Selective Nonparametric Regression via Testing

Fedor Noskov FNOSKOV@HSE.RU
HSE University,
Institute for Information Transmission Problems RAS,
and Moscow Institute of Science and Technology (MIPT), Moscow, Russia

Alexander Fishkov ALEXANDER.FISHKOV@SKOLTECH.RU
Skolkovo Institute of Science and Technology (Skoltech), Moscow, Russia

Maxim Panov PANOV.MAXIM@GMAIL.COM
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE

Editors: Berrin Yanıkoğlu and Wray Buntine

Abstract

Prediction with the possibility of abstention (or selective prediction) is an important problem for error-critical machine learning applications. While well-studied in the classification setup, selective approaches to regression are much less developed. In this work, we consider the nonparametric heteroskedastic regression problem and develop an abstention procedure via testing the hypothesis on the value of the conditional variance at a given point. Unlike existing methods, the proposed one allows to account not only for the value of the variance itself but also for the uncertainty of the corresponding variance predictor. We prove non-asymptotic bounds on the risk of the resulting estimator and show the existence of several different convergence regimes. Theoretical analysis is illustrated with a series of experiments on simulated and real-world data.

Keywords: nonparametric regression, selective regression, prediction with abstention, hypothesis testing

1. Introduction

In many machine learning applications, there exists a possibility to reject the prediction of the model and entrust it to the human or other model. Abstention is usually done based on the estimation of uncertainty in predicted value. In classification problems uncertainty might be measured via the probability of wrong prediction while for regression it corresponds to the expected error. In both cases, the estimation of these quantities is usually much harder than the solution of the initial prediction problem. In this work, we target the problem of regression with abstention (or selective regression) in nonparametric setup.

Related Works. There is a large variety of literature regarding classification with reject option. Most likely, the problem was firstly studied by Chow in papers (Chow, 1957, 1970). Moreover, in the article (Chow, 1970) he introduced a risk function used in the majority of forthcoming works including the present one. Herbei and Wegkamp (2006) studied an optimal procedure for this risk and proved consistency for the proposed plugin rule. Then the research was focused on investigation of either empirical risk minimization among a class of hypotheses (Bartlett and Wegkamp, 2008; Cortes et al., 2016) or on other

© 2023 F. Noskov, A. Fishkov & M. Panov.

types of risk (Denis and Hebiri, 2020; El-Yaniv and Wiener, 2010; Lei, 2014). Benefits of abstention for online and active learning were studied in (Neu and Zhivotovskiy, 2020) and (Puchkin and Zhivotovskiy, 2021) correspondingly. Besides, the problem was studied in a number of more practical works; see, for example, (Grandvalet et al., 2009; Geifman and El-Yaniv, 2019; Nadeem et al., 2009). Finally, conformal prediction approach (Vovk et al., 1999; Shafer and Vovk, 2008) has recently been applied to the classification with reject option (Linusson et al., 2018; Johansson et al., 2023).

Unfortunately, methods for selective regression are much less developed. Zaoui et al. (2020) suggested an approach to regression via a plugin rule. In papers (Shah et al., 2022) and (Salem et al., 2022), authors proposed new approaches of neural network learning for better uncertainty capturing. In (Jiang et al., 2020), the authors suggested an uncertainty measure for regression based on blending and a method to select samples with the least risk given some coverage.

Setup. In this work, we focus on the selective algorithms for regression problems with heteroskedastic noise. We assume that the data (X, Y) is coming from a standard regression model $Y = f(X) + \varepsilon$ with target function f and i.i.d. noise ε. Covariate X is assumed to follow some distribution $p(\cdot)$. The noise variance depends on the input point: $\sigma^2(\mathbf{x}) = \mathrm{Var}[Y \mid X = \mathbf{x}]$. The Chow model (Chow, 1970) assumes that the cost for abstention is given by a fixed value $\lambda > 0$, while for prediction the mean squared risk is paid. The abstention procedure for such a problem can be constructed based on the estimate of the variance $\widehat{\sigma}^2(\mathbf{x})$. The abstention rule $\widehat{\alpha}(\mathbf{x})$ proposed by Zaoui et al. (2020) is given by $\widehat{\alpha}(\mathbf{x}) = \mathrm{I}\left\{\widehat{\sigma}^2(\mathbf{x}) \geqslant \lambda\right\}$. The resulting method was proved to be consistent, and the corresponding rate of convergence was derived under standard nonparametric assumptions on functions f and σ. However, the analysis was done only for the risk averaged over the covariate distribution $p(\mathbf{x})$, while one may expect that the convergence properties at a given point $\mathbf{x}$ may significantly depend on the difference between the variance $\widehat{\sigma}^2(\mathbf{x})$ and cost of abstention λ. Moreover, the performance of the estimator $\widehat{\alpha}(\mathbf{x}) = \mathrm{I}\left\{\widehat{\sigma}^2(\mathbf{x}) \geqslant \lambda\right\}$ depends on how accurately $\widehat{\sigma}^2(\mathbf{x})$ estimates the true variance $\sigma^2(\mathbf{x})$. In particular, $\widehat{\sigma}^2(\mathbf{x})$ might give unreliable predictions in the areas of design space where there is little to no train data. Such situations arise when there is a covariate shift between train and test data. In this work, we aim to conduct in-depth theoretical analysis for the pointwise estimation risk for the considered problem and propose the abstention procedure that would be more robust to covariate shifts than the one based on the plugin rule.

