Biometrika (2014), **xx**, x, *pp*. 1–15
© 2007 Biometrika Trust
Printed in Great Britain

Classification with Confidence

BY JING LEI

Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania, 15215, U.S.A

jinglei@andrew.cmu.edu

SUMMARY

A framework of classification is developed with a notion of confidence. In this framework, a classifier consists of two tolerance regions in the predictor space with a specified coverage level for each class. The classifier also produces an ambiguous region where the classification needs further investigation. Theoretical analysis reveals interesting structures of the confidence-ambiguity trade-off, and characterizes the optimal solution by extending the Neyman–Pearson lemma. We provide general estimating procedures, with rates of convergence, based on estimates of the conditional probabilities. The method can be easily implemented with good robustness, as illustrated through theory, simulation and a data example.

Some key words: Classification; Confidence level; Consistency; Neyman–Pearson lemma; Level set

1. INTRODUCTION

In the binary classification problem, the data consist of independent pairs of random variables $(Y_i, X_i) \in \{0,1\} \times \mathcal{X}$ $(i = 1, ..., n)$ from a common distribution P, where $\mathcal{X}$ is usually a subset of $\mathbb{R}^d$. The classical inference task is to find a mapping $f : \mathcal{X} \mapsto \{0,1\}$ with small misclassification probability $\mathrm{pr}\{f(X) \neq Y\}$. The misclassification probability is minimized by the Bayes classifier $f_{\mathrm{Bayes}}(x) = \mathbb{I}\{\eta(x) \geq 1/2\}$, where $\eta(x) = \mathrm{pr}(Y = 1 \mid X = x)$ and $\mathbb{I}(\cdot)$ is the indicator function. While the misclassification probability serves as a good assessment of the overall performance, it does not directly provide confidence for a classification decision on each observation, and classification results near the boundary can be much worse than elsewhere.

In this work we develop a new framework of classification with the notion of confidence and efficiency. Let P_j be the conditional distribution of X given $Y = j$, for $j = 0, 1$. Given $\alpha_0, \alpha_1 \in (0,1)$, we look for classification regions C_0 and C_1 such that $\mathcal{X} = C_0 \cup C_1$, and $P_j(C_j) \geq 1 - \alpha_j$, $(j = 0, 1)$. Here α_j, chosen by the user, is the tolerated noncoverage rate for class j. Now C_0 and C_1 may overlap and give an ambiguous region $C_{01} \equiv C_0 \cap C_1$. The sets C_0 and C_1 define a set-valued classifier

$$h(x) = \begin{cases} \{1\}, & x \in C_1 \backslash C_{01}, \\ \{0\}, & x \in C_0 \backslash C_{01}, \\ \{0,1\}, & x \in C_{01}. \end{cases} \tag{1}$$

We measure the efficiency of h by $\mathrm{pr}(X \in C_{01})$, which is called the ambiguity of h. A good classifier can be found by solving the optimization problem

$$\begin{aligned} &\text{minimize } \ \mathrm{pr}(X \in C_0 \cap C_1) \\ &\text{subject to } \ C_0 \cup C_1 = \mathcal{X}, \ \ P_j(C_j) \geq 1 - \alpha_j, \ j = 0, 1. \end{aligned} \tag{2}$$

The quality of a classifier is measured by the coverage and ambiguity. There is a trade-off between these two competing quantities.

The proposed framework allows customized error control for both classes, which can be useful in many applications, especially when the cost of misclassification is high in one class or both. For example, in medical screening, a misclassification may result in missed medical care or a waste of resources (Hanczar & Dougherty, 2008; Nadeem et al., 2010). In gene expression studies for biomarker identification, it is important to control the level of false negatives, which is hard to achieve using conventional methods because true biomarkers are rare. Definitive classifiers, such as Neyman–Pearson classification (Scott & Nowak, 2005; Han et al., 2008; Rigollet & Tong, 2011; Tong, 2013), usually cannot guarantee accuracy for both classes.

Set-valued classifiers with an associated level of confidence were considered by Shafer & Vovk (2008) and Vovk et al. (2009). The relationship between several notions of confidence is discussed in Vovk (2013) and Lei & Wasserman (2014). The novel parts of our framework include the notion of ambiguity as a measure of statistical efficiency, formulating the problem by minimizing ambiguity with a confidence guarantee, and deriving optimal solutions and estimation methods.

Another related topic is classification with a reject option, where the classifier may not output a definitive classification. Chow (1970) considered this under a decision-theoretic framework. Herbei & Wegkamp (2006) and Yuan & Wegkamp (2010) studied plug-in methods and empirical risk minimization. In our framework, the ambiguous region corresponds to the cases where the classifier does not give definitive output. Despite this similarity, classification with reject option still aims at minimizing an overall misclassification risk, while our framework allows the user to customize the desired level of confidence for each class.

2. THE FRAMEWORK

2·1. *Class-specific coverage*

Assume that P_j, the distribution of X given $Y = j$, is continuous with density function p_j for $j = 0, 1$. Let $\pi_j = \mathrm{pr}(Y = j)$, $(j = 0, 1)$. Then

$$\eta(x) = \mathrm{pr}(Y = 1 \mid X = x) = \frac{\pi_1 p_1(x)}{\pi_0 p_0(x) + \pi_1 p_1(x)}.$$

We assume that $\eta(X)$ is a continuous random variable, and that π_0, π_1 are positive constants. See Remark 2 for the case that $\eta(X)$ is not continuous. For any C_0, C_1, the ambiguity is

$$P(C_0 \cap C_1) = \pi_0 P_0(C_{01}) + \pi_1 P_1(C_{01}) \leq \pi_0 P_0(C_1) + \pi_1 P_1(C_0).$$

According to the Neyman–Pearson lemma, the set C_0 that minimizes $P_1(C_0)$ subject to $P_0(C_0^c) \leq \alpha_0$ is a level set of η, and a similar result holds for C_1, where C_0^c denotes the complement of C_0. The following theorem says that C_0 and C_1 constructed from the level sets of η are indeed optimal for problem (2). It can be viewed as an extension of the Neyman–Pearson lemma.

THEOREM 1. *Fix* $0 < \alpha_0, \alpha_1 < 1$. *A solution to the optimization problem* (2) *is*

$$C_0 = \{x : \eta(x) \leq t_0\}, \qquad C_1 = \{x : \eta(x) \geq t_1\} \cup C_0^c,$$

where $t_0 = t_0(\alpha_0)$, $t_1 = t_1(\alpha_1)$ *are chosen such that* $P_0\{\eta(X) \leq t_0\} = 1 - \alpha_0$, *and* $P_1\{\eta(X) \geq t_1\} = 1 - \alpha_1$.

Remark 1. If α_0 and α_1 are small, then we expect that $t_1 < t_0$ and all sets C_0 and C_1 satisfying the constraints in (2) must overlap. In this paper we focus on the case $t_1 < t_0$ and $C_0 \cup C_1 = \mathcal{X}$.

When $t_1 > t_0$, the solution to (2) is not unique, and the additional part C_0^c is included in the definition of C_1 in Theorem 1 so that $C_0 \cup C_1 = \mathcal{X}$. If $t_1 > t_0$, one may change the optimization problem to $\mathrm{minimize}\ P(C_0 \cup C_1)$, subject to $P_j(C_j^c) \leq \alpha_j$ $(j = 0, 1)$ and $C_{01} = \emptyset$. It follows from the proof of Theorem 1 that the unique optimal solution to this new problem is $C_0 = \{x : \eta(x) \leq t_0\}$, $C_1 = \{x : \eta(x) \geq t_1\}$. Now $\mathcal{X} \backslash (C_0 \cup C_1) \neq \emptyset$, which corresponds to a region of empty classification, indicating instances that are atypical in both classes and may even suggest a new, unseen class.

Remark 2. When the distribution of $\eta(X)$ is not continuous at t_0 or t_1, define $L(t) \equiv \{x : \eta(x) \leq t\}$, $L(t^-) \equiv \{x : \eta(x) < t\}$, and let $t_0 = \inf\{t : P_0\{L(t)\} \geq 1 - \alpha_0\}$. If $P_0\{L(t_0^-)\} < 1 - \alpha_0 < P_0\{L(t_0)\}$, we can choose any C_0 such that $L(t_0^-) \subseteq C_0 \subseteq L(t_0)$ and $P_0(C_0) = 1 - \alpha_0$, provided that X is continuous on $L(t_0)\backslash L(t_0^-)$. When X is not continuous on $L(t_0)\backslash L(t_0^-)$, we can use randomized rules to achieve exact coverage. The set C_1 can be constructed similarly.

2·2. *Overall coverage*

When the overall coverage is of interest, the optimization problem becomes

$$\begin{aligned} &\mathrm{minimize}\ \ P(C_0 \cap C_1) \\ &\mathrm{subject\ to}\ \ C_0 \cup C_1 = \mathcal{X},\ \ \pi_0 P_0(C_0) + \pi_1 P_1(C_1) \geq 1 - \alpha. \end{aligned} \tag{3}$$

We assume that $\alpha < \min(\pi_0, \pi_1)$, otherwise one can classify everything to the majority class to achieve the desired coverage with no ambiguity.

For $(\alpha_0, \alpha_1) \in (0, 1)^2$, let $\nu(\alpha_0, \alpha_1)$ be the optimal value in problem (2), which is the minimum ambiguity. The next lemma, proved in Appendix A·1, characterizes the solution to (3).

