. Because data from class 0 is needed to estimate both the class 0 density and the threshold level, we split the class 0 data into two pieces. Therefore, suppose available data include class 0 samples $\mathcal{S}_0 = \left\{ X_1^-, \ldots, X_n^- \right\}$, $\tilde{\mathcal{S}}_0 = \left\{ X_{n+1}^-, \ldots, X_{2n}^- \right\}$, and a class 1 sample $\mathcal{S}_1 = \left\{ X_1^+, \ldots, X_m^+ \right\}$. Also assume that given samples $\mathcal{S}_0$ and $\mathcal{S}_1$, the variables in $\tilde{\mathcal{S}}_0$ are independent. Our mission is still to construct a plug-in classifier based on the optimal test output by the Neyman-Pearson Lemma and to show that it has desirable theoretical properties regarding type I and II errors.

First estimate p_0 and p_1 respectively from $\mathcal{S}_0$ and $\mathcal{S}_1$ by kernel estimators,

$$\hat{p}_0(x) = \frac{1}{n h_n^d} \sum_{i=1}^{n} K \left(\frac{X_i^- - x}{h_n} \right) \qquad \text{and} \qquad \hat{p}_1(x) = \frac{1}{m h_m^d} \sum_{i=1}^{m} K \left(\frac{X_i^+ - x}{h_m} \right),$$

where h_n and h_m denote the bandwidths. Since p_0 is unknown, $\hat{C}_\alpha$ can not be defined trivially as in the p_0 known case. And, it turns out that detecting the right threshold level is important to proving theoretical properties of the plug-in classifier. There is one essential piece of intuition. We know that having fast diminishing excess type II error demands a low noise condition, such as the margin assumption. On the other hand, if there are enough sample points around the optimal threshold level, we can approximate the threshold C^*_α accurately. Approximating the optimal threshold level is not a problem in the classical setting, because in that setting, the Bayes classifier is $\mathbb{I}(\eta(x) \geq 1/2)$, and the threshold level $1/2$ on the regression function η is known. Therefore, estimating the optimal threshold with the NP paradigm introduces new technical challenges. The following level α detection condition addresses this concern.

Condition 2 (level α detection condition) *The function f satisfies the level α detection condition (with respect to P_0 ($X \sim P_0$)) if there exist positive constants C_1 and γ, such that for any δ in a small right neighborhood of 0,*

$$P_0\left(C_\alpha^* - \delta \le f(X) \le C_\alpha^*\right) \wedge P_0\left(C_\alpha^* \le f(X) \le C_\alpha^* + \delta\right) \ge C_1 \delta^\gamma.$$

Definition 3 *Fix $\delta \in (0,1)$, for $d_n = 2\sqrt{2\frac{\log(2en)+\log(2/\delta)}{n}}$, let $\hat{C}_\alpha$ be the smallest C such that*

$$\frac{1}{n}\sum_{i=n+1}^{2n} \mathbb{I}\left(\frac{\hat{p}_1(X_i^-)}{\hat{p}_0(X_i^-)} \ge C\right) \le \alpha - d_n.$$

Having $\hat{p}_1, \hat{p}_0$ and $\hat{C}_\alpha$, we propose a plug-in classifier motivated by the Neyman-Pearson Lemma in statistical hypothesis testing:

$$\hat{\phi}(x) = \mathbb{I}\left(\frac{\hat{p}_1(x)}{\hat{p}_0(x)} \ge \hat{C}_\alpha\right). \tag{3}$$

Unlike the previous setup where p_0 was known, we now need to bound the type I error of $\hat{\phi}$ first.

Proposition 2 *With probability at least $1-\delta$ regarding the samples $\mathcal{S}_0$, $\tilde{\mathcal{S}}_0$ and $\mathcal{S}_1$, type I error of the plug-in classifier $\hat{\phi}$ defined in* (3) *is bounded from above by α, that is,*

$$R_0(\hat{\phi}) \le \alpha.$$

Proof Note that $R_0(\hat{\phi}) = P_0\left(\frac{\hat{p}_1(X)}{\hat{p}_0(X)} \ge \hat{C}_\alpha\right)$ and $\frac{1}{n}\sum_{i=n+1}^{2n} \mathbb{I}\left(\frac{\hat{p}_1(X_i^-)}{\hat{p}_0(X_i^-)} \ge \hat{C}_\alpha\right) \le \alpha - d_n$. Let

$$\mathcal{A}_t = \left\{ \sup_{c\in\mathbb{R}} \left| P_0\left(\frac{\hat{p}_1(X)}{\hat{p}_0(X)} \ge c\right) - \frac{1}{n}\sum_{i=n+1}^{2n} \mathbb{I}\left(\frac{\hat{p}_1(X_i^-)}{\hat{p}_0(X_i^-)} \ge c\right) \right| \ge t \right\}.$$

Then it is enough to show that, $\mathbb{P}(\mathcal{A}_{d_n}) \le \delta$.

Note that $\mathbb{P}(\mathcal{A}_t) = \mathbb{E}(\mathbb{P}(\mathcal{A}_t | \mathcal{S}_0, \mathcal{S}_1))$. Keep $\mathcal{S}_0$ and $\mathcal{S}_1$ fixed, and define $\hat{f}(x) = \frac{\hat{p}_1(x)}{\hat{p}_0(x)}$. Let Q be the conditional distribution of $Z = \hat{f}(X)$ given $\mathcal{S}_0$ and $\mathcal{S}_1$, where $X \sim P_0$, and Q^n denote the conditional joint distribution of $(Z_{n+1}^-, \ldots, Z_{2n}^-) = (\hat{f}(X_{n+1}^-), \ldots, \hat{f}(X_{2n}^-))$. Because half lines in $\mathbb{R}$ have VC dimension 1, by taking $t = d_n = 2\sqrt{2\frac{\log(2en)+\log(2/\delta)}{n}}$, the VC inequality[1] implies that

$$\mathbb{P}(\mathcal{A}_{d_n} | \mathcal{S}_0, \mathcal{S}_1) \quad = Q^n\left(\sup_c \left| Q(Z \ge c) - \tfrac{1}{n}\textstyle\sum_{i=n+1}^{2n} \mathbb{I}(Z_i^- \ge c) \right| \ge d_n\right) \quad \le \delta.$$

Therefore,

$$\mathbb{P}(\mathcal{A}_{d_n}) = \mathbb{E}(\mathbb{P}(\mathcal{A}_{d_n} | \mathcal{S}_0, \mathcal{S}_1)) \le \delta.$$

■

The next theorem addresses the excess type II error of $\hat{\phi}$.

1. For the readers' convenience, a simple corollary of VC inequality is quoted: let $\mathcal{G}$ be a class of classifiers with VC dimension l, then with probability at least $1-\delta$, $\sup_{g\in\mathcal{G}} |R(g) - R_n(g)| \le 2\sqrt{2\frac{l\log(2en/l)+\log(2/\delta)}{n}}$, where n is the sample size and R_n denotes the empirical risk.

Theorem 2 *Let $\hat{\phi}$ be the plug-in classifier defined as in* (3)*. Assume that the class conditional densities p_0 and p_1 satisfy Condition 1, $p_0 \in \mathcal{P}_\Sigma(\beta, L, [-1,1]^d)$ and the kernel K is β-valid and L'-Lipschitz. Also assume that the likelihood ratio p_1/p_0 satisfies the level α detection condition for some $\underline{\gamma} \geq \bar{\gamma}$. Then for any $\delta \in (0,1)$ and any sample sizes m,n such that*

$$\max\left(\sqrt{\frac{\log(m/\delta)}{mh_m^d}}, \sqrt{\frac{\log(n/\delta)}{nh_n^d}}\right) < 1\,,$$

where the bandwidths $h_n = (\frac{\log n}{n})^{\frac{1}{2\beta+d}}$ and $h_m = (\frac{\log m}{m})^{\frac{1}{2\beta+d}}$, it holds with probability $1-3\delta$,

$$R_1(\hat{\phi}) - R_1(\phi^*) \leq 2C_0\left[\left(2d_n/C_1\right)^{1/\underline{\gamma}} + 2T_{m,n}\right]^{1+\bar{\gamma}} + 2C_\alpha^* d_n\,,$$

where $d_n = 2\sqrt{2\frac{\log(2en)+\log(2/\delta)}{n}}$, $T_{m,n} = \frac{\delta_1 + \|p_1\|_\infty \delta_0/\mu_{\min}}{\mu_{\min}-\delta_0}$, $\delta_0 = C_2\sqrt{\frac{\log(n/\delta)}{nh_n^d}}$,
$\delta_1 = C_3\sqrt{\frac{\log(m/\delta)}{mh_m^d}}$, C_2 and C_3 are the same as C in Lemma 1 applied to p_0 and p_1 respectively.