The main **contributions** of our paper are the following.

- We show the natural way to construct the abstention rule for nonparametric heteroskedastic regression based on the hypothesis testing on the variance value at a given point. We implement the method via Nadaraya-Watson kernel estimates of regression and variance functions.

- We prove the accurate finite sample bounds for the risk of the resulting estimator. Our results show that the behavior of the risk significantly depends on the relative values of the variance $\sigma^2(\mathbf{x})$ and the abstention cost λ. The proposed method shows favorable performance over the plugin approach of Zaoui et al. (2020), see Table 1.

- We illustrate the theoretical findings by experiments with simulated and real-world data.

The paper is organized as follows. We introduce the setup of the study in Section 2. We propose a new abstention procedure based on hypothesis testing on the values of the conditional variance in Section 3. Theoretical properties of the developed method are studied in Section 4. Finally, Section 5 illustrates our experimental findings and Section 6 concludes the study.

2. Regression with Abstention

Let us start by formalizing the problem. We assume that we observe pairs (X, Y) with covariate $X \in \mathbb{R}^d$ and output $Y \in \mathbb{R}$. The regression task is to estimate $\mathbb{E}_Y[Y \mid X = \mathbf{x}]$ via some function $\widehat{f}(\mathbf{x})$, where $\mathbb{E}_Y[\cdot \mid X = \mathbf{x}]$ means the expectation over the distribution $Y \mid X = \mathbf{x}$. For the case of regression with abstention, for each $\mathbf{x}$ we decide to accept or to reject the prediction $\widehat{f}(\mathbf{x})$. Thus, we introduce an indicator of abstention $\widehat{\alpha}(\mathbf{x})$ which is equal to 1 if the prediction $\widehat{f}(\mathbf{x})$ was *rejected*. The intuition suggests accepting the prediction if the expected squared error $\mathbb{E}_Y[(\widehat{f}(X) - Y)^2 \mid X = \mathbf{x}]$ is not too large, say less than some λ.

That leads to a natural definition of risk which is a variant of the risk proposed in (Chow, 1970):

$$\mathcal{R}_\lambda(\mathbf{x}) = \mathbb{E}_Y\left[(\widehat{f}(X) - Y)^2 \, \mathrm{I}\{\widehat{\alpha}(\mathbf{x}) = 0\} \mid X = \mathbf{x}\right] + \lambda \, \mathrm{I}\{\widehat{\alpha}(\mathbf{x}) = 1\},$$

where $\mathrm{I}\{\cdot\}$ is an indicator function. The introduced risk has a natural interpretation. If we abstain from prediction then we should pay the fixed cost λ. Otherwise, we pay the expected squared error. Note that the provided risk is not the only option for the problem. For instance, people also considered coverage risk, see (Jiang et al., 2020).

Given a risk function, the following question rises up. What are the estimators that minimize it in each point? We formulate the answer as a proposition.

Proposition 1 *Define $f(\mathbf{x}) = \mathbb{E}[Y \mid X = \mathbf{x}]$ and $\sigma^2(\mathbf{x}) = \mathrm{Var}[Y \mid X = \mathbf{x}]$. Then, f is the optimal estimator of $Y \mid X = \mathbf{x}$ and $\alpha(\mathbf{x}) = \mathrm{I}\left\{\sigma^2(\mathbf{x}) \geqslant \lambda\right\}$ is the optimal abstention function.*

The risk related to the pair $\{f(\mathbf{x}), \alpha(\mathbf{x})\}$ we denote by $\mathcal{R}^*_\lambda(\mathbf{x})$.

3. Abstention via Testing of Variance Values

The setup considered in previous section was previously explored in (Zaoui et al., 2020) where it was proposed to use plugin approach, i.e. use some estimators $\widehat{f}$ and $\widehat{\sigma}^2$ of the population counterparts f and σ^2 directly in the rule given by Proposition 1. Their approach leads to consistent estimators in large sample regime. However, for finite samples not only $\widehat{f}$ can be imperfect but also the variance estimator $\sigma^2(\mathbf{x})$ can become unreliable if $\mathbf{x}$ lies far away from the train set under nonparametric setting. Basically, we might start rejecting or accepting the predictions based on the variance estimate which is far off from the actual variance values.

In this work, we aim to work with this issue by considering the uncertainty in the variance estimator $\sigma^2(\mathbf{x})$ itself. We propose a natural way to take this into account via testing between the following hypotheses:

$$H_0\colon \sigma^2(\mathbf{x}) \geqslant \lambda \text{ vs. } H_1\colon \sigma^2(\mathbf{x}) < \lambda.$$

This problem assumes that it is safer to reject a good prediction than to accept a bad one. It is the standard situation for many applications of selective machine learning.

Construction of the test requires some assumptions on the data which will be the same for the train and the test set. Thus, we introduce the studied model.

Model 1 *Given a sample $X \in \mathbb{R}^d$, the observed label Y is normally distributed with the mean $f(X)$ and the variance $\sigma^2(X)$ for some functions $f \colon \mathbb{R}^d \to \mathbb{R}$, $\sigma^2 \colon \mathbb{R}^d \to \mathbb{R}_+$.*

The normality of the noise is not an obligatory requirement, but it allows computing some constants precisely. In our analysis, we mostly use concentration inequalities that can be naturally extended to sub-Gaussian setting. We will work under general nonparametric assumptions on functions f and σ^2, see the details in Section 4.