LEMMA 1. *Suppose that (C_0^*, C_1^*) achieves the minimum of problem (3). Then (C_0^*, C_1^*) is a minimizer of problem (2) with $(\alpha_0, \alpha_1) = (\alpha_0^*, \alpha_1^*)$, where (α_0^*, α_1^*) solves the optimization problem*

$$\mathrm{minimize}\ \nu(\alpha_0, \alpha_1),\ \ \mathrm{subject\ to}\ \ 0 \leq \alpha_0, \alpha_1 \leq 1,\ \ \pi_0\alpha_0 + \pi_1\alpha_1 = \alpha. \tag{4}$$

According to Lemma 1 and Theorem 1, a strategy to solve (3) is to first obtain (α_0^*, α_1^*) by solving (4), and then solve (2) with (α_0^*, α_1^*). How do we solve (4)? For any $\alpha_0 \in (0, \alpha/\pi_0)$, the constraint requires $\alpha_1 = \alpha_1(\alpha_0) = (\alpha - \pi_0\alpha_0)/\pi_1$, so the objective function ν is a function of α_0. The next lemma shows that ν is a convex function when $\eta(X)$ has a continuous distribution. It confirms the intuition that the best α_0 must balance errors from both classes and the ambiguity will first decrease and then increase as α_0 varies from 0 to α/π_0. It also lays the foundation of our algorithm to solve (4), because strong convexity enables us to approximate α_0^* by minimizing $\hat{\nu}(\cdot)$, a good approximation to $\nu(\cdot)$.

LEMMA 2. *Let $\nu(\alpha_0) = \nu\{\alpha_0, \alpha_1(\alpha_0)\}$. If the distributions of $\eta(X)$ under P_0 and P_1 are continuous, then $\nu(\alpha_0)$ is a convex function.*

The proof of Lemma 2 in Appendix A·1 gives explicit expressions for $\nu'(\alpha_0)$ and $\nu''(\alpha_0)$. Moreover, if $\eta(X)$ has positive density on its support, then $\lim_{\alpha_0 \downarrow 0} \nu'(\alpha_0) < 0$ and $\lim_{\alpha_0 \uparrow (\alpha/\pi_0)} \nu'(\alpha_0) > 0$, and there is an $\alpha_0^* \in (0, \alpha/\pi_0)$ that minimizes $\nu(\alpha_0)$. Finally, if $\nu(\alpha_0) > 0$ for all $\alpha_0 \in (0, \alpha/\pi_0)$ then ν is strictly convex and the minimizer α_0^* is unique.

3. ESTIMATION PROCEDURES

3·1. *Estimation with class-specific coverage*

Let $\{(X_i, Y_i) : 1 \le i \le n\}$ be a random sample from P. Recall that $\mathbb{I}(\cdot)$ is the indicator function. For $j \in \{0,1\}$, let $n_j = \#\{i : Y_i = j\}$ and define

$$\hat{\pi}_j = n_j/n, \;\; \hat{P}_j(C) = n_j^{-1} \sum_{i:Y_i=j} \mathbb{I}(X_i \in C).$$

For $j = 0,1$, let $X_{j,1}, ..., X_{j,n_j}$ be the sample predictors in class j. Given $\alpha_j \in (0,1)$ and any estimate $\hat{\eta}(x)$ of $\eta(x) = \mathrm{pr}(Y = 1 \mid X = x)$, the classification region C_j is estimated by

$$\hat{C}_0 = \left\{x : \hat{\eta}(x) \le \hat{t}_0(\alpha_0)\right\}, \;\; \hat{C}_1 = \left\{x : \hat{\eta}(x) \ge \hat{t}_1(\alpha_1)\right\}, \tag{5}$$

where $\hat{t}_0(\alpha_0)$ is the $\lfloor n_0\alpha_0 \rfloor$th largest value in $\{\hat{\eta}(X_{0,1}), ..., \hat{\eta}(X_{0,n_0})\}$, and $\hat{t}_1(\alpha_1)$ is the $\lfloor n_1\alpha_1 \rfloor$th smallest value in $\{\hat{\eta}(X_{1,1}), ..., \hat{\eta}(X_{1,n_1})\}$. Here $\hat{t}_j(\alpha_j)$ is an empirical version of t_j given in Theorem 1. The performance of $\hat{C}_0$ and $\hat{C}_1$ depends on the accuracy of $\hat{\eta}$ and the regularity of P_0, P_1. We will assume that $\hat{\eta}$ satisfies the following accuracy property.

DEFINITION 1. *An estimator $\hat{\eta}$ is (δ_n, ρ_n)-accurate if* $\mathrm{pr}(\|\hat{\eta} - \eta\|_\infty \ge \delta_n) \le \rho_n$.

As we will see from examples in Section 3·3, many common estimators $\hat{\eta}$ are (δ_n, ρ_n)-accurate with $\delta_n \to 0$ and $\rho_n \to 0$.

For the regularity of P_j, let G_j be the cumulative distribution function of $\eta(X)$ under P_j. Then we have $t_0 = G_0^{-1}(1 - \alpha_0)$ and $t_1 = G_1^{-1}(\alpha_1)$. We consider the following margin condition.

(MA) There exist positive constants b_1, b_2, ϵ_0, and γ such that for $j = 0,1$ and all $\epsilon \in [-\epsilon_0, \epsilon_0]$,

$$b_1|\epsilon|^\gamma \le |G_j(t_j + \epsilon) - G_j(t_j)| \le b_2|\epsilon|^\gamma. \tag{6}$$

Condition (MA) characterizes the steepness of G_j near the cut-off levels. Similar margin conditions have been used in the estimation of density level sets (Polonik, 1995; Tsybakov, 1997), and classification (Audibert & Tsybakov, 2007; Tong, 2013). Compared with other versions of margin condition in the literature, condition (MA) has an extra lower bound part, because our method does not assume any knowledge of the cut-off value t_j and must estimate it from the data.

THEOREM 2. *If $\hat{\eta}$ is (δ_n, ρ_n)-accurate, and (MA) holds, then for each $r > 0$ there exists a positive constant c such that with probability at least* $1 - \rho_n - n^{-r}$,

$$P_j(\hat{C}_j \triangle C_j) \le c\left\{\delta_n^\gamma + \left(\frac{\log n}{n}\right)^{\frac{1}{2}}\right\}, \;\; j = 0,1. \tag{7}$$

The proof is given in Appendix A·2. Compared with related results in density level set estimation, for example, Rigollet & Vert (2009), Theorem 2 has an extra $(\log n/n)^{1/2}$ term because we need to estimate the cut-off value t_j. A large value of γ means that P_j has little mass near the contour $\{x : \eta(x) = t_j\}$, so estimating t_j is difficult and the convergence rate will be dominated by $(\log n/n)^{1/2}$. We further discuss the sharpness of Theorem 2 in Example 1 below. Some related results in the study of tolerance regions can be found in Cadre et al. (2013) and Lei et al. (2013).

3·2. *Estimation with total coverage*

Now consider overall coverage control as described in Section 2·2. First let $\hat{\pi}_j = n_j/n$ for $j = 0, 1$ and $\hat{\alpha}_+ = \alpha/\hat{\pi}_0$. For all $\alpha_0 \in [0, \hat{\alpha}_+]$, let $\hat{\alpha}_1 = \hat{\alpha}_1(\alpha_0) = (\alpha - \hat{\pi}_0\alpha_0)/\hat{\pi}_1$, and construct $(\hat{C}_0, \hat{C}_1)$ with class-specific coverages $1 - \alpha_0$ and $1 - \hat{\alpha}_1$, respectively, using the method given in Section 3·1. Let $\hat{\nu}(\alpha_0) = n^{-1}\sum_i \mathbb{I}(X_i \in \hat{C}_{01})$ be the empirical ambiguity, and $\hat{\alpha}_0^* = \arg\min_{\alpha_0 \in [\underline{\alpha}, \bar{\alpha}]} \hat{\nu}(\alpha_0)$, where $[\underline{\alpha}, \bar{\alpha}] \subset (0, \alpha/\hat{\pi}_0)$ is an interval to be chosen by the user. The estimated classification sets $\hat{C}_0$, $\hat{C}_1$ are those corresponding to $\alpha_0 = \hat{\alpha}_0^*$. The reason for not searching over the entire range $[0, \alpha/\hat{\pi}_0]$ is to avoid complicated conditions on the function G_j near 0 and 1. In practice one can choose $[\underline{\alpha}, \bar{\alpha}] = [0.01\alpha/\hat{\pi}_0, 0.99\alpha/\hat{\pi}_0]$.

THEOREM 3. *Assume that* $0 < b_1 \le G_j'(t_j) \le b_2 < \infty$ *for all* α_0 *in an open interval containing* $[\underline{\alpha}, \bar{\alpha}]$, *that* $\alpha_0^* \in [\underline{\alpha}, \bar{\alpha}]$ *with* $\nu''(\alpha_0^*) > 0$, *and that* $\hat{\eta}$ *is* (δ_n, ρ_n)*-accurate. Then for any* $r > 0$, *there exists constant* $c > 0$ *such that* $|\hat{\alpha}_0^* - \alpha_0^*| \le c\{\delta_n^{1/2} + (\log n/n)^{1/4}\}$ *with probability at least* $1 - \rho_n - n^{-r}$ *for* n *large enough. Moreover, with same probability, the resulting* $\hat{C}_j$ $(j = 0, 1)$ *satisfy* $P_j(\hat{C}_j \Delta C_j^*) \le c'\{\delta_n^{1/2} + (\log n/n)^{1/4}\}$ *for another constant* c'.