In particular, there exists some positive $\bar{C}$, such that for all $n,m \geq 1/\delta$,

$$R_1(\hat{\phi}) - R_1(\phi^*) \leq \bar{C}\left[\left(\frac{\log n}{n}\right)^{\min\left(\frac{1}{2}, \frac{1+\bar{\gamma}}{2\underline{\gamma}}, \frac{\beta(1+\bar{\gamma})}{2\beta+d}\right)} + \left(\frac{\log m}{m}\right)^{\frac{\beta(1+\bar{\gamma})}{2\beta+d}}\right]. \tag{4}$$

Proof

Denote by $G^* = \left\{\frac{p_1}{p_0} < C_\alpha^*\right\}$ and $\hat{G} = \left\{\frac{\hat{p}_1}{\hat{p}_0} < \hat{C}_\alpha\right\}$. Then

$$\begin{aligned}
&\int_{G^* \triangle \hat{G}} \left|\frac{p_1}{p_0} - C_\alpha^*\right| dP_0 \\
= &\int_{G^* \cap \hat{G}^c} \left(C_\alpha^* - \frac{p_1}{p_0}\right) dP_0 + \int_{G^{*c} \cap \hat{G}} \left(\frac{p_1}{p_0} - C_\alpha^*\right) dP_0 \\
= &\int_{G^*} \left(C_\alpha^* - \frac{p_1}{p_0}\right) dP_0 + \int_{\hat{G}} \left(\frac{p_1}{p_0} - C_\alpha^*\right) dP_0 \\
= &P_1(\hat{G}) - P_1(G^*) + C_\alpha^*\left[P_0(G^*) - P_0(\hat{G})\right].
\end{aligned}$$

Therefore the excess type II error can be decomposed in two parts,

$$P_1(\hat{G}) - P_1(G^*) = \int_{G^* \triangle \hat{G}} \left|\frac{p_1}{p_0} - C_\alpha^*\right| dP_0 + C_\alpha^*\left[P_0(G^{*c}) - P_0(\hat{G}^c)\right]. \tag{5}$$

Recall that $P_0(G^{*c}) = \alpha$ and $P_0(\hat{G}^c)$ is type I error of $\hat{\phi}$. From the above decomposition, we see that to control the excess type II error, type I error of $\hat{\phi}$ should be not only smaller than the level α, but also not far from α. This is intuitively correct, because having a small type I error amounts to having a very tight constraint set, which leads to significant deterioration in achievable type II error. Fortunately, this is not the case here with high probability. Note that by the definition of $\hat{C}_\alpha$, for any positive number l, the following holds for all $n \geq 1$,

$$\frac{1}{n}\sum_{i=n+1}^{2n} \mathbb{I}\left(\frac{\hat{p}_1(X_i^-)}{\hat{p}_0(X_i^-)} \geq \hat{C}_\alpha - l\right) > \alpha - d_n\,.$$

By the same argument as in the proof of Proposition 2, there exists an event $\bar{\mathcal{E}}_l$ regarding the samples $\mathcal{S}_0$, $\tilde{\mathcal{S}}_0$ and $\mathcal{S}_1$ with $\mathbb{P}(\bar{\mathcal{E}}_l) \geq 1-\delta$, such that on this event,

$$P_0\left(\frac{\hat{p}_1(X)}{\hat{p}_0(X)} \geq \hat{C}_\alpha - l\right) \geq \alpha - 2d_n\,. \tag{6}$$

To control the second part of R.H.S. in(5), let $\hat{G}_l = \{\frac{\hat{p}_1}{\hat{p}_0} < \hat{C}_\alpha - l\}$, then

$$P_0(G^{*c}) - P_0(\hat{G}^c) = \inf_{l>0}\left[P_0(G^{*c}) - P_0(\hat{G}_l^c)\right] \leq \alpha - (\alpha - 2d_n) = 2d_n\,, \tag{7}$$

on the event $\bar{\mathcal{E}} := \cap_{l>0}\bar{\mathcal{E}}_l$, and $\mathbb{P}(\bar{\mathcal{E}}) = \lim_{l\to 0}\mathbb{P}(\bar{\mathcal{E}}_l) \geq 1-\delta$.

Therefore, it remains to control the first part of R.H.S. in (5). Define an event regarding samples $\mathcal{S}_0$ and $\mathcal{S}_1$:

$$\mathcal{E} = \{\|\hat{p}_0 - p_0\|_\infty < \delta_0, \|\hat{p}_1 - p_1\|_\infty < \delta_1\}\,.$$

Lemma 1 implies that $\mathbb{P}(\mathcal{E}) \geq 1-2\delta$. We restrict ourselves to $\mathcal{E}\cap\bar{\mathcal{E}}$ for the rest of the proof. Note that the first part of R.H.S. in (5) can be decomposed by

$$\int_{G^*\triangle\hat{G}}\left|\frac{p_1}{p_0} - C_\alpha^*\right| dP_0 \quad = \quad \int_{G^{*c}\cap\hat{G}}\left|\frac{p_1}{p_0} - C_\alpha^*\right| dP_0 + \int_{G^*\cap\hat{G}^c}\left|\frac{p_1}{p_0} - C_\alpha^*\right| dP_0\,.$$

We will focus on bounding the integral over $G^*\cap\hat{G}^c$, because that over $G^{*c}\cap\hat{G}$ can be bounded similarly. Note that,

$$\left|\frac{\hat{p}_1}{\hat{p}_0} - \frac{p_1}{p_0}\right| \leq \left|\frac{\hat{p}_1}{\hat{p}_0} - \frac{p_1}{\hat{p}_0}\right| + \left|\frac{p_1}{\hat{p}_0} - \frac{p_1}{p_0}\right| \leq \frac{1}{|\hat{p}_0|}|\hat{p}_1 - p_1| + \left|\frac{p_1}{p_0}\right|\cdot\frac{|p_0-\hat{p}_0|}{|\hat{p}_0|} < \frac{\|p_1\|_\infty\delta_0/\mu_{\min} + \delta_1}{\mu_{\min} - \delta_0} = T_{m,n}\,.$$

The above inequality together with (6) implies that

$$\alpha - 2d_n \leq P_0\left(\frac{\hat{p}_1(X)}{\hat{p}_0(X)} \geq \hat{C}_\alpha - l\right) \leq P_0\left(\frac{p_1(X)}{p_0(X)} \geq \hat{C}_\alpha - l - T_{m,n}\right)\,. \tag{8}$$

We need to bound $\hat{C}_\alpha$ in terms of C_α^*. This is achieved through the following steps. First, we determine some $c_n > 0$ such that

$$P_0\left(\frac{p_1(X)}{p_0(X)} \geq C_\alpha^* + c_n\right) \leq \alpha - 2d_n\,,$$

which follows if the next inequality holds

$$2d_n \leq P_0\left(C_\alpha^* < \frac{p_1(X)}{p_0(X)} < C_\alpha^* + c_n\right)\,.$$

By the level α detection condition, it is enough to take $c_n = (2d_n/C_1)^{1/\gamma}$. Therefore in view of inequality (8),

$$P_0\left(\frac{p_1(X)}{p_0(X)} \geq C_\alpha^* + (2d_n/C_1)^{1/\gamma}\right) \leq \alpha - 2d_n \leq P_0\left(\frac{p_1(X)}{p_0(X)} \geq \hat{C}_\alpha - l - T_{m,n}\right)\,.$$

This implies that

$$\hat{C}_\alpha \le C^*_\alpha + (2d_n/C_1)^{1/\gamma} + l + T_{m,n}\,.$$

Since the above holds for all $l > 0$, we have

$$\hat{C}_\alpha \le C^*_\alpha + (2d_n/C_1)^{1/\gamma} + T_{m,n}\,.$$

For any positive $L_{m,n}$, we can decompose $G^{*c} \cap \hat{G}$ by

$$G^{*c} \cap \hat{G} = \left\{ \frac{p_1}{p_0} \ge C^*_\alpha, \frac{\hat{p}_1}{\hat{p}_0} < \hat{C}_\alpha \right\} = A_1 \cap A_2\,,$$

where

$$A_1 = \left\{ C^*_\alpha + L_{m,n} > \frac{p_1}{p_0} \ge C^*_\alpha, \frac{\hat{p}_1}{\hat{p}_0} < \hat{C}_\alpha \right\}, \text{ and } A_2 = \left\{ \frac{p_1}{p_0} \ge C^*_\alpha + L_{m,n}, \frac{\hat{p}_1}{\hat{p}_0} < \hat{C}_\alpha \right\}.$$

By the margin assumption,

$$\int_{A_1} \left| \frac{p_1}{p_0} - C^*_\alpha \right| dP_0 \le L_{m,n} P_0(A_1) \le C_0 (L_{m,n})^{1+\bar{\gamma}}.$$

Note that

$$A_2 \subset A_3 := \left\{ \frac{p_1}{p_0} \ge C^*_\alpha + L_{m,n}, \frac{\hat{p}_1}{\hat{p}_0} < C^*_\alpha + (2d_n/C_1)^{1/\gamma} + T_{m,n} \right\}.$$

Take $L_{m,n} = (2d_n/C_1)^{1/\gamma} + 2T_{m,n}$, then $P_0(A_2) = P_0(A_3) = 0$. Therefore,

$$\int_{G^{*c} \cap \hat{G}} \left| \frac{p_1}{p_0} - C^*_\alpha \right| dP_0 \le C_0 (L_{m,n})^{1+\bar{\gamma}}.$$

Similarly, we can bound the integral over $G^* \cap \hat{G}^c$, so

$$\int_{G^* \triangle \hat{G}} \left| \frac{p_1}{p_0} - C^*_\alpha \right| dP_0 \le 2C_0 (L_{m,n})^{1+\bar{\gamma}}. \tag{9}$$

Finally, note that $\mathbb{P}(\mathcal{E} \cap \bar{\mathcal{E}}) \ge 1 - 3\delta$. So (5), (7) and (9) together conclude the proof.