3.1. Construction of the Test

Nonparametric estimation offers a variety of tools for regression such as kNN, splines or kernel methods (Tsybakov, 2009). In this work, we stick to kernel approaches and employ celebrated Nadaraya-Watson (NW) method that estimates a function at a point $\mathbf{x}$ via weighted mean of its neighbours. Below, we introduce the method formally.

Let μ be the Lebesgue measure in $\mathbb{R}^d$. For a kernel $K \colon \mathbb{R}^d \to \mathbb{R}_+$, $\int_{\mathbb{R}^d} K(\mathbf{t}) d\mu(\mathbf{t}) = 1$, NW method computes weights of samples $X_1, \ldots, X_n$ at the point $\mathbf{x}$ as

$$\omega_i(\mathbf{x}) = \frac{K\left(\frac{\mathbf{x}-X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{\mathbf{x}-X_i}{h}\right)}, \tag{1}$$

where h is a bandwidth. Typically, h tends to 0 as n tends to infinity. Then, it computes the estimated mean

$$\widehat{f}_n(\mathbf{x}) = \sum_{i=1}^n \omega_i(\mathbf{x}) Y_i$$

of the conditional distribution $Y \mid X = \mathbf{x}$. This approach can be extended for computing the estimator of variance $\mathrm{Var}[Y \mid X = \mathbf{x}]$:

$$\widehat{\sigma}_n^2(\mathbf{x}) = \sum_{i=1}^n \omega_i(\mathbf{x}) Y_i^2 - \left(\sum_{i=1}^n \omega_i(\mathbf{x}) Y_i\right)^2.$$

Generally, estimates for mean and variance can use different kernels and bandwidths. However, we stick to the single choice in this work to make the results simpler and more illustrative.

In the paper (Fan and Yao, 1998), it was shown that under some assumptions on h, n and $K(\cdot)$, we have

$$\sqrt{nh^d}\left(\widehat{\sigma}_n^2(\mathbf{x}) - \sigma^2(\mathbf{x})\right) \underset{nh^d \to \infty}{\longrightarrow} \mathcal{N}\left(0, \sigma_V^2\right), \tag{2}$$

Algorithm 1: Acceptance testing

```
Input: samples $\{(X_i, Y_i)\}_{i=1}^n$, bandwidth $h$, parameters $\lambda, \beta, a$
Output: accept or reject the regression result
Calculate $\widehat{p}_n(\mathbf{x}), \widehat{\sigma}_n^2(\mathbf{x})$
if $\widehat{p}_n(\mathbf{x}) \geqslant \frac{4a}{nh^d}$ and criterion (3) holds then
    accept results of the regression
else
    reject
end
```

where $\sigma_V^2(\mathbf{x}) = \sigma^4(\mathbf{x}) \frac{2\|K\|_2^2}{p(\mathbf{x})}$, $\|K\|_2^2 = \int_{\mathbb{R}^d} K^2(\mathbf{t}) d\mu(\mathbf{t})$ and $p(\cdot)$ is a marginal density of covariates X. Thus, we obtain

$$\lim \sup_{nh^d \to \infty} \mathbb{P}\left(\sigma^2(\mathbf{x}) - \widehat{\sigma}_n^2(\mathbf{x}) \geqslant z_{1-\beta} \sigma^2(\mathbf{x}) \sqrt{\frac{2\|K\|_2^2}{nh^d p(\mathbf{x})}}\right) \leqslant \beta.$$

This convergence result allows to construct confidence sets with the guaranteed asymptotic coverage. Since $\sigma^2(\mathbf{x}) \geqslant \lambda$ under the null hypothesis, we obtain the test

$$\widehat{\sigma}_n^2(\mathbf{x}) \leqslant \lambda \left(1 - z_{1-\beta} \|K\|_2 \sqrt{\frac{2}{nh^d p(\mathbf{x})}}\right). \tag{3}$$

Due to the Slutsky lemma, if we replace $p(\mathbf{x})$ with some consistent estimator $\widehat{p}_n(\mathbf{x})$, the above still will be the test of asymptotic significance level β. For the density estimator, $\widehat{p}_n(\mathbf{x})$ we suggest the nonparametric estimator $\widehat{p}_n(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{\mathbf{x} - X_i}{h}\right)$.

3.2. Abstention Algorithm

The derived test allows to construct the procedure of regression with reject option. The only remaining thing we should check before applying the test is that $\widehat{p}_n(\mathbf{x})$ is not zero. Let a and b be such numbers that $K(\mathbf{t}) \geqslant a \cdot \mathrm{I}\{\|\mathbf{t}\| \leqslant b\}$ for all $\mathbf{t} \in \mathbb{R}^d$. For theoretical purposes we demand $\widehat{p}_n(\mathbf{x})$ to be greater than $4a/(nh^d)$ for any accepted point $\mathbf{x}$, see the details in Section C. From the construction of the test, it also follows that the prediction $\widehat{f}_n(\mathbf{x})$ is rejected independent of the value $\widehat{\sigma}_n^2(\mathbf{x})$ if $\widehat{p}_n(\mathbf{x}) \leqslant \frac{2z_{1-\beta}^2}{nh^d} \int_{\mathbb{R}^d} K^2(\mathbf{t}) d\mu(\mathbf{t})$. The resulting procedure is summarized in Algorithm 1.