The bound on G_j' is equivalent to assuming that (MA) holds with $\gamma = 1$ for all t_j as α_0 ranges from $\underline{\alpha}$ to $\bar{\alpha}$. The condition $\nu''(\alpha_0^*) > 0$ is satisfied when $\nu(\alpha_0) > 0$ for all α_0, as implied by Lemma 2. The rate is slower than that in Theorem 2, because the optimal α_0^* is unknown and needs to be estimated first.

3·3. *Examples*

We consider two examples of $\hat{\eta}$: local polynomial regression (Audibert & Tsybakov, 2007), and ℓ_1-penalized logistic regression (van de Geer, 2008).

Example 1. Assume that $\mathcal{X} = [-1, 1]^d$. For any $s = (s_1, ..., s_d) \in (\mathbb{Z}^+)^d$, where $\mathbb{Z}^+$ is the set of nonnegative integers, and $z \in \mathbb{R}^d$, define $z^s = z_1^{s_1} \times ... \times z_d^{s_d}$ and $|s| = \sum_{j=1}^d |s_j|$. Let K be a kernel function and $\tau > 0$ a bandwidth. The local polynomial estimator (Tsybakov, 2009) is based on the intuition of approximating $\eta(\cdot)$ at x using a polynomial of order ℓ,

$$\eta_x(z) = \sum_{s:|s|\le\ell} v_{s,x}\left(\frac{z - x}{\tau}\right)^s .$$

The local coefficients $(v_{s,x})_{s:|s|\le\ell}$ are estimated by weighted least squares,

$$(\hat{v}_{s,x})_{s:|s|\le\ell} = \arg\min_v \sum_{i=1}^n \left\{Y_i - \sum_{s:|s|\le\ell} v_s\left(\frac{X_i - x}{\tau}\right)^s\right\}^2 K\left(\frac{X_i - x}{\tau}\right),$$

and the local polynomial estimator $\hat{\eta}(x)$ is the value of $\hat{\eta}_x(z)$ evaluated at $z = x$,

$$\hat{\eta}(x) = \hat{v}_{0,x} . \qquad (8)$$

PROPOSITION 1. *In the setting of Example* 1, *assume that (a) the marginal density of* X *is bounded and bounded away from zero, and (b)* $\eta(\cdot)$ *belongs to a Hölder class with smoothness parameter* β. *Then there exist choices of kernel* K *and bandwidth* $\tau = \tau_n$ *such that for any* $r > 0$, *the local polynomial estimator* $\hat{\eta}$ *given in equation* (8) *is* (δ_n, ρ_n)*-accurate with*

$$\delta_n = c\left(\frac{\log n}{n}\right)^{\beta/(2\beta+d)}, \quad \rho_n = n^{-r},$$

where c is a positive constant depending on r.

Despite its familiar form (Stone, 1982; Tsybakov, 2009), we could not find a proof of this result in the literature. For completeness, we give more details of Proposition 1, including a proof, in the Supplementary Material.

Combining Proposition 1 and Theorem 2, if (MA) holds with $\gamma = 1$, then with probability at least $1 - 2n^{-r}$ we have

$$P_j(\hat{C}_j \triangle C_j) \leq c \left(\frac{\log n}{n}\right)^{\beta/(2\beta+d)}, \quad j = 0, 1\,.$$

When P_1 has uniform density and P_0 has bounded density, estimating C_0 is equivalent to a density level set problem, where the cut-off density level is implicitly determined by α_0. In this case, our upper bound on $P_0(\hat{C}_0 \Delta C_0)$ matches the minimax rate of density level set estimation given by Rigollet & Vert (2009). A related rate of convergence has been obtained for the excess risk of plug-in rules in Neyman–Pearson classification in a recent paper (Tong, 2013).

Example 2. Let $\mathcal{X} = [-1, 1]^d$ and $\mathrm{var}\{X(k)\} = \sigma^2$ for all $1 \leq k \leq d$, where $X(k)$ is the kth coordinate of X. Assume that $\eta(X)/\{1 - \eta(X)\} = \exp(X^T\theta^*)$ for some $\theta^* \in \mathbb{R}^d$. The coefficient θ^* and function $\eta(\cdot)$ can be estimated by first estimating θ using the ℓ_1-penalized logistic regression

$$\hat{\theta} = \arg\min n^{-1} \sum_{i=1}^{n} \left[-Y_i X_i^T \theta + \log\{1 + \exp(X_i^T \theta)\}\right] + \lambda_n \|\theta\|_1. \tag{9}$$

Such regression is typically used when d is large, and the number of true signals, $\|\theta^*\|_0$, is relatively small. The convergence of ℓ_1-penalized logistic regression has been extensively studied. The following result is a consequence of Example 1 in van de Geer (2008).

PROPOSITION 2. *Under the conditions in Example* 2*, assume that the smallest eigenvalue of $E(XX^T)$ is bounded away from zero by a constant. Let $\hat{\eta}(x) = \exp(x^T\hat{\theta})/\{1 + \exp(x^T\hat{\theta})\}$ with $\hat{\theta}$ given by* (9)*. Then one can choose λ_n such that $\hat{\eta}$ satisfies $(\delta_{n,d}, \rho_{n,d})$-accuracy with*

$$\delta_{n,d} \leq c_1 \left(\frac{\log d}{n}\right)^{\frac{1}{4}}, \quad \rho_{n,d} \leq c_2 \left\{\frac{1}{d} + \left(\frac{\log d}{n}\right)^{\frac{1}{2}} \|\theta^*\|_0\right\}$$

for positive constants c_1, c_2.

In the setting of Example 2, assume that the conditions of Proposition 2 hold and that (MA) holds with $\gamma = 1$. Let $\hat{\eta}$ and $\hat{C}_j$ be the output of ℓ_1-penalized regression with appropriate choice of λ_n. We can choose r large enough in Theorem 2, so that

$$P_j(\hat{C}_j \triangle C_j) \leq c_1 \left\{\left(\frac{\log d}{n}\right)^{\frac{1}{4}} + \left(\frac{\log n}{n}\right)^{\frac{1}{2}}\right\}, \quad j = 0, 1$$

with probability at least $1 - c_2\{d^{-1} + (\log d/n)^{1/2}\|\theta^*\|_0\}$, for some constants c_1, c_2.

4. ROBUSTNESS AND TUNING PARAMETER SELECTION

4·1. *Robust error control*

The theoretical results presented in Theorems 2 and 3 presuppose that $\hat{\eta}$ is an accurate approximation to η. In practice, such an assumption can be violated, for example, when local polynomial estimation is applied to a non-smooth distribution, or when the logistic regression model is not a good approximation to the truth. The method described in Section 3 can be easily modified such that the coverage is controlled on average in the sense that

$$E\left\{P_j(\hat{C}_j)\right\} \geq 1-\alpha_j\,, \tag{10}$$

for any distribution P and any estimator $\hat{\eta}$.

The modified method uses sample splitting. First the index set $\{1,...,n\}$ is randomly split into two parts, $\mathcal{I}_1$ and $\mathcal{I}_2$. Then $\hat{\eta}$ is estimated from the fitting subsample $\{(Y_i, X_i) : i \in \mathcal{I}_1\}$, and evaluated on the ranking subsample $\{X_i : i \in \mathcal{I}_2\}$, yielding $\mathcal{Z}_j = \{\hat{\eta}(X_i) : i \in \mathcal{I}_2, Y_i = j\}$, $j = 0, 1$. Let $\hat{t}_j$ be the $\lfloor |\mathcal{Z}_j|\alpha_j \rfloor$th largest or smallest value in $\mathcal{Z}_j$, according to $j = 0$ or 1. The estimator is $\hat{C}_0 = \{x : \hat{\eta}(x) \leq \hat{t}_0\}$, $\hat{C}_1 = \{x : \hat{\eta}(x) \geq \hat{t}_1\}$. The extension to total coverage is straightforward.

PROPOSITION 3. *For $j = 0, 1$, let $\hat{C}_j$ be given by the sample splitting procedure. If $(Y_i, X_i)_{i=1}^n$ are independent and identically distributed, then the corresponding set-valued classifier satisfies (*10*).*

Proposition 3 is based on a result in conformal prediction (Vovk et al., 2005). A proof can be found in Lemma 2.1 of Lei et al. (2014). The only assumptions are independence and identical distribution. Proposition 3 requires no assumption on the estimator $\hat{\eta}$ or the underlying distribution. When the class proportion $\mathrm{pr}(Y = 0)$ is different in the training and testing sample, the class-specific error control in eq. (10) still holds as long as the distribution P_j remains the same. Theorems 2 and 3 hold for $\hat{C}_j$ obtained by the sample splitting procedure when $|\mathcal{I}_1|$ and $|\mathcal{I}_2|$ are lower bounded by a constant fraction of n.

4·2. *Cross-validated ambiguity*

In our examples, the local polynomial estimator involves two tuning parameters: the polynomial degree and the bandwidth; and penalized logistic regression depends crucially on λ_n. These tuning parameters must be chosen in a data-driven manner. A natural idea is to choose a tuning parameter that minimizes the empirical ambiguity. To avoid overfitting, we can split the sample and use one part to find $\hat{\eta}$ and the other to choose the tuning parameter. Let $\Lambda = \{\lambda_1, ..., \lambda_T\}$ be the set of candidate tuning parameters. For each $\lambda \in \Lambda$, we can fit $\hat{\eta}_\lambda$ using the first subsample and compute $(\hat{C}_0, \hat{C}_1)$ and empirical ambiguity $\hat{\mathcal{A}}(\lambda)$ from the second subsample. Then we choose $\hat{\lambda}^* = \arg\min \hat{\mathcal{A}}(\lambda)$. This procedure can also be modified to a V-fold cross validation. As shown in Section 6, the sample splitting approach gives good empirical performance.