■

Now we briefly discuss the above result. Same as the p_0 known setup, the coefficient $\bar{\gamma}$ from the margin assumption has influence on the convergence rate of the excess type II error. The larger the $\bar{\gamma}$, the easier the classification problem, and hence the faster the convergence of the excess type II error. The coefficient $\underline{\gamma}$ in the detection condition works differently. The larger the $\underline{\gamma}$, the more difficult it is to detect the optimal decision boundary, and hence the harder the classification problem. Take it to the extreme $\underline{\gamma} \to \infty$ (keep $\bar{\gamma}$ fixed), which holds when the amount of data around the optimal threshold level goes to zero,

$$\left(\frac{\log n}{n} \right)^{\frac{1+\bar{\gamma}}{2\underline{\gamma}}} \to \left(\frac{\log n}{n} \right)^0 = 1\,.$$

In other words, the upper bound in (4) is uninformative when we have a null level α detection condition.

In anomaly detection applications, let class 0 represent the normal class, and class 1 represent the anomaly class. We have in mind $n \gg m$, that is, the normal sample size is much bigger than that of the anomaly, and so $\log n/n$ is dominated by $\log m/m$. Therefore, the right hand side of (4) is of the order $\left[(\frac{\log n}{n})^{\min(1/2,(1+\bar{\gamma})/(2\underline{\gamma}))} + (\frac{\log m}{m})^{\beta(1+\bar{\gamma})/(2\beta+d)} \right]$. Compared with the p_0 known setup, the extra term $\left(\frac{\log n}{n}\right)^{\min(1/2,(1+\bar{\gamma})/(2\underline{\gamma}))}$ arises from estimating the threshold level C^*_α. Let $n \to \infty$, which amounts to knowing p_0, this term vanishes, and the upper bound reduces to the same as in the previous subsection. When $\underline{\gamma} < 1+\bar{\gamma}$, we have $1/2 < (1+\bar{\gamma})/(2\underline{\gamma})$, so $\underline{\gamma}$ does not show up in the upper bound. Finally, for fixed $\underline{\gamma}(\geq 1+\bar{\gamma})$, β and d, we can calculate explicitly an order relation between m and n, such that $(\frac{\log m}{m})^{\frac{\beta}{2\beta+d}} \geq (\frac{\log n}{n})^{\frac{1}{2\underline{\gamma}}}$. A sufficient condition for this inequality is that $n \geq \left(\frac{m}{\log m}\right)^{\beta\underline{\gamma}/(4\beta+2d)}$. Intuitively, this says that if the normal class sample size is large enough compared to the anomaly class sample size, lack of precise knowledge on normal class density p_0 does not change the type II error rate bound, up to a multiplicative constant.

4. Plug-in Based on the Regression Function

In this section, instead of plugging in class conditional densities p_1 and p_0, we target the regression function $\eta(x) = \mathbb{E}(Y|X=x)$ directly. As will be illustrated, this version of plug-in estimator allows us to handle a different assumption on sampling scheme. Recall that the rationality behind plugging in for p_1/p_0 lies in the Neyman-Pearson Lemma for hypothesis testing. A simple derivation shows that a thresholding rule on p_1/p_0 can be translated into a thresholding rule on η. Indeed, let $\pi = \mathbb{P}(Y=1)$, then we have

$$\eta(x) = \mathbb{P}(Y=1|X=x) = \frac{\pi \cdot p_1/p_0(x)}{\pi \cdot p_1/p_0(x) + (1-\pi)}.$$

When $\pi < 1$, the function $x \mapsto \frac{\pi x}{\pi x + (1-\pi)}$ is strictly monotone increasing on $\mathbb{R}^+$. Therefore, there exists a positive constant D^*_α depending on α, such that

$$\left\{ x \in [-1,1]^d : \frac{p_1(x)}{p_0(x)} \geq C^*_\alpha \right\} = \{ x \in [-1,1]^d : \eta(x) \geq D^*_\alpha \}.$$

Moreover, the oracle thresholds C^*_α and D^*_α are related by $D^*_\alpha = \frac{\pi C^*_\alpha}{\pi C^*_\alpha + (1-\pi)}$. Parallel to the previous section, we address both the p_0 known and p_0 unknown setups. In both setups, we assume that we have access to an i.i.d. sample $\bar{\mathcal{S}} = \{(X_1,Y_1),\ldots,(X_m,Y_m)\}$.

4.1 Class 0 Density p_0 Known

This part is similar to the p_0 known setup in Section 3. The essential technical difference is that we need a uniform deviation result on the Nadaraya-Watson estimator $\hat{\eta}$ (based on the sample $\bar{\mathcal{S}}$) instead of those on $\hat{p}_0$ and $\hat{p}_1$. Recall that $\hat{\eta} = \sum_{i=1}^m Y_i K\left(\frac{X_i - x}{h}\right) / \sum_{i=1}^m K\left(\frac{X_i - x}{h}\right)$ can written as $\hat{f}/\hat{p}$,

where

$$\hat{p}(x) = \frac{1}{mh^d}\sum_{i=1}^{m} K\left(\frac{X_i - x}{h}\right) \text{ and } \hat{f}(x) = \frac{1}{mh^d}\sum_{i=1}^{m} Y_i K\left(\frac{X_i - x}{h}\right),$$

in which h is the bandwidth. Denote by $p = \pi p_1 + (1-\pi)p_0$ and $f = \eta \cdot p$, then $\eta = f/p$. For a given level α, define $\hat{D}_\alpha$ and D^*_α respectively by

$$P_0(\hat{\eta}(X) \geq \hat{D}_\alpha) = \alpha \quad \text{and} \quad P_0(\eta(X) \geq D^*_\alpha) = \alpha\,.$$

For simplicity, we assume that $\hat{D}_\alpha$ and D^*_α exist and are unique. Note that the oracle classifier of level α is $\phi^*(x) = \mathbb{I}(\eta(x) \geq D^*_\alpha)$, so a plug-in classifier motivated by the Neyman-Pearson Lemma is

$$\tilde{\phi}(x) = \mathbb{I}(\hat{\eta}(x) \geq \hat{D}_\alpha)\,. \tag{10}$$

Since $\hat{D}_\alpha$ is constructed to meet the level α exactly, the excess type I error of $\tilde{\phi}$ vanishes, that is,

$$R_0(\tilde{\phi}) - R_0(\phi^*) = 0\,.$$

The following theorem addresses type II error of $\tilde{\phi}$.

Condition 3 *Suppose that p the marginal density of X and η the regression function satisfy:*

(i) There exist positive constants $\mu'_{\min}$ and $\nu'_{\max}(<1)$, such that $p \geq \mu'_{\min}$ and $\eta \leq \nu'_{\max}$,

(ii) $f = \eta \cdot p \in \mathcal{P}_\Sigma(\beta, L, [-1,1]^d)$ and $p \in \mathcal{P}_\Sigma(\beta, L, [-1,1]^d)$,

(iii) The regression function η satisfies the margin assumption of order $\bar{\gamma}$ with respect to probability distribution P_0 at the level D^*_α.

Proposition 3 *Let $\tilde{\phi}$ be the plug-in classifier defined by* (10)*. Assume that p and η satisfy condition 3 and that the kernel K is β-valid and L'-Lipschitz. Then there exists a positive $\tilde{D}$, such that for any $\delta \in (0,1)$ and any sample size m satisfying $\sqrt{\frac{\log(m/\delta)}{mh^d}} < 1$, it holds with probability $1-\delta$,*

$$R_1(\tilde{\phi}) - R_1(\phi^*) \leq \tilde{D}\left(\frac{\log(3m/\delta)}{mh^d}\right)^{\frac{1+\bar{\gamma}}{2}},$$

where $h = \left(\frac{\log m}{m}\right)^{\frac{1}{2\beta+d}}$. Furthermore, there exists a positive D such that for any $m \geq 1/\delta$, it holds with probability $1-\delta$,

$$R_1(\tilde{\phi}) - R_1(\phi^*) \leq D\left(\frac{\log m}{m}\right)^{\frac{\beta(1+\bar{\gamma})}{2\beta+d}}.$$

Proof First note that the excess type II error

$$R_1(\tilde{\phi}) - R_1(\phi^*) = \int_{G^* \triangle \hat{G}} \left|\frac{p_1}{p_0} - C^*_\alpha\right| dP_0 = P_1(\hat{G}) - P_1(G^*)\,,$$

where $G^* = \{\eta < D_\alpha^*\}$ and $\hat{G} = \left\{\hat{\eta} < \hat{D}_\alpha\right\}$, and $G^* \triangle \hat{G} = (G^* \cap \hat{G}^c) \cup (G^{*c} \cap \hat{G})$.