The proposed procedure was designed for abstract features in $\mathbb{R}^d$. However, in machine learning applications we often have quite complex features as images or texts, and neural networks are usually used for their processing. The considered method might be coupled with neural networks by applying it to some embedding space induced by a neural network.

4. Theoretical guarantees

In this section, we provide theoretical guarantees for our algorithm. There are some natural assumptions that should hold to obtain our results.

Assumption 1 *The Hessian of the function f exists and is bounded by H_f. Moreover, f is L_f-Lipschitz.*

Assumption 1 helps to reduce the bias in the estimation of f. Roughly speaking, if the kernel is symmetric then $\mathbb{E}\widehat{f}_n(\mathbf{x}) - f(\mathbf{x})$ has order at most h^2 times the second derivative of f. Otherwise, h times the Lipschitz constant may appear in the decomposition of the bias $\mathbb{E}\widehat{f}_n(\mathbf{x}) - f(\mathbf{x})$. We also impose the similar assumption for $\sigma^2(\mathbf{x})$.

Assumption 2 *The Hessian of the function σ^2 exists and is bounded by H_σ. Moreover, σ^2 is L_{σ^2}-Lipschitz.*

As was previously mentioned, the bias term of order h vanishes if the kernel K is symmetric. Besides, to estimate f at a point $\mathbf{x}$, the kernel should aggregate well the neighborhood of $\mathbf{x}$. Thus, its support should cover some ball in $\mathbb{R}^d$. But the kernel should not rely on the response provided by far points, so we require exponential tail for the kernel. The most common assumption is that the support of the kernel is bounded, but it is not the case of the Gaussian kernel which is widely used. Formally, the case of the Gaussian kernel implies that $\widehat{p}_n(\mathbf{x})$ is non-zero over the whole space $\mathbb{R}^d$ but we start considering a point $\mathbf{x}$ as explored only if it has estimated density at least $\Theta\big((nh^d)^{-1}\big)$. That allows to derive standard bias-variance decomposition and has a natural interpretation in terms of regression with abstention.

Assumption 3 *For the kernel $K\colon \mathbb{R}^d \to \mathbb{R}_+$, there exist constants a and b such that*

$$K(\mathbf{t}) \geqslant a\,\mathrm{I}\{\|\mathbf{t}\| \leqslant b\}$$

holds for all $\mathbf{t} \in \mathbb{R}^d$. The kernel is symmetric, i.e. $K(\mathbf{t}) = K(-\mathbf{t})$. Moreover, there are constants R_K and r_K such that for all $\mathbf{t}$, it holds that

$$K(\mathbf{t}) \leqslant R_K e^{-r_K\|\mathbf{t}\|}.$$

Finally, we impose some conditions on the density $p(\mathbf{x})$. In the classical nonparametric studies, it is usually assumed that the support of $p(\mathbf{x})$ has positive Lebesgue measure so we do not consider nonparametric low-dimensional manifold estimation. We define

$$\mathcal{S}_q = \{\mathbf{x} \in \mathbb{R}^d \mid p(\mathbf{x}) > q\}.$$

We denote the support $\mathrm{cl}(\mathcal{S}_0)$ by $\mathcal{S}$ and the boundary of $\mathcal{S}$ by $\partial\mathcal{S}$. Inside the support $\mathcal{S}$ we require $p(\mathbf{x})$ to be Lipschitz. That also helps to suppress summands of order h in the bias of our estimator. So the density can be non-continuous at the boundary like the uniform distribution, but it will not affect the inference inside the support.

Assumption 4 *The density p of X is L_p-Lipschitz in $\mathcal{S}_0$ and bounded by C_p.*

To bound the excess risk at a point $\mathbf{x}$, we need its neighborhood to be explored a bit. So there should be large enough probability mass in a ball of radius h around $\mathbf{x}$. Thus, we require $p(\mathbf{x})$ to be larger than Ch and the Euclidean distance to the boundary $d(\mathbf{x}, \partial\mathcal{S})$ to be at least $C'h$. If $\partial\mathcal{S} = \varnothing$, then $d(\mathbf{x}, \partial\mathcal{S})$ is assumed to be infinite.

Finally, note that $\mathcal{R}_\lambda(\cdot)$ depends on the training set. We bound the mean of the excess risk over all training sets $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ where X_i are i.i.d. samples from the density $p(\cdot)$ and Y_i generated according to Model 1.

In the theorem below, we study the upper bounds for the risk. The notation $\lesssim$ means that the corresponding inequality holds with some multiplicative constant that is independent of n, h, β and $p(\mathbf{x})$. The formulation with all the constants presented in the explicit way is given in Supplementary Material, see Theorem 4.