The output of the sample splitting procedure introduced here is an optimal $\hat{\lambda}^*$, which is then used to obtain $\hat{\eta} = \hat{\eta}_{\hat{\lambda}^*}$. Such a sample splitting for tuning parameter selection should not be confused with that introduced in Section 4·1. Theoretically speaking, in order to perform tuning parameter selection and achieve finite sample coverage, one needs to split the training sample into a fitting subsample and a ranking subsample as in Section 4·1, and further split the fitting subsample to obtain an estimate of $\hat{\eta}$ using a data driven tuning parameter. However, in practice splitting into two subsamples usually works well enough, where the ranking subsample is used for both selecting tuning parameters and estimating cut-off values.

Table 1. *Coverage (%) and ambiguity (%) under a Gaussian mixture model.*

	spar $= 1$			spar $= 0.5$		
	Noncover 0	Noncover 1	Ambiguity	Noncover 0	Noncover 1	Ambiguity
Ideal	5.00	5.00	31.1	5.00	5.00	31.1
Method 1	4.83 (0.10)	4.76 (0.20)	34.1 (0.65)	5.74 (0.12)	8.55 (0.21)	26.8 (0.46)
Method 2	4.55 (0.17)	4.19 (0.28)	39.1 (1.15)	4.44 (0.17)	4.43 (0.22)	46.9 (1.03)
Method 3	0.87 (0.04)	9.24 (0.29)	34.1 (0.68)	1.82 (0.07)	13.5 (0.26)	26.4 (0.46)

Percentage of coverage and ambiguity for different methods under a Gaussian mixture model: $(X \mid |Y = 0) \sim N(-1, 1)$, $(X \mid Y = 1) \sim N(1, 1)$, $\pi_0 = 0.75$, $\pi_1 = 0.25$. Method 1: classification with confidence; Method 2: classification with confidence and robust implementation; Method 3: classification with rejection. Reported are the average numbers in percentage over 100 independent samples, with estimated standard error in parenthesis. $\mathrm{spar} = 1$ corresponds to cross validation tuning, and $\mathrm{spar} = 0.5$ is chosen for illustration.

5. SIMULATION

We present a simulation study to illustrate the difference between the proposed framework and classification with rejection, and to demonstrate the finite sample performance of the robust implementation introduced in Section 4·1.

To highlight the main idea, we consider a simple Gaussian mixture model: $(X \mid Y = 0) \sim N(-1, 1)$, $(X \mid Y = 1) \sim N(1, 1)$, $\pi_0 = 3/4$, $\pi_1 = 1/4$. The target coverage is $\alpha_0 = \alpha_1 = 0.05$. The conditional probability function η is estimated using the R function smooth.spline. Three plug-in methods are compared: classification with confidence, the robust implementation described in Section 4, and classification with rejection; all are based on the same estimated $\hat{\eta}$. In our notation, an estimate given by the classification with rejection approach is $\hat{C}_0 = \{x : \hat{\eta}(x) \leq 0.5 + k\}$, $\hat{C}_1 = \{x : \hat{\eta}(x) \geq 0.5 - k\}$ for some $k \in [0, 0.5)$. In our simulation k is chosen such that $\hat{C}_0 \cap \hat{C}_1$ contains the same number of sample points as in the ambiguous region in the classification with confidence approach.

The R function smooth.spline requires a smoothness tuning parameter spar. We use two values of spar: 1 and 0.5, with a sample size of $n = 500$. In this case, $\mathrm{spar} = 1$ is a reasonably good choice of smoothing parameter. The median value of cross-validated spar over 100 simulations is 1.006. In the second estimate, we intentionally use a bad value of tuning parameter $\mathrm{spar} = 0.5$ to illustrate the finite sample coverage guarantee of the robust implementation. The simulation is repeated 100 times and the results are summarized in Table 1.

The simulation results reveal a fundamental difference between the proposed framework and classification with rejection. When the two methods are set to have ambiguous regions of similar size, the proposed method approximately achieves the target class-wise coverage levels, while classification with rejection tends to have imbalanced performance between classes. When the smoothness parameter is chosen inappropriately, the proposed method no longer has the asymptotic coverage guarantee. One can still use the robust implementation, which gives correct coverage on average, at the expense of having a larger ambiguous region.

6. DATA EXAMPLE

6·1. *The hand-written zip code data*

In this section we illustrate our method on the zip code data. The data consists of 16×16 pixel images of hand-written digits. Each image has a label in $\{0, 1, ..., 9\}$. This data set has been used to test pattern recognition (Le Cun et al., 1990) and classification algorithms (Hastie

Table 2. *Coverage (%) and ambiguity (%) in zip code data.*

	Noncover 0	Noncover 1	Total Noncover	Ambiguity	$\lambda_n \times 10^3$
ℓ_1-Logistic	1.4	11.4	3.3	0	1.84
Class-specific coverage	5.5	6.0	5.6	1.9	1.49
Overall coverage	0.9	10.08	2.8	3.3	1.49

Performance summary of classification with confidence, with comparison to the standard ℓ_1-penalized logistic regression. $\alpha_0 = \alpha_1 = 0.05$ for class-specific error control; $\alpha = 0.03$ for overall error control.

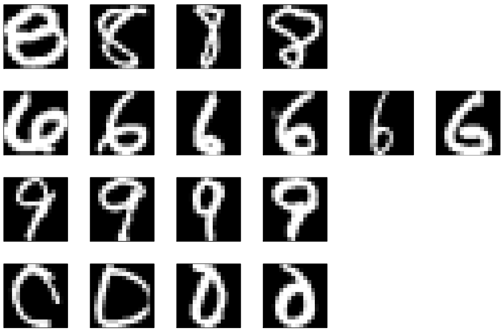

Fig. 1. Ambiguous images in the ranking subsample (18 out of 1014). Class-specific coverage $\alpha_0 = \alpha_1 = 0.05$.

et al., 2009). Examples of the images can be found in Le Cun et al. (1990). We choose this data example because the ambiguous cases are easy to interpret visually.

In order to make an imbalanced binary classification problem, we use the subset of data containing digits $\{0, 6, 8, 9\}$, which are digits with circles. Images corresponding to digits 0, 6, and 9 are labeled as class 0, and those of digit 8 are labeled as class 1. The training sample contains 3044 images, with 542 in class 1. The test sample contains 872 images, with 166 in class 1.

6·2. *Class-specific coverage*

We apply ℓ_1-penalized logistic regression on the training data set with $\alpha_0 = \alpha_1 = 0.05$. The robust implementation is carried out by using two-thirds of training data as the fitting subsample, and the remainder as the ranking subsample. The tuning parameter λ_n is selected by minimizing the empirical ambiguity in the ranking subsample. The resulting set-valued classifier $\hat{h}$ is then validated on the testing data.

Table 2 summarizes the performance of our method and compares it with the ordinary binary classifier given by the ℓ_1-penalized logistic regression using standard 5-fold cross-validation. We look at four measurements: noncoverage for class 0; noncoverage for class 1; total noncoverage; ambiguity.

Although ordinary binary classification gives a small overall mis-classification rate with no ambiguity, the mis-classfication rate for class 1 is much higher. Our procedure gives noncoverage rates near the nominal levels for both classes, with a small ambiguity. Figure 1 displays the 18 ambiguous images in the ranking subsample. Many of these images exhibit some pattern of abnormal hand-writing. They are hard to classify by algorithm but can be classified visually.

Figure 2 provides some further insights into the performance of our procedure, supporting the theory developed earlier. The empirical coverage is close to the nominal level for almost the entire range of λ_n. This agrees with Proposition 3. In the right panel, the empirical ambi-

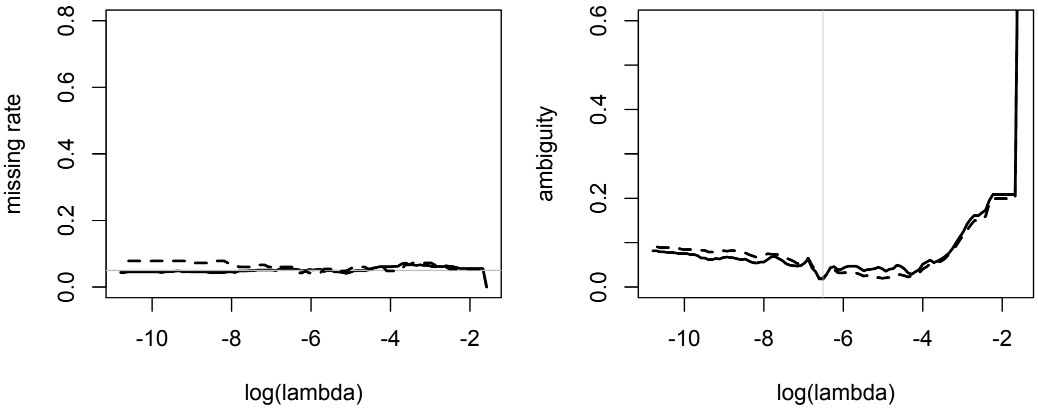

Fig. 2. Effects of tuning parameter on non-coverage rates and ambiguity using class-specific coverage with $\alpha_0 = \alpha_1 = 0.05$. Left panel: non-coverage rate as a function of $\log(\lambda)$ for class 0 (solid) and class 1 (dash). The grey line marks the nominal noncoverage rate of 0.05. Right panel: empirical ambiguity as a function of $\log(\lambda)$ for training sample (solid) and test sample (dash). The grey line indicates the value of $\log(\lambda)$ that minimizes empirical ambiguity in the training sample.

guity curves calculated from the test and training data are very close, because in both cases the empirical ambiguity is evaluated using a random sample independent of $\hat{\eta}$.