Define an event regarding the sample $\bar{S}$,

$$\mathcal{E} = \{\|\hat{\eta} - \eta\|_\infty < \delta_1/2\},$$

where $\delta_1 = D_1\sqrt{\frac{\log(3m/\delta)}{mh^d}}$ and D_1 is a constant chosen as in Lemma 2. From this point to the end of the proof, we restrict ourselves to $\mathcal{E}$. Decompose

$$G^{*c} \cap \hat{G} = \left\{\eta \geq D_\alpha^*, \hat{\eta} < \hat{D}_\alpha\right\} = A_1 \cup A_2,$$

where

$$A_1 = \left\{D_\alpha^* + \delta_1 \geq \eta \geq D_\alpha^*, \hat{\eta} < \hat{D}_\alpha\right\},$$

and

$$A_2 = \left\{\eta > D_\alpha^* + \delta_1, \hat{\eta} < \hat{D}_\alpha\right\}.$$

Now we need to control the distance of $\left|\frac{p_1}{p_0} - C_\alpha^*\right|$ in terms of $|\eta - D_\alpha^*|$. This can be achieved by recalling

$$\eta = \frac{\pi p_1/p_0}{\pi p_1/p_0 + 1 - \pi} \text{ and } D_\alpha^* = \frac{\pi C_\alpha^*}{\pi C_\alpha^* + 1 - \pi},$$

and the assumption that $\eta \leq \nu'_{\max}(< 1)$ (also $D_\alpha^* \leq \nu'_{\max}$ should follow). Indeed, let

$$f(x) = \frac{\pi x}{\pi x + (1-\pi)}, 0 < x < \nu'_{\max}.$$

Then,

$$g(x) = f^{-1}(x) = \frac{1-\pi}{\pi}\frac{x}{1-x}, 0 < x < \frac{\pi\nu'_{\max}}{\pi\nu'_{\max} + 1 - \pi}.$$

Since $|g'(x)| \leq \frac{\pi}{1-\pi}\left(\frac{\pi\mu'_{\max}+1-\pi}{\pi\nu'_{\max}}\right)^2 =: U$, g is Lipschitz with Lipschitz constant U. Therefore,

$$|\eta - D_\alpha^*| \leq \delta_1 \implies \left|\frac{p_1}{p_0} - C_\alpha^*\right| \leq U\delta_1.$$

This implies

$$\int_{A_1}\left|\frac{p_1}{p_0} - C_\alpha^*\right| dP_0 \leq U\delta_1 P_0(A_1) \leq C_0 U \delta_1^{1+\bar{\gamma}},$$

where the second inequality follows from the margin assumption. To bound the integral over A_2, we control the distance between $\hat{D}_\alpha$ and D_α^*. Indeed,

$$\begin{aligned}\alpha &= P_0(\hat{\eta}(X) \geq \hat{D}_\alpha) = P_0(\eta(X) \geq D_\alpha^*) \\ &\geq P_0(\hat{\eta}(X) \geq D_\alpha^* + \delta_1/2).\end{aligned}$$

This implies that $\hat{D}_\alpha \leq D_\alpha^* + \delta_1/2$. So

$$P_0(A_2) \leq P_0(\eta > D_\alpha^* + \delta_1, \hat{\eta} < D_\alpha^* + \delta_1/2) = 0.$$

Therefore,

$$\int_{G^{*c}\cap\hat{G}}\left|\frac{p_1}{p_0} - C_\alpha^*\right| \leq C_0 U \delta_1^{1+\bar{\gamma}}.$$

Similarly bound the integral over $G^* \cap \hat{G}^c$, then

$$\int_{G^* \triangle \hat{G}} \left| \frac{p_1}{p_0} - C_\alpha^* \right| \leq 2C_0 U \delta_1^{1+\bar{\gamma}}.$$

Lemma 2 implies that $\mathbb{P}(\mathcal{E}) \geq 1-\delta$. This concludes the proof. ■

Note that Lemma 2 is a uniform deviation result on the Nadaraya-Watson estimator, and it is the first result of such kind to the best of our knowledge.

4.2 Class 0 Density p_0 Unknown

In this subsection, the assumption of knowledge on p_0 is relaxed. Suppose in addition to the mixed sample $\bar{\mathcal{S}} = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$, we have access to a class 0 sample $\mathcal{S}_0 = \{X_1^-, \ldots, X_n^-\}$. Moreover, assume that variables in $\mathcal{S}_0$ are independent given $\bar{\mathcal{S}}$. As in the p_0 known case, the notation $\hat{\eta}$ denotes the Nadaraya-Watson estimator based on the sample $\bar{\mathcal{S}}$. Just like $\hat{C}_\alpha$ in Definition 3, we need to define the threshold level $\hat{D}_\alpha$ carefully.

Definition 4 *Fix $\delta \in (0,1)$, for $d_n = 2\sqrt{2\frac{\log(2en)+\log(2/\delta)}{n}}$, let $\hat{D}_\alpha$ be the smallest L such that*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(\hat{\eta}(X_i^-) \geq L\right) \leq \alpha - d_n .$$

Unlike the previous setup where p_0 is known, we now bound the excess type I error.

Proposition 4 *With probability at least $1-\delta$, type I error of the plug-in classifier $\tilde{\phi}$ defined in* (10) *is bounded from above by α, that is,*

$$R_0(\tilde{\phi}) \leq \alpha .$$

The proof is omitted due to similarity to that of Proposition 2. A similar α level detection condition can be formulated can be formulated for the regression function, but we omit it as C_α^* is simply replaced by D_α^*. The next theorem address the excess type II error of $\tilde{\phi}$: $R_1(\tilde{\phi}) - R_1(\phi^*)$.

Theorem 3 *Let $\tilde{\phi} = \mathbb{I}(\hat{\eta} \geq \hat{D}_\alpha)$ be defined as in* (10)*. Assume condition 3 and the regression function η satisfies the level α detection condition for some $\gamma (\geq \bar{\gamma})$. Take the bandwidth $h = (\frac{\log m}{m})^{1/(2\beta+d)}$ in the Nadaraya-Watson estimator $\hat{\eta}$, where the kernel K is β-valid and L'-Lipschitz. Then there exists a positive constant $\bar{C}$, such that for any $\delta \in (0,1)$ and any $m, n \geq 1/\delta$, it holds with probability $1-2\delta$,*

$$R_1(\tilde{\phi}) - R_1(\phi^*) \leq \bar{C}\left[\left(\frac{\log n}{n}\right)^{\min\left(\frac{1}{2}, \frac{1+\bar{\gamma}}{2\gamma}\right)} + \left(\frac{\log m}{m}\right)^{\frac{\beta(1+\bar{\gamma})}{2\beta+d}}\right].$$

Proof Let $G^* = \{\eta < D_\alpha^*\}$ and $\hat{G} = \{\hat{\eta} < \hat{D}_\alpha\}$, then the excess type II error of $\tilde{\phi}$ can be decomposed by

$$P_1(\hat{G}) - P_1(G^*) = \int_{G^* \triangle G^*} \left| \frac{p_1}{p_0} - C_\alpha^* \right| dP_0 + C_\alpha^* + \left[P_0(G^{*c}) - P_0(\hat{G}^c)\right]. \tag{11}$$

Recall that $P_0(G^{*c}) = \alpha$ and $P_0(\hat{G}^c)$ is type I error of $\tilde{\phi}$. By the definition of $\hat{D}_\alpha$, for any positive number l, the following holds for all $n \geq 1$,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\hat{\eta}(X_i^-) \geq \hat{D}_\alpha - l) > \alpha - d_n\,,$$

where $d_n = 2\sqrt{2\frac{\log(2en)+\log(2/\delta)}{n}}$. By the same argument as in the proof of Proposition 2, there exists an event $\bar{\mathcal{E}}_l$ regarding the samples $\mathcal{S}_0$, $\tilde{\mathcal{S}}_0$ and $\mathcal{S}_1$ with $\mathbb{P}(\bar{\mathcal{E}}_l) \geq 1-\delta$, such that on this event,

$$P_0\left(\hat{\eta} \geq \hat{D}_\alpha - l\right) \geq \alpha - 2d_n\,.$$

To control the second part of R.H.S. in(11), let $\hat{G}_l = \{\hat{\eta} < \hat{D}_\alpha - l\}$, then

$$P_0(G^{*c}) - P_0(\hat{G}^c) = \inf_{l>0}\left[P_0(G^{*c}) - P_0(\hat{G}_l^c)\right] \leq \alpha - (\alpha - 2d_n) = 2d_n\,, \tag{12}$$

on the event $\bar{\mathcal{E}} := \cap_{l>0}\bar{\mathcal{E}}_l$, and $\mathbb{P}(\bar{\mathcal{E}}) = \lim_{l\to 0}\mathbb{P}(\bar{\mathcal{E}}_l) \geq 1-\delta$. Therefore, it remains to control the first part of R.H.S. in (11). Define an event regarding the sample $\bar{\mathcal{E}}$,

$$\mathcal{E} = \{\|\hat{\eta} - \eta\|_\infty < \delta_1/2\}\,,$$

where $\delta_1 = D_1\sqrt{\frac{\log(3m/\delta)}{mh^d}}$ and D_1 is a constant chosen as in Lemma 2. Lemma 2 implies that $\mathbb{P}(\mathcal{E}) \geq 1-\delta$. We restrict ourselves to $\mathcal{E}\cap\bar{\mathcal{E}}$ for the rest of the proof. Note that the first part of R.H.S. of (11) can be decomposed by

$$\int_{G^*\triangle\hat{G}}\left|\frac{p_1}{p_0} - C_\alpha^*\right| dP_0 \quad = \quad \int_{G^{*c}\cap\hat{G}}\left|\frac{p_1}{p_0} - C_\alpha^*\right| dP_0 + \int_{G^*\cap\hat{G}^c}\left|\frac{p_1}{p_0} - C_\alpha^*\right| dP_0\,.$$

We will focus the integral over $G^*\cap\hat{G}^c$, as that over $G^{*c}\cap\hat{G}$ can be bounded similarly. Because $|\hat{\eta}-\eta| < \delta_1/2$, for all $l > 0$,

$$\alpha - 2d_n \leq P_0\left(\hat{\eta} \geq \hat{D}_\alpha - l\right) \leq P_0\left(\eta \geq \hat{D}_\alpha - l - \delta_1/2\right)\,.$$