Theorem 2 *Suppose that Assumptions 1-4 hold. Define $\Delta(\mathbf{x}) = |\sigma^2(\mathbf{x}) - \lambda|$. Let $\mathcal{E}_\lambda(\mathbf{x})$ be the excess risk of the estimator $\widehat{f}_n(\mathbf{x})$ and the abstention rule $\widehat{\alpha}_n(\mathbf{x})$ introduced in Algorithm 1. Let $\mathbb{E}_\mathcal{D}$ be the expectation with respect to training dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$, where $X_1, \dots, X_n$ are i.i.d. samples from then density $p(\cdot)$. Then*

- *if $\sigma^2(\mathbf{x}) \geqslant \lambda$ and $\Delta(\mathbf{x}) \leqslant C_1\{nh^d p(\mathbf{x})\}^{-1} + C_2 h^2/p(\mathbf{x}) - C_3 z_{1-\beta}\{nh^d p(\mathbf{x})\}^{-1/2}$, we have*
$$\mathbb{E}_\mathcal{D}(\mathcal{E}_\lambda(\mathbf{x})) \lesssim \{nh^d p(\mathbf{x})\}^{-1} + h^4 p^{-2}(\mathbf{x}) + \Delta(\mathbf{x}),$$
- *if $\sigma^2(\mathbf{x}) \geqslant \lambda$ and $\Delta(\mathbf{x}) \geqslant C_1\{nh^d p(\mathbf{x})\}^{-1} + C_2 h^2/p(\mathbf{x}) - C_3 z_{1-\beta}\{nh^d p(\mathbf{x})\}^{-1/2}$, we have*
$$\mathbb{E}_\mathcal{D}(\mathcal{E}_\lambda(\mathbf{x})) \lesssim \Delta(\mathbf{x}) \exp\Big(-\Omega(nh^{d+2} p(\mathbf{x}))\Big)$$
- *if $\sigma^2(\mathbf{x}) \geqslant \lambda$ and $\Delta(\mathbf{x}) \geqslant C_1\{nh^d p(\mathbf{x})\}^{-1} + C_2 h - C_3 z_{1-\beta}\{nh^d p(\mathbf{x})\}^{-1/2}$, we have*
$$\mathbb{E}_\mathcal{D}(\mathcal{E}_\lambda(\mathbf{x})) \lesssim \exp\{-nh^d p(\mathbf{x})\},$$
- *if $\sigma^2(\mathbf{x}) < \lambda$ and $\Delta(\mathbf{x}) \leqslant C_1'\{nh^d p(\mathbf{x})\}^{-1} + C_2' h^2/p(\mathbf{x}) + C_3' z_{1-\beta}\{nh^d p(\mathbf{x})\}^{-1/2}$, we have*
$$\mathbb{E}_\mathcal{D}(\mathcal{E}_\lambda(\mathbf{x})) \lesssim \{nh^d p(\mathbf{x})\}^{-1} + h^4 p^{-2}(\mathbf{x}) + \Delta(\mathbf{x}),$$
- *if $\sigma^2(\mathbf{x}) < \lambda$ and $\Delta(\mathbf{x}) \gg C_1'\{nh^d p(\mathbf{x})\}^{-1} + C_2' h^2/p(\mathbf{x}) + C_3' z_{1-\beta}\{nh^d p(\mathbf{x})\}^{-1/2}$, we have*
$$\mathbb{E}_\mathcal{D}(\mathcal{E}_\lambda(\mathbf{x})) \lesssim \{nh^d p(\mathbf{x})\}^{-1} + h^4 p^{-2}(\mathbf{x}).$$

Let us note that Theorem 2 applies not only to Algorithm 1 but also to the plugin estimator proposed by Zaoui et al. (2020). Indeed, by setting $\beta = 0.5$ one gets $z_{1-\beta} = 0$ and we obtain plugin approach as a particular instance of our algorithm. While Theorem 2 determines only the upper bound of the risk, it satisfactorily captures the real behavior of Algorithm 1, see experimental evaluation in Section 5. Below, we discuss different estimation regimes implied by Theorem 2.

For beginning, we consider the case when $\sigma^2(\mathbf{x}) > \lambda$. In most of the applications, we assume that $nh^d \to \infty$ as n tends to infinity. Typically, h is chosen to minimize bias-variance trade-off so $h = \Theta\big(n^{-1/(d+4)}\big)$. Assume additionally that $\beta < 0.5$ and

$$C_1\{nh^d p(\mathbf{x})\}^{-1} + C_2 h^2/p(\mathbf{x}) - C_3 z_{1-\beta}\{nh^d p(\mathbf{x})\}^{-1/2} < 0,$$

where constants C_1, C_2, C_3 come from the first case of Theorem 2. This inequality can be satisfied if $h = C_\beta n^{-1/(d+4)} p^{1/2}(\mathbf{x})$ for a small enough constant C_β that depends on β. We refer to this condition as *"undersmoothing"* since it requires the bias to be significantly less than the variance. Moreover, a similar condition is required to ensure (2). Then, our approach provably becomes very efficient. Indeed, in that case the condition $\Delta(\mathbf{x}) \leqslant C_1\{nh^d p(\mathbf{x})\}^{-1} + C_2 h^2/p(\mathbf{x}) - C_3 z_{1-\beta}\{nh^d p(\mathbf{x})\}^{-1/2}$ can be simplified as $\Delta(\mathbf{x}) < 0$ so it never holds. Thus, for any $\mathbf{x}$ such that $\sigma^2(\mathbf{x}) > \lambda$, the expected excess risk converges exponentially. But if one chooses larger h, the advantages of our algorithm remain, since it becomes to converge exponentially earlier than the plugin.