6·3. *Overall coverage*

In controlling overall coverage as discussed in Sections 2·2 and 3·2, we choose $\alpha = 0.03$. The method described in Section 3·2 is applied to find $\hat{C}_0$ and $\hat{C}_1$ by searching for the best $\hat{\alpha}_0^*$ that minimizes the empirical ambiguity in the ranking subsample. The tuning parameter is selected by minimizing estimated ambiguity with the same data split as described in the previous subsection. The result is summarized in the bottom row of Table 2, where the overall error rate is bounded by the nominal level.

Figure 3 shows the test data noncoverage rate and ambiguity as a function of λ_n. The tuning parameter has little effect on noncoverage rates, as claimed in Proposition 3, but does affect the ambiguity. The ambiguity is systematically higher in the test sample than in the training sample, because we choose $\hat{\alpha}_0^*$ by minimizing empirical ambiguity in the ranking subsample, which is then plugged in for the test sample without searching for a new $\hat{\alpha}_0^*$. Nevertheless, minimizing empirical ambiguity still leads to a reasonable choice of λ_n. In the right panel of Figure 3, the plotted curve does reflect a convex shape with a unique minimum.

7. DISCUSSION

The confidence level of classification is quite similar to the significance level in statistical hypothesis testing. Such a connection between classification and testing has appeared in the literature. The proof of Lemma 2 of this paper uses ideas in Sun & Cai (2007, 2009), where the multiple testing problem is formulated via a classification problem.

Although the optimal solution depends on ranking and thresholding $\eta(x)$, our framework is applicable in more general settings. For example, many popular classification methods, such as

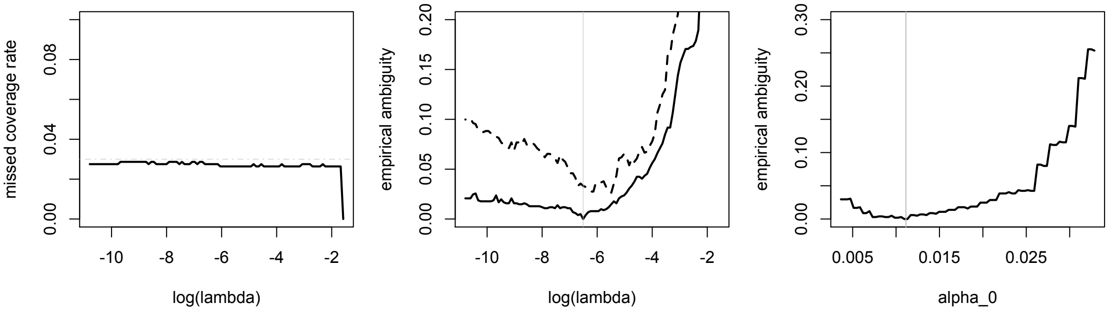

Fig. 3. Total coverage control for zip code data with $\alpha = 0.03$. Left panel: total error as a function of $\log(\lambda)$. The grey line marks the nominal noncoverage rate of 0.03. Middle panel: empirical ambiguity as a function of $\log(\lambda)$ for training data (solid) and test data (dash). The grey line marks the value of $\log(\lambda)$ that minimizes the empirical ambiguity in the training sample. Right panel: empirical ambiguity as a function of α_0 with $\lambda_n = 1.49 \times 10^{-3}$. The grey line marks the value of α_0 that minimizes the empirical ambiguity.

empirical risk minimization, are based on thresholding a function $\hat{f}(x)$. Most of the arguments developed in this paper are still applicable when f, the population version of $\hat{f}$, is roughly a monotone function of η in a neighborhood of the cut-off points.

There are interesting extensions in both theory and methodology. First, it would be interesting to ask what would happen to the ambiguity if a good approximation to the ideal procedure is not available. For example, naive Bayes classifiers have been shown to give competitive performance (Bickel & Levina, 2004) but are also known to be biased except under a naive Bayes model. The same question could also be asked for classifiers with other ranking scores, such as data depth (Li et al., 2012).

Another extension is the multi-class case. In a k-class problem, the optimal procedure shall be based on the vector $\eta(x) = \{\eta_\ell(x) : 1 \leq \ell \leq k\}$ with $\eta_\ell(x) = \mathrm{pr}(Y = \ell \mid X = x)$. One can define the ambiguity of a set-valued classifier in different ways. For example, both $\mathcal{A}(h) = \mathrm{pr}\{|h(X)| \geq 2\}$ (Vovk et al., 2005, Section 3) and $\mathcal{A}(h) = E\{|h(X)|\}$ agree with the definition used in this work when $k = 2$. It is an open problem to find the corresponding ideal classifiers and consistent estimates. Moreover, the option of $h(x) = \emptyset$ for certain values of x is interesting for both practical and theoretical concerns. An example of possible applications is clinical decision-making such as dynamic treatment regimes (Laber et al., 2014).

ACKNOWLEDGEMENT

This research was partially supported by the U.S. National Science Foundation and the U.S. National Institutes of Health. The author is grateful to the editor, associate editor, and three reviewers for insightful comments that helped improving this paper. The author thanks Professor Larry Wasserman for many inspiring discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material available online includes a proof of Proposition 1.

TECHNICAL DETAILS

A·1. *Proofs of Section* 2

Proof. (Theorem 1) Let C_0, C_1 be the classification regions given in Lemma 1. Let S_0, S_1 be the corresponding regions of any other classifier such that $P_j(S_j) \geq 1 - \alpha_j$, for $j = 0, 1$. Without loss of generality, we assume $t_0 \geq t_1$ since otherwise $C_0 \cap C_1 = \emptyset$ and the claim is true. We also assume that $S_0 \cup S_1 = \mathcal{X}$ because otherwise we can always substitute S_1 by $S_1' = S_1 \cup [\mathcal{X} \backslash (S_0 \cup S_1)]$ without affecting the result and argument. Let $S_{01} = S_0 \cap S_1$. Using the fact that $\alpha_0 = P_0(C_0) \leq P_0(S_0)$ we have

$$P_0(C_{01} \cap S_0^c) - P_0(C_0^c \cap S_{01}) \leq P_0(C_0^c \cap S_1^c) - P_0(C_1^c \cap S_0^c). \tag{A1}$$

Canceling $P_0(C_{01} \cap S_{01})$ in $P_0(C_{01})$ and $P_0(S_{01})$, we have

$$\begin{aligned} P_0(C_{01}) - P_0(S_{01}) &= P_0(C_{01} \cap S_1^c) + P_0(C_{01} \cap S_0^c) - P_0(C_0^c \cap S_{01}) - P_0(C_1^c \cap S_{01}) \\ &\leq P_0(C_{01} \cap S_1^c) + P_0(C_0^c \cap S_1^c) - P_0(C_1^c \cap S_0^c) - P_0(C_1^c \cap S_{01}) \\ &= P_0(C_{01} \cap S_1^c) + P_0(C_0^c \cap S_1^c) - [P_0(C_1^c \cap S_0^c) + P_0(C_1^c \cap S_{01})] \\ &= P_0(C_1 \cap S_1^c) - P_0(C_1^c \cap S_1) \\ &\leq \frac{1 - t_1}{t_1} P_1(C_1 \cap S_1^c) - \frac{1 - t_1}{t_1} P_1(C_1^c \cap S_1) \leq 0. \end{aligned}$$

In the above derivation, the first inequality follows from (A1). The second uses the fact that $C_1 \cap S_1^c \subseteq C_1$, $C_1^c \cap S_1 \subseteq C_1^c$ and the definition of C_1, as the likelihood ratio dP_0/dP_1 is greater than $(1 - t_1)/t_1$ on C_1^c and smaller than $(1 - t_1)/t_1$ on C_1. The last holds because $P_1(S_1) \geq 1 - \alpha_1 = P_1(C_1)$.

The proof is completed by realizing that $P_1(C_{01}) \leq P_1(S_{01})$ can be derived using the same argument. □

Proof. (Lemma 1) If (C_0^*, C_1^*) does not minimize problem (2) with (α_0, α_1) specified by (α_0^*, α_1^*), let (C_0', C_1') be the minimizer so that $P(C_{01}') < P(C_{01})$. On the other hand, (C_0', C_1') is feasible for (2) and hence it is also feasible for (3). This contradicts the fact that (C_0^*, C_1^*) is optimal for (3).