We need to bound $\hat{D}_\alpha$ in terms of D_α^*. This is achieved through the following steps. First, we determine some $c_n > 0$ such that

$$P_0\left(\eta \geq D_\alpha^* + c_n\right) \leq \alpha - 2d_n\,,$$

which follows if the next inequality holds

$$2d_n \leq P_0\left(D_\alpha^* < \eta < D_\alpha^* + c_n\right)\,.$$

By the level α detection condition, it is enough to take $c_n = \left(2d_n/C_1\right)^{1/\gamma}$. Therefore,

$$P_0\left(\eta \geq D_\alpha^* + \left(2d_n/C_1\right)^{1/\gamma}\right) \leq \alpha - 2d_n \leq P_0\left(\eta \geq \hat{D}_\alpha - l - \delta_1/2\right)\,.$$

This implies that

$$\hat{D}_\alpha \leq D_\alpha^* + \left(2d_n/C_1\right)^{1/\gamma} + l + \delta_1/2\,.$$

Since the above holds for all $l > 0$, we have

$$\hat{D}_\alpha \leq D^*_\alpha + (2d_n/C_1)^{1/\gamma} + \delta_1/2\,.$$

We can decompose $G^{*c} \cap \hat{G}$ by

$$G^{*c} \cap \hat{G} = \left\{\eta \geq D^*_\alpha, \hat{\eta} < \hat{D}_\alpha\right\} = A_1 \cap A_2\,,$$

where

$$A_1 = \left\{D^*_\alpha + (2d_n/C_1)^{1/\gamma} + \delta_1 > \eta \geq D^*_\alpha, \hat{\eta} < \hat{D}_\alpha\right\}, \text{ and}$$

$$A_2 = \left\{\eta \geq D^*_\alpha + (2d_n/C_1)^{1/\gamma} + \delta_1, \hat{\eta} < \hat{D}_\alpha\right\}.$$

Let $U := \frac{\pi}{1-\pi}\left(\frac{\pi\mu'_{\max}+1-\pi}{\pi\nu'_{\max}}\right)^2$, through the same derivation as the p_0 known case,

$$|\eta - D^*_\alpha| \leq (2d_n/C_1)^{1/\gamma} + \delta_1 \implies \left|\frac{p_1}{p_0} - C^*_\alpha\right| \leq U\left((2d_n/C_1)^{1/\gamma} + \delta_1\right).$$

By the margin assumption,

$$\int_{A_1}\left|\frac{p_1}{p_0} - C^*_\alpha\right| dP_0 \leq U\left((2d_n/C_1)^{1/\gamma} + \delta_1\right) P_0(A_1) \leq C_0 U\left[(2d_n/C_1)^{1/\gamma} + \delta_1\right]^{1+\bar{\gamma}}.$$

Note that

$$A_2 \subset A_3 := \left\{\eta \geq D^*_\alpha + (2d_n/C_1)^{1/\gamma} + \delta_1, \hat{\eta} < D^*_\alpha + (2d_n/C_1)^{1/\gamma} + \delta_1/2\right\},$$

but $|\eta - \hat{\eta}| < \delta_1/2$ on $\mathcal{E} \cap \bar{\mathcal{E}}$, so $P_0(A_2) = P_0(A_3) = 0$. Therefore,

$$\int_{G^{*c}\cap\hat{G}}\left|\frac{p_1}{p_0} - C^*_\alpha\right| dP_0 \leq C_0 U\left[(2d_n/C_1)^{1/\gamma} + \delta_1\right]^{1+\bar{\gamma}}.$$

Similarly, the integral over $G^* \cap \hat{G}^c$ can be bounded, so we have

$$\int_{G^*\triangle\hat{G}}\left|\frac{p_1}{p_0} - C^*_\alpha\right| dP_0 \leq 2C_0 U\left[(2d_n/C_1)^{1/\gamma} + \delta_1\right]^{1+\bar{\gamma}}. \tag{13}$$

Finally, note that $\mathbb{P}(\mathcal{E} \cap \bar{\mathcal{E}}) \geq 1 - 2\delta$. So (11), (12) and (13) together conclude the proof.

■

In anomaly detection applications, normal samples are considered abundant, that is, $n \gg m$, which implies that $(\frac{\log n}{n})^{\frac{1}{2\beta+d}} \leq (\frac{\log m}{m})^{\frac{1}{2\beta+d}}$. Then the upper bounds for the excess type II errors in Theorem 2 and Theorem 3 are of the same order. Having access to the mixture (contaminated) sample $\bar{\mathcal{S}}$ looks like a weaker condition than having access to a pure anomaly sample. However, this

does not seem to be the case in our settings. The essence is revealed by observing that the density ratio p_1/p_0 and the regression function η play the same role in the oracle NP classifier at level α:

$$\phi^*(x) = \mathbb{I}(p_1/p_0(x) \geq C_\alpha^*) = \mathbb{I}(\eta(x) \geq D_\alpha^*)\,.$$

A plug-in classifier depends upon an estimate of either p_1/p_0 or η. Being able to estimate the anomaly density p_1 is not of particular advantage, because only the ratio p_1/p_0 matters. Strictly speaking, the conditions for the two theorems are not the same, but one advantage of targeting the regression function seems to be that we do not have to split the normal example into two, with one to estimate p_0 and the other to estimate the optimal threshold C_α^*.

Acknowledgments

The author thanks Philippe Rigollet for thought-provoking discussions, and anomalous referees for constructive advice. This research was conducted with generous support at Statlab, Princeton University and Department of Mathematics, MIT.

Appendix A. Technical Lemmas

The appendix includes two important technical lemmas and their proofs. Lemma 1 is a uniform deviation result on kernel density estimators, and Lemma 2 is a uniform division result on Nadaraya-Watson estimators.

Lemma 1 *Let $p \in \mathcal{P}_\Sigma(\beta, L, [-1,1]^d)$ and the kernel K be β-valid and L'-Lipschitz. Denote by $\hat{p}(x) = \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)$ the kernel density estimator of p based on the i.i.d sample $\{X_1, \ldots, X_n\}$, where h is the bandwidth. For any $\varepsilon \in (0,1)$, if the sample size n is such that $\sqrt{\frac{\log(n/\varepsilon)}{nh^d}} < 1$, it holds*

$$\mathbb{P}(\|\hat{p} - p\|_\infty \geq \delta) \leq \varepsilon\,,$$

where

$$\delta = (32c_2 d + \sqrt{48dc_1})\sqrt{\frac{\log(n/\varepsilon)}{nh^d}} + 2Lc_3 h^\beta + \frac{1}{nh^d}\frac{\sqrt{d}L'}{nh} + (L + \tilde{C}\sum_{1 \leq |s| \leq \lfloor\beta\rfloor}\frac{1}{s!})d^{\beta/2}n^{-2\beta}\,,$$

where $c_1 = \|p\|_\infty \|K\|^2$, $c_2 = \|K\|_\infty + \|p\|_\infty + \int |K||t\|^\beta dt$, $c_3 = \int |K||t\|^\beta dt$, and $\tilde{C}$ is such that $\tilde{C} \geq \sup_{1 \leq |s| \leq \lfloor\beta\rfloor} \sup_{x \in [-1,1]^d} |D^s p(x)|$.

Let $h = \left(\frac{\log n}{n}\right)^{\frac{1}{2\beta+d}}$, it is enough to take $\delta = C\sqrt{\frac{\log(n/\varepsilon)}{nh^d}}$, where

$$C = 32c_2 d + \sqrt{48dc_1} + 2Lc_3 + \sqrt{d}L' + L + \tilde{C}\sum_{1 \leq |s| \leq \lfloor\beta\rfloor}\frac{1}{s!}\,.$$

Proof Divide each coordinate of the hypercube $[-1,1]^d$ into $2M$ equally spaced subintervals. Then $[-1,1]^d$ is subdivided into $(2M)^d$ small hypercubes, with a total of $(2M+1)^d$ vertices. We denote the collection of these vertices by G. Note that for any $\delta > 0$,

$$\mathbb{P}\left(\|\hat{p} - p\|_\infty \geq \delta\right) \leq \mathbb{P}\left(M_1 + M_2 + M_3 \geq \delta\right), \tag{14}$$

where

$$M_1 = \sup_{\|x-x'\| \le \frac{\sqrt{d}}{M}} \frac{1}{nh^d} \left| \sum_{i=1}^{n} \left(K(\frac{X_i - x}{h}) - K(\frac{X_i - x'}{h}) \right) \right|,$$
$$M_2 = \sup_{\|x-x'\| \le \frac{\sqrt{d}}{M}} |p(x) - p(x')|,$$
$$M_3 = \sup_{x \in G} \left| \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{X_i - x}{h}) - p(x) \right|.$$

Note that because K is L'-Lipschitz,

$$M_1 \le \sup_{\|x-x'\| \le \frac{\sqrt{d}}{M}} \frac{1}{nh^d} \sum_{i=1}^{n} \left| \left(K(\frac{X_i - x}{h}) - K(\frac{X_i - x'}{h}) \right) \right| \le \frac{1}{nh^d} \frac{n\sqrt{d}L'}{Mh} = \frac{1}{nh^d} \frac{\sqrt{d}L'}{nh}.$$