For the plugin, our upper bound can not achieve exponential convergence rates while $\Delta(\mathbf{x}) \leqslant C_1\{nh^d p(\mathbf{x})\}^{-1} + C_2 h^2/p(\mathbf{x})$. That matches our observations for synthetic data, see Figure 1(*b*) and Figure 3(*c*).

To explain the behaviour of estimators for $\sigma^2(\mathbf{x}) \leqslant \lambda$, we impose the following proposition.

Proposition 3 *For any pair of estimators $(\widehat{f}, \widehat{\alpha})$ the expected excess risk can be decomposed as follows:*

$$\mathbb{E}_{\mathcal{D}}\mathcal{R}_\lambda(\mathbf{x}) - \mathcal{R}^*(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}\left[\big(\widehat{f}(\mathbf{x}) - f(\mathbf{x})\big)^2 \mathrm{I}\left\{\widehat{\alpha}(\mathbf{x}) = 0\right\}\right] + \Delta(\mathbf{x}) \cdot \mathbb{P}\big(\widehat{\alpha}(\mathbf{x}) \neq \alpha(\mathbf{x})\big).$$

In our case

$$\begin{aligned}\mathbb{P}\big(\widehat{\alpha}(\mathbf{x}) \neq \alpha(\mathbf{x})\big) &\leqslant \mathbb{P}\left(\widehat{\sigma}_n^2(\mathbf{x}) \leqslant \lambda\left[1 - \frac{C z_{1-\beta}}{\sqrt{nh^d \widehat{p}_n(\mathbf{x})}}\right]\right) \\ &\leqslant \mathbb{P}\left(\widehat{\sigma}_n^2(\mathbf{x}) - \sigma^2(\mathbf{x}) \leqslant \Delta(\mathbf{x}) - \frac{C\lambda z_{1-\beta}}{\sqrt{nh^d \widehat{p}_n(\mathbf{x})}}\right) =: \mathbf{P}(\mathbf{x})\end{aligned}$$

The whole set $\{\mathbf{x} \mid \sigma^2(\mathbf{x}) \leqslant \lambda\}$ can be divided into two sets. Roughly speaking, one is $\mathcal{A} = \{\mathbf{x} \mid \Delta(\mathbf{x}) \lesssim (nh^d)^{-1/2}\}$ and the other is $\mathcal{B} = \{\mathbf{x} \mid \Delta(\mathbf{x}) \gg (nh^d)^{-1/2}\}$. While $\mathbf{x} \in \mathcal{A}$, the leading term of the excess risk is $\Delta(\mathbf{x}) \cdot \mathbf{P}(\mathbf{x})$ that has order $(nh^d)^{-1/2}$. The factor $\mathbf{P}(\mathbf{x})$ does not go to zero, since, informally, $\sqrt{nh^d}\big(\widehat{\sigma}_n^2(\mathbf{x}) - \sigma^2(\mathbf{x})\big) \approx \mathcal{N}\left(0, \sigma_V^2\right)$ due to (2) and so the difference between λ and $\sigma^2(\mathbf{x})$ can not be captured by $\widehat{\sigma}_n^2(\mathbf{x})$. While this argument is not strict from the theoretical point of view, one may prove anticoncentration bounds via the Carbery-Wright theorem. On Figure 1(*b*), one may observe sets $\mathcal{A}$ for different n as hills on the left of the point where $\sigma^2(\mathbf{x}) = \lambda$.

But if $\mathbf{x} \in \mathcal{B}$, bias and variance suppress term $\Delta(\mathbf{x}) \cdot \mathbf{P}(\mathbf{x})$ and we obtain usual rates of convergence for nonparametric estimators. Since for small n the set $\mathcal{A}$ is large there maybe some warm-up when we see slower rates of convergence on plots. So in each point, the convergence may have two phases: one is when $\mathbf{x} \in \mathcal{A}$ and the other is when $\mathbf{x} \in \mathcal{B}$. That is how we explain two phases on Figure 3(*b*) for $\mathbf{x} \in \{-1.6, -0.5, 0.3\}$.

We summarize the behaviour of our estimator and estimator proposed by Zaoui et al. (2020) in Table 1.

4.1. Sketch of the Proof

We start by the following bound on the kernel values:

$$a\,\mathrm{I}\left\{X_i \in \mathcal{B}_{bh}(\mathbf{x})\right\} \leqslant K\left(\frac{X_i - \mathbf{x}}{h}\right),$$

$\sigma^2(\mathbf{x}) \geqslant \lambda$	$\Delta(\mathbf{x}) < C_1 h^2/p(\mathbf{x})$	$C_2 h > \Delta(\mathbf{x}) > C_3 h^2/p(\mathbf{x})$	$\Delta(\mathbf{x}) \gg h$
testing-based	$O(h^2)/p(\mathbf{x}) \cdot \exp\{-\Omega(nh^{d+2}/p(\mathbf{x}))\}$	$O(h) \cdot \exp\{-\Omega(nh^{d+2}p(\mathbf{x}))\}$	$\exp\{-\Omega(nh^{d}p(\mathbf{x}))\}$
plugin	$O\left(h^2/p(\mathbf{x})\right)$		