For the second claim, consider a different optimization problem

$$\begin{aligned} &\arg\min \ \nu(\alpha_0, \alpha_1) \\ &\text{subject to } 0 \leq \alpha_0, \alpha_1 \leq 1, \ \ \pi_0\alpha_0 + \pi_1\alpha_1 \leq \alpha. \end{aligned} \tag{4'}$$

It follows immediately from the first part that the optimal value of (4’) equals the optimal value of (3). The two problems (4) and (4’) are equivalent because $\nu(\alpha_0, \alpha_1)$ is decreasing in both α_0 and α_1 and hence the minimum of (4’) is achieved when $\pi_0\alpha_0 + \pi_1\alpha_1 = \alpha$. □

Proof. (Lemma 2) Let $G_j(\cdot)$ be the cumulative distribution function of $\eta(X)$ when $X \sim P_j$ $(j = 0, 1)$, and $g_j(\cdot)$ be the corresponding density function. For a given $\alpha_0 \in (0, \alpha/\pi_0)$, let $\alpha_1 = (\alpha - \pi_0\alpha_0)/\pi_1$ according to the constraint in problem (4). Then the corresponding t_0 and t_1 in Theorem 1 are $t_0 = G_0^{-1}(1 - \alpha_0)$ and $t_1 = G_1^{-1}(\alpha_1)$. Let $C_0 = \{x : \eta(x) \geq t_0\}$ and $C_1 = \{x : \eta(x) \leq t_1\}$. Define

$$\begin{aligned} \mu(\alpha_0) =& \pi_0\{G_0(t_0) - G_0(t_1)\} + \pi_1\{G_1(t_0) - G_1(t_1)\} \\ =& - \pi_0 G_0(t_1) + \pi_1 G_1(t_0) + \pi_0 - \alpha\,. \end{aligned}$$

Differentiating over α_0, we have

$$\mu'(\alpha_0) = - \pi_0 \frac{dG_0(t_1)}{d\alpha_0} + \pi_1 \frac{dG_1(t_0)}{d\alpha_0} = -\pi_0 g_0(t_1) \frac{dt_1}{d\alpha_0} + \pi_1 g_1(t_0) \frac{dt_0}{d\alpha_0} = \frac{\pi_0^2}{\pi_1} \frac{g_0(t_1)}{g_1(t_1)} - \pi_1 \frac{g_1(t_0)}{g_0(t_0)}\,.$$

According to Corollary 1 of Sun & Cai (2009),

$$\frac{g_1(t)}{g_0(t)} = \frac{t}{1 - t} \frac{\pi_0}{\pi_1}\,, \ \ \mu'(\alpha_0) = \pi_0 \frac{1 - t_1}{t_1} - \pi_0 \frac{t_0}{1 - t_0}, \ \ \mu''(\alpha_0) = \frac{\pi_0^2}{\pi_1} \frac{1}{t_1^2 g_1(t_1)} + \pi_0 \frac{1}{(1 - t_0)^2 g_0(t_0)},$$

so $\mu(\alpha_0)$ is convex on the interval $[0, \alpha/\pi_0]$. The claimed result follows by realizing that $\nu(\alpha_0) = \max\{\mu(\alpha_0), 0\}$ is also convex. □

A·2. *Proofs for Section* 3

Let $\hat{G}_j$ be the empirical cumulative distribution function of $\eta(X_{j,1}),...,\eta(X_{j,n_j})$. Throughout this subsection we will focus on the event

$$E_r = \left\{ \|\hat{\eta}-\eta\|_\infty \le \delta_n,\ |\hat{\pi}_j - \pi_j| \le c(\log n_j/n_j)^{1/2},\ \sup_t |G_j(t) - \hat{G}_j(t)| \le c(\log n_j/n_j)^{1/2} \right\},$$

which has probability at least $1-\rho_n - n^{-r}$ for some constant c depending on r only. To see this, using Hoeffding's inequality, $\mathrm{pr}\{|\hat{\pi}_j - \pi_j| \ge c(\log n/n)^{1/2}\} \le n^{-r}/2$. Then using the result of Massart (1990) we have $\mathrm{pr}\{\sup_t |G_j(t) - \hat{G}_j(t)| \ge c(\log n_j/n_j)^{1/2}\} \le n^{-r}/2$. In both cases c depends on r only. The value of constant c, as well as c_1, c_2 below, may vary from one line to another.

The following lemma shows that $\hat{t}_j$ is a good approximation to t_j.

LEMMA A1. *On E_r, the $\hat{t}_j(\alpha_j)$ $(j=0,1)$ used in eq.* (5) *satisfy, for n large enough,*

$$|\hat{t}_j(\alpha_j) - t_j| \le c\left\{\delta_n + \left(\frac{\log n}{n}\right)^{\frac{1}{2\gamma}}\right\}.$$

Proof. We prove the result for $j=1$. The argument for $j=0$ is the same.

Let $\hat{P}_1$ be the empirical probability distribution that assigns probability $1/n_1$ at each $X_{1,i}$ for $i = 1,...,n_1$. Denote $L^\ell(t) = \{x \in \mathcal{X} : \eta(x) \le t\}$, and $\hat{L}^\ell(t) = \{x \in \mathcal{X} : \hat{\eta}(x) \le t\}$. Then

$$\hat{P}_1\{\hat{L}^\ell(t)\} \le \hat{P}_1\{L^\ell(t+\delta_n)\} = \hat{G}_1(t+\delta_n) \le G_1(t+\delta_n) + c\left(\frac{\log n_1}{n_1}\right)^{1/2},\ \ t\in[0,1].$$

Let $t_1' = t_1 - \delta_n - \{2cb_1^{-1}(\log n_1/n_1)^{1/2}\}^{1/\gamma}$, where b_1 is the constant in (MA). For n and n_1 large enough we have $n_1^{-1} < c(\log n_1/n_1)^{1/2}$, and $\delta_n + \{2cb_1^{-1}(\log n_1/n_1)^{1/2}\}^{1/\gamma} \le \epsilon_0$, with ϵ_0 defined in (MA). Then

$$\begin{aligned}\hat{P}_1\{\hat{L}^\ell(t_1')\} \le & G_1\left[t_1 - \{2cb_1^{-1}(\log n_1/n_1)^{1/2}\}^{1/\gamma}\right] + c\left(\frac{\log n_1}{n_1}\right)^{1/2} \le G_1(t_1) - c(\log n_1/n_1)^{1/2} \\ = & \alpha_1 - c(\log n_1/n_1)^{1/2} < \alpha_1 - n_1^{-1} \le \lfloor n_1\alpha_1 \rfloor n_1^{-1} \le \hat{P}_1\{\hat{L}^\ell(\hat{t}_1)\}.\end{aligned}$$

where the second last step uses (MA). Therefore,

$$\hat{t}_1 \ge t_1 - \delta_n - \{2cb_1^{-1}(\log n_1/n_1)^{1/2}\}^{1/\gamma}.$$

Similarly, one can show that $\hat{t}_1 \le t_1 + \delta_n + \{2cb_1^{-1}(\log n_1/n_1)^{1/2}\}^{1/\gamma}$. The claimed result follows by picking a larger constant c so that n_1 can be replaced by n. □

Proof. (Theorem 2) We give the proof for $j=1$. Under assumptions in the theorem, and according to Lemma 1, we have, on event E_r,

$$\begin{aligned}P_1(\hat{C}_1\backslash C_1) = & P_1\{\hat{\eta}(X) \ge \hat{t}_1, \eta(X) < t_1\} \le P_1\left\{t_1 - 2\delta_n - c\left(\frac{\log n}{n}\right)^{1/2\gamma} \le \eta(X) < t\right\} \\ \le & b_2\left\{2\delta_n + c\left(\frac{\log n}{n}\right)^{1/2\gamma}\right\}^\gamma \le 2^\gamma b_2\left\{2^\gamma\delta_n^\gamma + c^\gamma\left(\frac{\log n}{n}\right)^{1/2}\right\},\end{aligned}$$

where the last inequality comes from condition (MA) and holds when n is large enough so that $2\delta_n + c(\log n/n)^{1/2\gamma} \le \epsilon_0$. The other parts can be argued similarly. □

Proof. (Theorem 3) Recall that on E_r, we have, for some c_1, c_2,

$$|\hat{t}_0 - t_0| \le c_1\{\delta_n + (\log n/n)^{1/2}\},\ \ \alpha_0 \in [\underline{\alpha}, \overline{\alpha}], \tag{A2}$$

and $|\tilde\pi_0 - \pi_0| \le c_2(\log n/n)^{1/2}$. Then for n large enough, Lemma 1 implies that

$$\begin{aligned}|\hat t_1(\hat\alpha_1) - t_1| \le& |\hat t_1(\hat\alpha_1) - t_1(\hat\alpha_1)| + |G_1^{-1}(\alpha_1) - G_1^{-1}(\hat\alpha_1)|\\ \le& c_1\{\delta_n + (\log n/n)^{1/2}\} + c_2'(\log n/n)^{1/2}\,,\ \ \alpha_0 \in [\underline{\alpha}, \overline{\alpha}]\end{aligned} \tag{A3}$$

where the first part follows from that (MA) holds with $\gamma = 1$, and the second part follows from that G_1^{-1} is Lipschitz as implied by (MA).

Let P_n be the empirical marginal distribution of X. Then for n large enough we have, on E_r,

$$\begin{aligned}\hat\nu(\alpha_0) - \nu(\alpha_0) =& P_n\left\{\hat t_1(\hat\alpha_1) \le \hat\eta(X) \le \hat t_0(\alpha_0)\right\} - P\left\{t_1 \le \eta(X) \le t_0\right\}\\ \le& P_n\left\{\hat t_1(\hat\alpha_1) - \delta_n \le \eta(X) \le \hat t_0(\alpha_0) + \delta_n\right\} - P\left\{t_1 \le \eta(X) \le t_0\right\}\\ =& P_n\left\{\hat t_1(\hat\alpha_1) - \delta_n \le \eta(X) \le \hat t_0(\alpha_0) + \delta_n\right\} - P\left\{\hat t_1(\hat\alpha_1) - \delta_n \le \eta(X) \le \hat t_0(\alpha_0) + \delta_n\right\}\\ &+ P\left\{\hat t_1(\hat\alpha_1) - \delta_n \le \eta(X) \le \hat t_0(\alpha_0) + \delta_n\right\} - P\left\{t_1 \le \eta(X) \le t_0\right\}\\ \le& c\{\delta_n + (\log n/n)^{1/2}\},\end{aligned}$$

where in the last inequality, the first part is bounded, up to a constant factor, by $(\log n/n)^{1/2}$ using standard empirical process theory, and the second part is bounded by $\delta_n + (\log n/n)^{1/2}$ using (A2), (A3), and the assumption that G_j has bounded density. The other direction $-\hat\nu(\alpha_0) + \nu(\alpha_0)$ can be treated similarly. The first claim follows from strong convexity and uniform approximation of $\hat\nu$ to ν.