To control M_2, note that if $\beta \le 1$,

$$|p(x) - p(x')| = |p(x) - p_{x'}(x)| \le L\|x - x'\|^{\beta}.$$

If $\beta > 1$, p is $\lfloor \beta \rfloor$-times continuously differentiable. In particular, for all s such that $1 \le |s| \le \lfloor \beta \rfloor$, $D^s p$ is continuous. Since $[-1,1]^d$ is compact, there exists a positive constant $\tilde{C}$, such that

$$\sup_{1 \le |s| \le \lfloor \beta \rfloor} \sup_{x \in [-1,1]^d} |D^s p(x)| \le \tilde{C}.$$

Therefore,

$$\begin{aligned} |p(x) - p(x')| &\le |p(x) - p_{x'}(x)| + |p_{x'}(x) - p(x')| \\ &= L\|x - x'\|^{\beta} + \sum_{1 \le |s| \le \lfloor \beta \rfloor} \left| \frac{(x' - x)^s}{s!} D^s p(x') \right| \\ &\le \left(L + \tilde{C} \sum_{1 \le |s| \le \lfloor \beta \rfloor} \frac{1}{s!} \right) \|x - x'\|^{\beta}. \end{aligned}$$

Putting together the $\beta \le 1$ and $\beta > 1$ cases yields $M_2 \le \left(L + \tilde{C} \sum_{1 \le |s| \le \lfloor \beta \rfloor} \frac{1}{s!} \right) d^{\beta/2} n^{-2\beta}$.

Define by $t = \delta - \frac{1}{nh^d} \frac{\sqrt{d}L'}{nh} - \left(L + \tilde{C} \sum_{1 \le |s| \le \lfloor \beta \rfloor} \frac{1}{s!} \right) d^{\beta/2} n^{-2\beta}$. Inequality (14) together with upper bounds on M_1 and M_2, implies that

$$\mathbb{P}\left(\|\hat{p} - p\|_{\infty} \ge \delta \right) \le \mathbb{P}\left(M_3 \ge t \right).$$

Use a union bound to control the tail probability of M_3,

$$\begin{aligned} \mathbb{P}(M_3 \ge t) &\le \sum_{x \in G} \mathbb{P}\left(\left| \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{X_i - x}{h}) - p(x) \right| \ge t \right). \\ &\le 2(2M+1)^d \exp\left(-\frac{h^d n t^2}{8(c_1 + c_2 t/6)} \right), \end{aligned}$$

for $t \geq 2Lc_3h^\beta$. The last inequality relies on the assumptions that $p \in \mathcal{P}_\Sigma(\beta, L, [-1,1]^d)$ and K is β-valid. It essentially follows along the same lines as the proof of Lemma 4.1 in Rigollet and Vert (2009), so we omit the derivation and refer the interested reader to this paper. For a given $\varepsilon \in (0,1)$, we would like to enforce

$$2(2M+1)^d \exp\left(-\frac{h^d n t^2}{8(c_1 + c_2 t/6)}\right) \leq \varepsilon .$$

Because $\log 2 + d\log(2M+1) \leq 6d\log n$, it is sufficient to have

$$6d\log n - \frac{nh^d t^2}{8(c_1 + c_2 t/6)} \leq \log \varepsilon .$$

It is clear there exists a positive t^* such that the above inequality attains equality, and that this inequality holds for all $t \geq t^*$. To get t^*, we restrict ourselves to $t > 0$, so that we have $c_1 + c_2 t/6 > 0$. Then we solve for t^* as the bigger root of a quadratic function in t:

$$t^* = \frac{1}{2nh^d}\left(8c_2 d\log n - \frac{4}{3}\log\varepsilon c_2 + \sqrt{(8c_2 d\log n - \frac{4}{3}\log\varepsilon c_2)^2 - 4nh^d(8c_1\log\varepsilon - 48dc_1\log n)}\right).$$

Observe that for positive a and b, $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, so it holds

$$\begin{aligned} t^* &\leq \frac{1}{2nh^d}\left(2(8c_2 d\log n - \frac{4}{3}\log\varepsilon c_2) + \sqrt{4nh^d(48dc_1\log n - 8c_1\log\varepsilon)}\right) \\ &\leq 32c_2 d\frac{\log\frac{n}{\varepsilon}}{nh^d} + \sqrt{48dc_1}\sqrt{\frac{\log\frac{n}{\varepsilon}}{nh^d}} \\ &\leq (32c_2 d + \sqrt{48dc_1})\sqrt{\frac{\log\frac{n}{\varepsilon}}{nh^d}}, \end{aligned}$$

in which the last inequality holds for n such that $\sqrt{\frac{\log(n/\varepsilon)}{nh^d}} < 1$. Then we can take

$$\delta = (32c_2 d + \sqrt{48dc_1})\sqrt{\frac{\log\frac{n}{\varepsilon}}{nh^d}} + 2Lc_3 h^\beta + \frac{1}{nh^d}\frac{\sqrt{d}L'}{nh} + \left(L + \tilde{C}\sum_{1\leq|s|\leq\lfloor\beta\rfloor}\frac{1}{s!}\right)d^{\beta/2}n^{-2\beta}.$$

When $h = \left(\frac{\log n}{n}\right)^{\frac{1}{2\beta+d}}$, we have $\sqrt{\frac{\log n}{nh^d}} = h^\beta = \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}}$, and $nh > 1$. Also, it is safe to assume that $d^{\beta/2}n^{-2\beta} \leq \sqrt{\log(n/\varepsilon)/(nh^d)}$, Therefore,

$$\begin{aligned} \delta &\leq (32c_2 d + \sqrt{48dc_1})\sqrt{\frac{\log\frac{n}{\varepsilon}}{nh^d}} + 2Lc_3\sqrt{\frac{\log n}{nh^d}} + \sqrt{d}L'\sqrt{\frac{\log n}{nh^d}} + \left(L + \tilde{C}\sum_{1\leq|s|\leq\lfloor\beta\rfloor}\frac{1}{s!}\right)\frac{d^{\beta/2}}{n^{2\beta}} \\ &\leq C\sqrt{\frac{\log\frac{n}{\varepsilon}}{nh^d}}, \end{aligned}$$

where $C = 32c_2 d + \sqrt{48dc_1} + 2Lc_3 + \sqrt{d}L' + L + \tilde{C}\sum_{1\leq|s|\leq\lfloor\beta\rfloor}\frac{1}{s!}$.

■

Lemma 2 *Denote by $\hat{\eta}$ the Nadaraya-Watson estimator of the regression function η based on an i.i.d. sample $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$. Let p be the marginal density of X, $p \in \mathcal{P}_\Sigma(\beta,L,[-1,1]^d)$, $f = \eta \cdot p \in \mathcal{P}_\Sigma(\beta,L,[-1,1]^d)$, and the kernel K be β-valid and L'-Lipschitz. Moreover, assume $p \geq \mu'_{\min}(>0)$ and the sample size n is such that $\sqrt{\frac{\log(n/\varepsilon)}{nh^d}} < 1$. Then for any $\varepsilon > 0$,*

$$\mathbb{P}(\|\hat{\eta}-\eta\|_\infty \geq \delta) \leq 3\varepsilon,$$

for

$$\begin{aligned}\delta &= \frac{1}{\mu'_{\min}-\delta'}\left(\delta' + (32d\|K\|_\infty + \sqrt{12d\|K\|^2\|p\|_\infty})\sqrt{\frac{\log(n/\varepsilon)}{nh^d}} + (c_4+c_5)h^\beta\right)\\ &+ \frac{1}{\mu'_{\min}-\delta'}\left(\frac{1}{nh^d}\frac{\sqrt{d}L'}{nh} + \left(L + \tilde{C}_1 \sum_{1\leq|s|\leq\lfloor\beta\rfloor}\frac{1}{s!}\right)\frac{d^{\beta/2}}{n^{2\beta}}\right),\end{aligned}$$

where δ' is the same as δ in Lemma 1, $c_4 = \frac{\|p\|_\infty L}{\mu'_{\min}}\left(1+\frac{\|f\|_\infty}{\mu'_{\min}}\right)\left(\int|K(z)|\cdot\|z\|^\beta dz\right)$ and $c_5 = L\int|K(z)|\cdot\|z\|^\beta dz$, and $\tilde{C}_1$ is such that $\tilde{C}_1 \geq \sup_{1\leq|s|\leq\lfloor\beta\rfloor}\sup_{x\in[-1,1]^d}|D^s p(x)|$.

In particular, when $h = (\frac{\log n}{n})^{\frac{1}{2\beta+d}}$, there exists some positive D, such that we can take $\delta = D\sqrt{\frac{\log(n/\varepsilon)}{nh^d}}$.