$\sigma^2(\mathbf{x}) < \lambda$	$\Delta(\mathbf{x}) \lesssim \left(nh^d p(\mathbf{x})\right)^{-1/2}$	$\Delta(\mathbf{x}) \gg \left(nh^d p(\mathbf{x})\right)^{-1/2}$
testing-based	$O(\{nh^d p(\mathbf{x})\}^{-1/2})$	$O(\{nh^d p(\mathbf{x})\}^{-1}) + O(h^4/p^2(\mathbf{x}))$
plugin		

Table 1: The upper bounds derived in Theorem 2, the case of undersmoothing.

where $\mathcal{B}_r(\mathbf{x})$ is a ball with radius r and center $\mathbf{x}$. These bounds allow us to deal with values of the kernel like they are Bernoulli random variable with certain mean. Thus, we show that with probability $1 - C\exp^{-\Omega(nh^d p(\mathbf{x}))}$, we have

$$ab^d \omega_d p(\mathbf{x}) \leqslant \widehat{p}_n(\mathbf{x}) \leqslant 2p(\mathbf{x}) + L_p h \int_{\mathbb{R}^d} \|\mathbf{t}\| K(\mathbf{t}) d\mu(\mathbf{t}),$$

see Propositions 5 and 19 in Supplementary Material. The bounds above are rough but they will be sufficient for our purposes.

For any L-Lipschitz function g we also can obtain the bound

$$\left| \sum_{i=1}^{n} g(X_i)\omega(X_i) - g(\mathbf{x}) \right| \leqslant \sum_{i=1}^{n} |g(X_i) - g(\mathbf{x})| \omega(X_i) \lesssim \frac{Lh}{p(\mathbf{x}) nh^d},$$

since $\omega(X_i)$ is, roughly speaking, $\frac{R_K e^{-r_K \|(X_i - \mathbf{x})/h\|}}{p(\mathbf{x}) nh^d}$ up to a constant, see Corollary 6 in Supplementary Material. This approximation is based on the fact that $K(\mathbf{t})$ is bounded above by a constant and the denominator of the weight with high probability is $\Omega\left(nh^d p(\mathbf{x})\right)$, see Proposition 5 in Supplementary Material. Finally, under some conditions on n, h and $p(\mathbf{x})$ we establish the concentration bounds for any function g which is Lipschitz and has bounded Hessian, see Corollary 9 in Supplementary Material.

If $\sigma^2(\mathbf{x}) \geqslant \lambda$ then $\mathrm{I}\{\widehat{\alpha}(\mathbf{x}) = 0\} = \mathrm{I}\{\widehat{\alpha}(\mathbf{x}) \neq \alpha(\mathbf{x})\}$. So we may bound

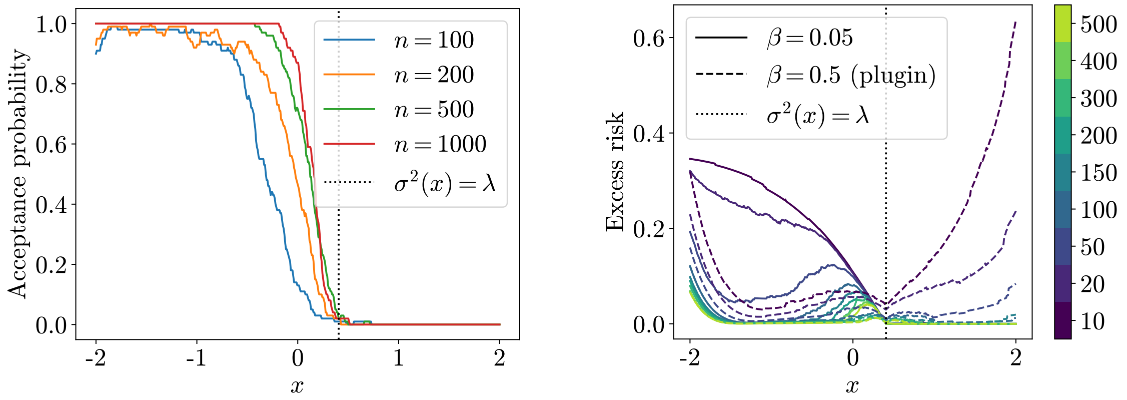
$$\mathbb{E}_{\mathcal{D}} \left(\widehat{f}(\mathbf{x}) - f(\mathbf{x})\right)^2 \mathrm{I}\{\widehat{\alpha}(\mathbf{x}) = 0\} \leqslant \sqrt{\mathbb{E}_{\mathcal{D}} \left(\widehat{f}(\mathbf{x}) - f(\mathbf{x})\right)^4 \mathrm{I}\{\widehat{\alpha}(\mathbf{x}) = 0\}} \cdot \mathbb{P}^{1/2}\left(\widehat{\alpha}(\mathbf{x}) \neq \alpha(\mathbf{x})\right).$$

We bound the 4-th moment above by integrating concentration inequalities. That results in standard bias-variance trade-off, see Lemma 18 in Supplementary Material. Thus, the rate of the excess risk is determined by the factor $\mathbb{P}^{1/2}(\widehat{\alpha}_n(\mathbf{x}) \neq \alpha(\mathbf{x}))$. It can be reformulated as

$$\mathbb{P}^{1/2}\left(|\widehat{\sigma}_n^2(\mathbf{x}) - \sigma^2(\mathbf{x})| \geqslant \Delta(\mathbf{x}) + O\left\{\left(nh^d p(\mathbf{x})\right)^{-1/2}\right\}\right).$$

The random value $\widehat{\sigma}_n^2(\mathbf{x})$ behaves like sub-exponential random variable with the mean $\sigma^2(\mathbf{x}) + o(1)$. Thus, under certain assumptions on $\Delta(\mathbf{x})$, we get exponential rates of convergence via the concentration argument, see Corollary 15 in Supplementary Material.