Next we prove the second part for $\hat C_0^*$. The proof for $\hat C_1^*$ is similar. On the event in the first claim, $\hat\alpha_0^* \in [\underline{\alpha}, \bar\alpha]$. By the assumptions, the claim of Lemma 1 holds uniformly for all $\alpha_0 \in [\underline{\alpha}, \bar\alpha]$. Thus

$$\begin{aligned}P_0(\hat C_0^* \Delta C_0^*) \le& P_0\{\hat C_0^* \Delta C_0(\hat\alpha_0^*)\} + P_0\{C_0(\hat\alpha_0^*)\Delta C_0^*\}\\ \le& c\{\delta_n + (\log n/n)^{1/2}\} + |\hat\alpha_0 - \alpha_0^*| \le c'\{\delta_n^{1/2} + (\log n/n)^{1/4}\}\,.\end{aligned}$$

REFERENCES

AUDIBERT, J.-Y. & TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers. *The Annals of Statistics* **35**, 608–633.

BICKEL, P. J. & LEVINA, E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10**, 989–1010.

CADRE, B., PELLETIER, B. & PUDLO, P. (2013). Estimation of density level sets with a given probability content. *Journal of Nonparametric Statistics* **25**, 261–272.

CHOW, C. K. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory* **16**, 41–46.

HAN, M., CHEN, D. & SUN, Z. (2008). Analysis to Neyman–Pearson classification with convex loss function. *Analysis in Theory and Applications* **24**, 18–28.

HANCZAR, B. & DOUGHERTY, E. R. (2008). Classification with reject option in gene expression data. *Bioinformatics* **24**, 1889–1895.

HASTIE, T. J., TIBSHIRANI, R. J. & FRIEDMAN, J. H. (2009). *The Elements of Statistical Learning*. New York: Springer-Verlag, 2nd ed.

HERBEI, R. & WEGKAMP, M. H. (2006). Classification with rejection option. *The Canadian Journal of Statistics* **34**, 709–721.

LABER, E. B., LIZOTTE, D. J. & FERGUSON, B. (2014). Set-valued dynamic treatment regimes for competing outcomes. *Biometrics* **70**, 53–61.

LE CUN, Y., BOSER, B., DENKER, J. S., HENDERSON, D., HOWARD, R. E., HUBBARD, W. & JACKEL, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems 2. Proceedings of the 1989 Conference*, D. S. Touretzky, ed. San Francisco, CA, USA: Morgan Kaufmann.

LEI, J., RINALDO, A. & WASSERMAN, L. (2014). A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence* , in press.

LEI, J., ROBINS, J. & WASSERMAN, L. (2013). Distribution free prediction set. *Journal of the American Statistical Association* **108**, 278–287.

LEI, J. & WASSERMAN, L. (2014). Distribution free prediction bands for nonparametric regression. *Journal of the Royal Statistical Society, Series B* **76**, 71–96.

LI, J., CUESTAS-ALBERTOS, J. A. & LIU, R. Y. (2012). DD-classifier: Nonparametric classification procedure based on dd-plot. *Journal of the American Statistical Association* **107**, 737–753.

MASSART, P. (1990). The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability* , 1269–1283.

NADEEM, M. S. A., ZUCKER, J.-D. & HANCZAR, B. (2010). Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. *Journal of Machine Learning Research-Proceedings Track* **8**, 65–81.

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

RIGOLLET, P. & TONG, X. (2011). Neyman–Pearson classification under a strict constraint. *Journal of Machine Learning Research-Proceedings Track* **19**, 595–614.

RIGOLLET, P. & VERT, R. (2009). Optimal rates for plug-in estimators of density level sets. *Bernoulli* **14**, 1154–1178.

SCOTT, C. & NOWAK, R. (2005). A Neyman–Pearson approach to statistical learning. *IEEE Transactions on Information Theory* **51**, 3806–3819.

SHAFER, G. & VOVK, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research* **9**, 371–421.

SUN, W. & CAI, T. (2007). Oracle and adaptive compund decision rules for false discovery rate control. *Journal of the American Statistical Association* **102**, 901–912.

SUN, W. & CAI, T. T. (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**, 393–424.

TONG, X. (2013). A plug-in approach to Neyman–Pearson classification. *Journal of Machine Learning Research* **14**, 3011–3040.

TSYBAKOV, A. B. (1997). On nonparametric estimation of density level sets. *The Annals of Statistics* **25**, 948–969.

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. New York: Springer.

VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics* **36**, 614–645.

VOVK, V. (2013). Conditional validity of inductive conformal predictors. *Machine Learning* **92**, 349–376.

VOVK, V., GAMMERMAN, A. & SHAFER, G. (2005). *Algorithmic Learning in a Random World*. New York: Springer.

VOVK, V., NOURETDINOV, I. & GAMMERMAN, A. (2009). On-line predictive linear regression. *The Annals of Statistics* **37**, 1566–1590.

YUAN, M. & WEGKAMP, M. (2010). Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research* **11**, 111–130.

Biometrika (2014), **xx**, x, *pp.* 1–4
© 2007 Biometrika Trust
Printed in Great Britain

Supplementary material for “Classification with Confidence”

BY JING LEI

Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania, 15215, U.S.A

jinglei@andrew.cmu.edu

1. BACKGROUND OF LOCAL POLYNOMIAL ESTIMATOR

Assume $\mathcal{X} = [-1, 1]^d$. For any $s = (s_1, ..., s_d) \in (\mathbb{Z}^+)^d$, where $\mathbb{Z}^+$ is the set of nonnegative integers, and $z \in \mathbb{R}^d$, define $|s| = \sum_{1\le k\le d} s_k$ and $z^d = z_1^{s_1} \ldots z_d^{s_d}$. Let K be a kernel function and $\tau > 0$ a bandwidth. The local polynomial estimator for $\eta(x)$ is

$$\hat\eta(x) = \hat v_0(x), \tag{1}$$

where

$$\hat v(x) = \{\hat v_s(x)\}_{s:|s|\le\ell} = \arg\min_v \sum_{i=1}^n \left\{ Y_i - \sum_{s:|s|\le\ell} v_s \left(\frac{X_i - x}{\tau}\right)^s \right\}^2 K\left(\frac{X_i - x}{\tau}\right).$$

To prove consistency of $\hat\eta(x)$, we assume that $\eta(x) = \mathrm{pr}(Y = 1 \mid X = x)$ belongs to a Hölder class $\Sigma(\beta, L)$, and that K is a valid kernel function. The following definitions are commonly used in nonparametric statistics (Tsybakov, 2009).

For any sufficiently smooth function $f : \mathbb{R}^d \mapsto \mathbb{R}$, define D^s to be the differential operator:

$$D^s f = \frac{\partial^{|s|} f}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}}(x_1, ..., x_d).$$

Given $\beta > 0$, for any function f that is $\lfloor\beta\rfloor$ times differentiable, denote its Taylor expansion of degree $\lfloor\beta\rfloor$ at x_0 by

$$f_{x_0}^{(\beta)}(x) = \sum_{|s|\le\beta} \frac{(x - x_0)^s}{s_1! \cdots s_d!} D^s f(x_0).$$

DEFINITION 1. *For constants $\beta > 0$, $L > 0$, define the Hölder class $\Sigma(\beta, L)$ to be the set of $\lfloor\beta\rfloor$-times differentiable functions on $\mathbb{R}^d$ such that,*

$$|f(x) - f_{x_0}^{(\beta)}(x)| \le L\|x - x_0\|^\beta. \tag{2}$$

A standard condition on the kernel is the notion of β-valid kernels.

DEFINITION 2. *For any $\beta > 0$, function $K : \mathbb{R}^d \mapsto \mathbb{R}^1$ is a β-valid kernel if (a) K is supported on $[-1, 1]^d$; (b) $\int K = 1$; (c) $\int |K|^r < \infty$, all $r \ge 1$; (d) $\int y^s K(y)dy = 0$ for all $1 \le |s| \le \beta$.*

Remark 1. The assumption that $\mathcal{X} = [0, 1]^d$ can be generalized to the (r_0, c_0)-regularity condition used in Audibert & Tsybakov (2007).