Proof Recall that $\hat{p}(x) = \frac{1}{nh^d}\sum_{i=1}^n K(\frac{X_i-x}{h})$ and $\hat{f}(x) = \frac{1}{nh^d}\sum_{i=1}^n Y_i K(\frac{X_i-x}{h})$, so

$$|\hat{\eta}-\eta| = \left|\frac{\hat{f}}{\hat{p}}-\frac{f}{p}\right| \leq \left|\frac{f}{\hat{p}}-\frac{f}{p}\right| + \left|\frac{\hat{f}}{\hat{p}}-\frac{f}{\hat{p}}\right| = |f|\frac{|\hat{p}-p|}{|\hat{p}||p|} + \frac{1}{|\hat{p}|}|\hat{f}-f|.$$

This implies that

$$\mathbb{P}(\|\hat{\eta}-\eta\|_\infty \geq \delta) \leq \mathbb{P}\left(\left\||f|\frac{|\hat{p}-p|}{|\hat{p}||p|}\right\|_\infty + \left\|\frac{1}{|\hat{p}|}|\hat{f}-f|\right\|_\infty \geq \delta\right).$$

Therefore, to claim $\mathbb{P}(\|\hat{\eta}-\eta\|_\infty \geq \delta) \leq 3\varepsilon$, it is enough to show that for δ_1 and δ_2 such that $\delta = \delta_1+\delta_2$, it holds

$$\text{i) } \mathbb{P}\left(\left\||f|\frac{|\hat{p}-p|}{|\hat{p}||p|}\right\|_\infty \geq \delta_1\right) \leq \varepsilon, \text{ and ii) } \mathbb{P}\left(\left\|\frac{1}{|\hat{p}|}|\hat{f}-f|\right\|_\infty \geq \delta_2\right) \leq 2\varepsilon, \tag{15}$$

where the quantities δ_1 and δ_2 will be specified later.

i). We prove the first inequality in (15). Because $\eta \leq 1$ and $\eta(x) = f(x)/p(x)$,

$$\mathbb{P}\left(\left\||f|\frac{|\hat{p}-p|}{|\hat{p}||p|}\right\|_\infty \geq \delta_1\right) \leq \mathbb{P}\left(\left\|\frac{\hat{p}-p}{\hat{p}}\right\|_\infty \geq \delta_1\right).$$

Denote an event regarding the sample $\mathcal{E} = \{\|\hat{p}-p\|_\infty < \delta'\}$, where δ' is the same as δ in Lemma 1. So by Lemma 1, $\mathbb{P}(\mathcal{E}) > 1-\varepsilon$. Moreover,

$$\begin{aligned}&\mathbb{P}(\|(\hat{p}-p)/\hat{p}\|_\infty \geq \delta_1)\\ \leq\ &\mathbb{P}\left(\|(\hat{p}-p)/\hat{p}\|_\infty \geq \delta_1, \|\hat{p}-p\|_\infty \geq \delta'\right) + \mathbb{P}\left(\|(\hat{p}-p)/\hat{p}\|_\infty \geq \delta_1, \|\hat{p}-p\|_\infty < \delta'\right)\\ \leq\ &\mathbb{P}(\|\hat{p}-p\|_\infty \geq \delta') + \mathbb{P}\left(\|\hat{p}-p\|_\infty \geq \delta_1(\mu'_{\min}-\delta'), \|\hat{p}-p\|_\infty < \delta'\right).\end{aligned}$$

Take $\delta_1 = \frac{\delta'}{\mu'_{\min} - \delta'}$, then we have $\delta' = \delta_1(\mu'_{\min} - \delta')$. So

$$\mathbb{P}\left(\left\| |f| \frac{|\hat{p} - p|}{|\hat{p}||p|} \right\|_\infty \geq \delta_1 \right) \leq \mathbb{P}(\|\hat{p} - p\|_\infty \geq \delta') \leq \varepsilon .$$

ii). Now we prove the second inequality in (15). Note that

$$\begin{aligned} & \mathbb{P}\left(\left\| \frac{\hat{f} - f}{\hat{p}} \right\|_\infty \geq \delta_2 \right) \\ = \; & \mathbb{P}\left(\left\| \frac{\hat{f} - f}{\hat{p}} \right\|_\infty \geq \delta_2, \|\hat{p} - p\|_\infty \geq \delta' \right) + \mathbb{P}\left(\left\| \frac{\hat{f} - f}{\hat{p}} \right\|_\infty \geq \delta_2, \|\hat{p} - p\|_\infty < \delta' \right) \\ \leq \; & \mathbb{P}(\|\hat{p} - p\|_\infty \geq \delta') + \mathbb{P}\left(\|\hat{f} - f\|_\infty \geq \delta_2(\mu'_{\min} - \delta') \right) \\ = \; & \varepsilon + \mathbb{P}\left(\|\hat{f} - f\|_\infty \geq \delta_2(\mu'_{\min} - \delta') \right) . \end{aligned}$$

Therefore, to bound the tail probability of $\|(\hat{f} - f)/\hat{p}\|_\infty$, it remains to show

$$\mathbb{P}(\|\hat{f} - f\|_\infty \geq \delta_2(\mu'_{\min} - \delta')) \leq \varepsilon .$$

Let G be the same collection of vertices of sub-cubes in $[-1,1]^d$ as in the proof of Lemma 1, and denote by $M = n^2$. Note that for every $\delta_3 > 0$, it holds

$$\mathbb{P}(\|\hat{f} - f\|_\infty \geq \delta_3) \quad \leq \quad \mathbb{P}(M_1 + M_2 + M_3 \geq \delta_3) ,$$

where

$$\begin{aligned} M_1 &= \sup_{\|x - x'\| \leq \frac{\sqrt{d}}{M}} |\hat{f}(x) - \hat{f}(x')| , \\ M_2 &= \sup_{\|x - x'\| \leq \frac{\sqrt{d}}{M}} |f(x) - f(x')| , \\ M_3 &= \sup_{x \in G} |\hat{f}(x) - f(x)| . \end{aligned}$$

The quantity M_1 can be controlled as follows:

$$\begin{aligned} M_1 \; &= \; \sup_{\|x - x'\| \leq \frac{\sqrt{d}}{M}} \left| \frac{1}{nh^d} \sum_{i=1}^n Y_i K(\frac{X_i - x}{h}) - \frac{1}{nh^d} \sum_{i=1}^n Y_i K(\frac{X_i - x'}{h}) \right| \\ &\leq \; \sup_{\|x - x'\| \leq \frac{\sqrt{d}}{M}} \frac{1}{nh^d} \sum_{i=1}^n \left| K(\frac{X_i - x}{h}) - K(\frac{X_i - x'}{h}) \right| \\ &\leq \; \frac{1}{nh^d} \frac{n\sqrt{d}L'}{Mh} = \frac{1}{nh^d} \frac{\sqrt{d}L'}{nh} . \end{aligned}$$

The quantity M_2 can be controlled similarly as its counterpart in proof for Lemma 1,

$$M_2 \leq \left(L + \tilde{C}_1 \sum_{1 \leq |s| \leq \lfloor \beta \rfloor} \frac{1}{s!} \right) d^{\frac{\beta}{2}} n^{-2\beta} .$$

Let $t = \delta_3 - \frac{1}{nh^d}\frac{\sqrt{d}L'}{nh} - \left(L + \tilde{C}_1 \sum_{1\le|s|\le\lfloor\beta\rfloor} \frac{1}{s!}\right) d^{\frac{\beta}{2}} n^{-2\beta}$, then

$$\mathbb{P}(\|\hat{f} - f\|_\infty \ge \delta_3) \le \mathbb{P}(M_3 \ge t)\,.$$

Use a union bound to control the tail probability of M_3:

$$\mathbb{P}(M_3 \ge t) \le \sum_{x\in G} \mathbb{P}\left(|\hat{f}(x) - f(x)| \ge t\right)\,.$$

For each fixed $x \in G$,

$$\begin{aligned}
\hat{f}(x) - f(x) &= \frac{1}{nh^d}\sum_{i=1}^n Y_i K(\frac{X_i - x}{h}) - \eta(x)\cdot p(x) \\
&= \frac{1}{nh^d}\sum_{i=1}^n Y_i K(\frac{X_i - x}{h}) - \mathbb{E}\left[\frac{1}{nh^d}\sum_{i=1}^n Y_i K(\frac{X_i - x}{h})\right] \\
&+ \mathbb{E}\left[\frac{1}{nh^d}\sum_{i=1}^n Y_i K(\frac{X_i - x}{h})\right] - \eta(x)\cdot p(x) \\
&= B_1(x) + B_2(x)\,,
\end{aligned}$$

where

$$\begin{aligned}
B_1(x) &= \frac{1}{nh^d}\sum_{i=1}^n Y_i K(\frac{X_i - x}{h}) - \mathbb{E}\left[\frac{1}{nh^d}\sum_{i=1}^n Y_i K(\frac{X_i - x}{h})\right]\,, \\
B_2(x) &= \mathbb{E}\left[\frac{1}{nh^d}\sum_{i=1}^n Y_i K(\frac{X_i - x}{h})\right] - \eta(x)\cdot p(x)\,.
\end{aligned}$$

This implies that

$$\mathbb{P}(M_1 \ge t) \le \sum_{x\in G} \mathbb{P}(|B_1(x)| + |B_2(x)| \ge t)\,.$$

The tail probability of $|B_1(x)|$ is controlled by invoking the Bernstein's inequality. Denote by $Z_i = Z_i(x) = \frac{1}{h^d} Y_i K(\frac{X_i - x}{h}) - \mathbb{E}\left[\frac{1}{h^d} Y_i K(\frac{X_i - x}{h})\right]$. It is clear that $\mathbb{E}(Z_i) = 0$, $|Z_i| \le 2\|K\|_\infty h^{-d}$. Moreover,

$$Var(Z_i) \le \mathbb{E}\left(h^{-2d} K^2(\frac{X_i - x}{h})\right) = \int h^{-d} K^2(y) p(x + yh) dy \le \|K\|^2 \|p\|_\infty h^{-d}\,.$$

Therefore for any $t_1 > 0$,

$$\begin{aligned}
&\sum_{x\in G} \mathbb{P}\left(|B_1(x)| \ge t_1\right) = \sum_{x\in G} \mathbb{P}\left(\frac{1}{n}|\sum_{i=1}^n Z_i| \ge t_1\right) \\
&\le 2(2M+1)^d \exp\left(-\frac{nt_1^2}{2\|K\|^2\|p\|_\infty h^{-d} + 4/3\|K\|_\infty h^{-d} t_1}\right)\,.
\end{aligned}$$

To have the last display bounded from above by $\varepsilon \in (0,1)$, we recycle the arguments in the proof for Lemma 1 to find out that t_1 should be greater than or equal to

$$t_1^* = \left(32d\|K\|_\infty + \sqrt{12d\|K\|^2\|p\|_\infty}\right)\sqrt{\frac{\log(n/\varepsilon)}{nh^d}},$$

provided the sample size n is such that $\sqrt{\frac{\log(n/\varepsilon)}{nh^d}} < 1$.

Decompose $B_2(x)$ into two parts,

$$\begin{aligned} B_2(x) &= \left\{\mathbb{E}\left[\tfrac{1}{nh^d}\textstyle\sum_{i=1}^n K(\tfrac{X_i-x}{h})\eta(X_i)\right] - \mathbb{E}[\hat{p}(x)\eta(x)]\right\} \\ &\quad + \left\{\mathbb{E}[\hat{p}(x)\eta(x)] - p(x)\eta(x)\right\}. \end{aligned}$$

Note that

$$\begin{aligned} &\left|\mathbb{E}\left[\frac{1}{nh^d}\sum_{i=1}^n K(\frac{X_i-x}{h})\eta(X_i)\right] - \mathbb{E}[\hat{p}(x)\eta(x)]\right| \\ &= \left|\int \frac{1}{h^d}K(\frac{y-x}{h})(\eta(y)-\eta(x))p(y)dy\right| \\ &= \left|\int K(z)[\eta(x+hz)-\eta(x)]p(x+hz)dz\right| \\ &\le \|p\|_\infty \int |K(z)|\cdot|\eta(x+hz)-\eta(x)|dz. \end{aligned} \tag{16}$$

Note that

$$\begin{aligned} |\eta(x+hz)-\eta(x)| &= \left|\frac{f(x+hz)}{p(x+hz)} - \frac{f(x)}{p(x)}\right| \\ &\le \frac{1}{\mu'_{\min}}\left|f(x+hz) - f(x)\frac{p(x+hz)}{p(x)}\right|. \end{aligned}$$

It follows from $p \in \mathcal{P}_\Sigma(L,\beta,[-1,1]^d)$ that

$$\left|\frac{p(x+hz)}{p(x)} - 1\right| \le \frac{L\|z\|^\beta h^\beta}{\mu'_{\min}}.$$

Therefore,

$$\begin{aligned} &|\eta(x+hz)-\eta(x)| \\ &\le \frac{1}{\mu'_{\min}}\left(|f(x+hz)-f(x)| + |f(x)|\cdot\frac{L}{\mu'_{\min}}\|z\|^\beta h^\beta\right) \\ &\le \left(1 + \frac{\|f\|_\infty}{\mu'_{\min}}\right)\frac{L}{\mu'_{\min}}\|z\|^\beta h^\beta, \end{aligned}$$

where the last inequality follows from $f \in \mathcal{P}_\Sigma(\beta,L,[-1,1]^d)$. The above inequality together with (16) implies that

$$\left|\mathbb{E}\left[\frac{1}{nh^d}\sum_{i=1}^n K(\frac{X_i-x}{h})\eta(X_i)\right] - \mathbb{E}[\hat{p}(x)\cdot\eta(x)]\right| \le c_4 h^\beta,$$

where $c_4 = \frac{\|p\|_\infty L}{\mu'_{\min}} \left(1 + \frac{\|f\|_\infty}{\mu'_{\min}}\right) \left(\int |K(z)| \cdot \|z\|^\beta dz\right)$.

Now we control the second part of $B_2(x)$. Because $p \in \mathcal{P}_\Sigma(\beta, L, [-1,1]^d)$, we have via similar lines to the proof of Lemma 4.1 in Rigollet and Vert (2009),

$$|\mathbb{E}[\hat{p}(x)\eta(x)] - p(x)\eta(x)| \leq |\eta(x)| \cdot |\mathbb{E}\hat{p}(x) - p(x)| \leq c_5 h^\beta,$$

where $c_5 = L\int |K(z)| \cdot \|z\|^\beta dz$. Therefore,

$$|B_2(x)| \leq (c_4 + c_5) h^\beta.$$

Taking $\tilde{t} = (32d\|K\|_\infty + \sqrt{12d\|K\|^2\|p\|_\infty})\sqrt{\frac{\log(n/\varepsilon)}{nh^d}} + (c_4 + c_5)h^\beta$, and $\delta_3 = \tilde{t} + \frac{1}{nh^d}\frac{\sqrt{d}L'}{nh} + \left(L + \tilde{C}_1 \sum_{1 \leq |s| \leq \lfloor\beta\rfloor} \frac{1}{s!}\right) d^{\beta/2} n^{-2\beta}$, we have

$$\mathbb{P}(\|\hat{f} - f\|_\infty \geq \delta_3) \leq \mathbb{P}(M_1 \geq \tilde{t}) \leq \sum_{x \in G} \mathbb{P}\left(|B_1(x)| \geq \tilde{t} - (c_4 + c_5)h^\beta\right) \leq \varepsilon.$$

Take $\delta_2 = \frac{\delta_3}{\mu'_{\min} - \delta'}$, we have $\mathbb{P}(\|\hat{f} - f\|_\infty \geq \delta_2(\mu'_{\min} - \delta')) \leq \varepsilon$.

To conclude, part i) and part ii) in (15) together close the proof.

■

References

M. Agyemang, K. Barker, and R. Alhajj. A comprehensive survey of numeric and symbolic outlier mining techniques. *Intelligent Data Analysis*, 6:521–538, 2006.

J. Audibert and A. Tsybakov. Fast learning rates for plug-in classifiers under the margin condition. *Annals of Statistics*, 35:608–633, 2007.

G. Blanchard, G. Lee, and G. Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11:2973–3009, 2010.

A. Cannon, J. Howse, D. Hush, and C. Scovel. Learning with the neyman-pearson and min-max criteria. *Technical Report LA-UR-02-2951*, 2002.

D. Casasent and X. Chen. Radial basis function neural networks for nonlinear fisher discrimination and neyman-pearson classification. *Neural Networks*, 16(5-6):529 – 535, 2003.

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 09:1–72, 2009.

C. Elkan. The foundations of cost-sensitive learning. *In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 973–978, 2001.

E. Giné, V. Koltchinskii, and L. Sakhanenko. Kernel density estimators: convergence in distribution for weighted sup norms. *Probability Theory and Related Fields*, 130:167–198, 2004.

M. Han, D. Chen, and Z. Sun. Analysis to Neyman-Pearson classification with convex loss function. *Analysis in Theory and Applications*, 24(1):18–28, 2008.

V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Rivew*, 2:85–126, 2004.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005. ISBN 0-387-98864-5.

J. Lei, A. Rinaldo, and L. Wasserman. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 2013.

O. Lepski. Multivariate density estimation under sup-norm loss: oracle approach, adaptation and independence structure. *Annals of Statistics*, 41(2):1005–1034, 2013.

E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *Annals of Statistics*, 27:1808–1829, 1999.

M. Markou and S. Singh. Novelty detection: a review-part 1: statistical approahces. *Signal Processing*, 12:2481–2497, 2003a.

M. Markou and S. Singh. Novelty detection: a review-part 2: network-based approahces. *Signal Processing*, 12:2499–2521, 2003b.

A. Patcha and J.M. Park. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, 12:3448–3470, 2007.

W. Polonik. Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *Annals of Statistics*, 23:855–881, 1995.

P. Rigollet and X. Tong. Neyman-pearson classification, convexity and stochastic constraints. *Journal of Machine Learning Research*, 12:2831–2855, 2011.

P. Rigollet and R. Vert. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, 15(4): 1154–1178, 2009.

C. Scott. Comparison and design of neyman-pearson classifiers. Unpublished, 2005.

C. Scott. Performance measures for Neyman-Pearson classification. *IEEE Transactions on Information Theory*, 53(8):2852–2863, 2007.

C. Scott and R. Nowak. A neyman-pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11):3806–3819, 2005.

B. Tarigan and S. van de Geer. Classifiers of support vector machine type with l1 complexity regularization. *Bernoulli*, 12:1045–1076, 2006.

A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32: 135–166, 2004.

A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

A. Tsybakov and S. van de Geer. Square root penalty: Adaptation to the margin in classification and in edge estimation. *Annals of Statistics*, 33:1203–1224, 2005.

D. Wied and R. Weiβbach. Consistency of the kernel density estimator: a survey. *Statistical Papers*, 53(1):1–21, 2010.

Y. Yang. Minimax nonparametric classification-part i: rates of convergence. *IEEE Transaction Information Theory*, 45:2271–2284, 1999.

B. Zadrozny, J. Langford, and N. Abe. Cost-sensitive learning by cost-proportionate example weighting. *IEEE International Conference on Data Mining*, page 435, 2003.