2. SUP-NORM ERROR BOUND FOR LOCAL POLYNOMIAL ESTIMATOR

THEOREM 1. *Assume that $\eta \in \Sigma(\beta, L)$, and that K is a valid kernel function of order β. Let $\hat\eta$ be the order $\lfloor\beta\rfloor$ local polynomial estimator using kernel function K and bandwidth τ. If the marginal density function p_X of X is uniformly bounded and bounded away from zero on $\mathcal{X} = [-1,1]^d$, then for any $r > 0$, there exists a constant c such that*

$$\mathrm{pr}\left[\sup_x |\hat\eta(x) - \eta(x)| \le c\left\{\left(\frac{\log n}{n\tau^d}\right)^{1/2} + \tau^\beta\right\}\right] \ge 1 - n^{-r}.$$

Proof. Let $U(z) = (z^s)_{s:|s|\le\ell}$ for all $z \in \mathbb{R}^d$, where $\ell = \lfloor\beta\rfloor$. Let M be the length of $U(z)$. The local polynomial estimator can be written as

$$\hat\eta(x) = U^T(0)H_{nx}^{-1}A_{nx}Y,$$

where

$$\begin{aligned} H_{nx}(s,s') =& \frac{1}{n\tau^d}\sum_{i=1}^n \left(\frac{X_i - x}{\tau}\right)^{s+s'} K\left(\frac{X_i - x}{\tau}\right), \\ A_{nx}(s,i) =& \frac{1}{n\tau^d}\left(\frac{X_i - x}{\tau}\right)^{s} K\left(\frac{X_i - x}{\tau}\right), \\ Y =& (Y_1, ..., Y_n)^T. \end{aligned}$$

By the reproducing property of local polynomial estimators, if $Q_n = \{q(X_1), ..., q(X_n)\}^T$ with q an order ℓ polynomial, then we have

$$U^T(0)H_{nx}^{-1}A_{nx}Q_n = q(x).$$

Then we have the following bias-variance decomposition of the estimation error:

$$\begin{aligned} \hat\eta(x) - \eta(x) &= U^T(0)H_{nx}^{-1}A_{nx}\{Y - \eta(x)e_n\} \\ &= U^T(0)H_{nx}^{-1}A_{nx}\{Y - \eta_n + \eta_n - \eta_{nx} + \eta_{nx} - \eta(x)e_n\} \\ &= U^T(0)H_{nx}^{-1}A_{nx}(\xi_n + R_{nx}), \end{aligned}$$

where e_n is a $n \times 1$ column vector of 1's, and η_n, η_{nx}, ξ_n, R_{nx} are vectors such that

$$\eta_n(i) = \eta(X_i),\ \ \eta_{nx}(i) = \eta_x(X_i),\ \ \xi_n = Y - \eta_n,\ \ R_{nx} = \eta_n - \eta_{nx},$$

and $\eta_x(\cdot)$ is the local polynomial approximation for $\eta(\cdot)$ at x. By the Hölder assumption, $|R_{nx}(i)| \le L\tau^\beta$ whenever $K\{(X_i - x)/\tau\} \ne 0$. In the last equation we used the reproducing property of the local polynomial estimator. The remainder of the proof consists of three main steps.

In the first step, we provide a uniform lower bound on the smallest eigenvalue of H_{nx}. Let ϕ_H be the smallest eigenvalue of matrix H. Then using Equation 6.3 in Audibert & Tsybakov (2007), there exist constants $c > 0$ and $\mu_0 > 0$, such that

$$\mathrm{pr}(\phi_{H_{nx}} \le 2\mu_0) \le 2M^2\exp(-cn\tau^d)$$

for all x, and all n large enough, $n\tau^d$ large enough.

Using continuity and differentiation, we can find some constant κ depending on β and K only, such that for all s, s' satisfying $|s|, |s'| \leq \ell$ and all X,

$$\left| \left(\frac{X-x}{\tau}\right)^{s+s'} K\left(\frac{X-x}{\tau}\right) - \left(\frac{X-x'}{\tau}\right)^{s+s'} K\left(\frac{X-x'}{\tau}\right) \right| \leq \kappa\tau^{-1}\|x-x'\|_\infty, \tag{3}$$

which implies that

$$\|H_{nx} - H_{nx'}\|_2 \leq \|H_{nx} - H_{nx'}\|_F \leq \kappa M \tau^{-(d+1)}\|x-x'\|_\infty. \tag{4}$$

Given $\delta > 0$, there exists a finite subset $\{x_1, ..., x_{N_\delta}\} \subset [0,1]^d$ that covers $[-1,1]^d$ with radius δ in ℓ_∞ norm with $N_\delta \leq 2^d\delta^{-d}$. Now let $\delta = \mu_0\tau^{d+1}/\kappa M$ where κ is the constant in (3) and (4). Then for any $x \in [0,1]^d$, there exists a j such that $\|H_{nx} - H_{nx_j}\|_2 \leq \mu_0$ and hence for n large enough

$$\begin{aligned} \mathrm{pr}\left(\inf_x \lambda_{H_{nx}} \leq \mu_0\right) &\leq \mathrm{pr}\left(\text{there exists } 1 \leq j \leq N_\delta : \lambda_{H_{nx_j}} \leq 2\mu_0\right) \\ &\leq (\delta/2)^{-d}\exp(-cn\tau^d) \\ &\leq c'\exp\left\{-cn\tau^d - d(d+1)\log\tau\right\} \leq c'\exp(-cn\tau^d/2). \end{aligned} \tag{5}$$

The second step is to control $\|A_{nx}\xi_n\|_2$. Let

$$a_{nx}(s) = \frac{1}{n\tau^d}\sum_{i=1}^n \left(\frac{X_i - x}{\tau}\right)^s K\left(\frac{X_i-x}{\tau}\right)\xi_n(i).$$

Recall that $\{X_1, \xi_n(1)\}, ..., \{X_n, \xi_n(n)\}$ is an independent and identically distributed sample. Let $f_{s,x}(X,\xi) = \tau\left\{(X-x)/\tau\right\}^s K\{(X-x)/\tau\}\xi$. Then $|f_{s,x}(X,\xi) - f_{s,x'}(X,\xi)| \leq \kappa\|x - x'\|_\infty$, by smoothness of K, where κ depends only on β and K. As a result, for any $\delta > 0$, there exists a δ-covering set, with cardinality at most $(2\kappa/\delta)^d$, for the function class $\mathcal{F} = \{f_{x,s} : x \in [0,1]^d\}$ in $L_2(P)$ norm for any probability measure P on $[-1,1]^d \times [-1,1]$. We also have $|f_{s,x}| \leq \tau$, $E(f_{s,x}) = 0$, and $\mathrm{var}(f_{s,x}) \leq c\tau^{d+2}$ for some constant c depending on the distribution of X only.

The function class $\mathcal{F}$ indexed by x satisfies the conditions of Theorem 2.1 and Corollary 2.2 in Giné & Guillou (2002), which give sharp tail probability bounds for the sup-norm of empirical processes. These results are consequences of powerful exponential concentration inequalities due to Talagrand (1994, 1996).

As a result, there exist constants c_1, c_2, c_3 depending only on κ, d, such that for all $c \geq c_1$, we have

$$\mathrm{pr}\left\{\sup_x \left|\frac{1}{n\tau^d}\sum_{i=1}^n \left(\frac{X_i-x}{\tau}\right)^s K\left(\frac{X_i-x}{\tau}\right)\xi_n(i)\right| \geq c\left(\frac{\log n}{n\tau^d}\right)^{1/2}\right\} \leq c_2 n^{-c_3 c}. \tag{6}$$

Thus with probability at least $1 - c_2Mn^{-c_3c}$, we have

$$\|A_{nx}\xi_n\|_2 \leq c\left(\frac{M\log n}{n\tau^d}\right)^{1/2}. \tag{7}$$

The third step is to control $\|A_{nx}R_{nx}\|_2$. Consider the sth coordinate of $A_{nx}R_{nx}$

$$|(A_{nx}R_{nx})(s)| = \left|\frac{1}{n\tau^d}\sum_{i=1}^n \left(\frac{X_i - x}{\tau}\right)^s K\left(\frac{X_i - x}{\tau}\right) R_{nx}(i)\right|$$
$$\leq L\|K\|_\infty \tau^\beta \frac{1}{n\tau^d}\sum_{i=1}^n \mathbb{I}(|X_i - x| < \tau),$$

where $\mathbb{I}(\cdot)$ is the indicator function. Because the density of X is bounded by $\bar{\mu} < \infty$, we have

$$\sup_x E\{\mathbb{I}(|X - x| \leq \tau)\} \leq (2\tau)^d \bar{\mu}\,.$$

On the other hand, the function class $\mathcal{F}' = \{\mathbb{I}(|\cdot - x| < \tau) : x \in [0,1]^d\}$ also satisfies the sup-norm tail bound conditions mentioned above. Therefore, there exist constants c_1, c_2, and c_3 such that when $(n\tau^d/\log n)^{1/2} \geq c_1$,

$$\mathrm{pr}\left\{\sup_x \frac{1}{n\tau^d}\sum_{i=1}^n \mathbb{I}(|X_i - x| < \tau) \geq (\bar{\mu} + 1)\right\} \leq c_2 n^{-c_3\left(\frac{n\tau^d}{\log n}\right)^{1/2}}. \quad (8)$$

As a result, with probability at least $1 - c_2 n^{-c_3(n\tau^d/\log n)^{1/2}}$ we have

$$\|A_{nx}R_{nx}\|_2 \leq L\|K\|_\infty M^{1/2}(\bar{\mu} + 1)\tau^\beta. \quad (9)$$

Finally, combining (5), (7) and (9), for any given $r > 0$ there exists constant c large enough such that with probability at least $1 - n^{-r}$

$$\sup_x |\hat{\eta}(x) - \eta(x)| \leq c\left\{\left(\frac{\log n}{n\tau^d}\right)^{1/2} + \tau^\beta\right\},$$

as claimed in Theorem 1. □

REFERENCES

AUDIBERT, J.-Y. & TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers. *The Annals of Statistics* **35**, 608–633.

GINÉ, E. & GUILLOU, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics* **38**, 907–921.

TALAGRAND, M. (1994). Sharper bounds for Gaussian and empirical processes. *The Annals of Probability* **22**, 28–76.

TALAGRAND, M. (1996). New concentration inequalities in product spaces. *Inventiones Mathematicae* **126**, 505–563.

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. New York: Springer.