If $\sigma^2(\mathbf{x}) < \lambda$, two terms from Proposition 3 demonstrate different behaviour. The first one can be bounded via standard bias-variance trade-off. The second one exponentially decreases if h is smaller than some constant and $\Delta(\mathbf{x})$ is larger than some decreasing function of n and h. The proof is similar to the case $\sigma^2(\mathbf{x}) > \lambda$.



(a) Acceptance probability.

(b) Expected excess risk.

Figure 1: Example with synthetic data: $X \sim \mathcal{U}(-2,2)$, $\sigma(x) = \mathtt{sigmoid}(x)$. We sample multiple datasets of each sample size n. Confidence level $\beta = 0.05$ and abstention cost $\lambda = 0.36$.

5. Experiments

5.1. How to choose λ and β

In practice, two natural questions arise: how to choose λ and how to choose β. Obviously, one may define λ from the formulation of the problem as an inappropriate level of noise. The case of β is a bit more sophisticated. From Algorithm 1, we infer that any $\mathbf{x}$ will be rejected if $\widehat{p}_n(\mathbf{x}) \leqslant \frac{2\|K\|_2^2 z_{1-\beta}^2}{nh^d}$. Thus, any such $\mathbf{x}$ is considered as outlier, and, hence, $z_{1-\beta}$ is a tolerance level for outliers. Additionally, the choice of β determines the trade-off between type I and type II errors.

5.2. Synthetic data

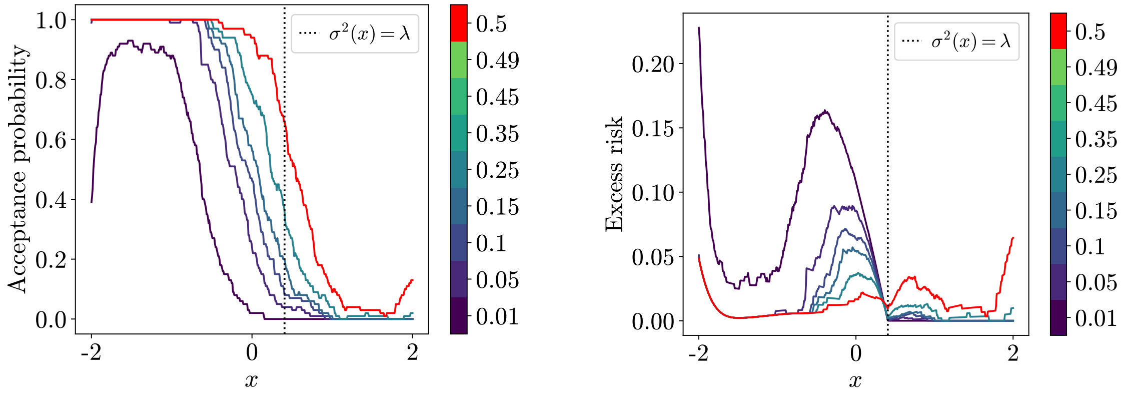
For the first part of the experiments we use one dimensional data with known simple functions as true mean and variance at each point:

$$Y = f(X) + \sigma(X)\varepsilon,\ X \sim p(\cdot),\ \varepsilon \sim \mathcal{N}(0,1). \tag{4}$$

Specifically, we consider normal and uniform distributions of the independent variable $p(\cdot) \in \{\mathcal{N}(0,1), \mathcal{U}(-2,2)\}$, a fixed mean function $f(x) = \frac{x^2}{4}$, and two choices of standard deviation: sigmoid function and Heaviside function. Parameter λ was fixed at 0.36 and parameter $\beta = 0.05$ unless otherwise noted. Optimal bandwidth was selected using leave-one-out cross-validation optimizing mean squared error of prediction by NW estimator. In all our experiments for each setting of hyperparameters we have generated 100 different random datasets from our data model and then averaged the results.

5.2.1. Convergence of estimates

We sampled 100 datasets of sizes $n \in \{100, 200, 500, 1000\}$ and for each $x \in [-2, 2]$ we estimate the fraction of predictions that are accepted by the proposed method. We present



(a) Acceptance probability.

(b) Expected excess risk.

Figure 2: Example with synthetic data: $X \sim \mathcal{U}(-2, 2)$, $\sigma(x) = \texttt{sigmoid}(x)$. Sample size $n = 100$, abstention cost $\lambda = 0.36$. We apply the proposed testing-based method for different values of β. Plugin approach corresponds to $\beta = 0.5$.

the resulting chart in Figure 1(a) for $X \sim \mathcal{U}(-2, 2)$, $\sigma(x) = \texttt{sigmoid}(x)$, additional charts are in Supplementary Material, Section D.1.1. The results demonstrate that for the area with $\sigma^2(x) > \lambda$ (to the right of the dashed line) the convergence is much faster than for the area with $\sigma^2(x) < \lambda$.

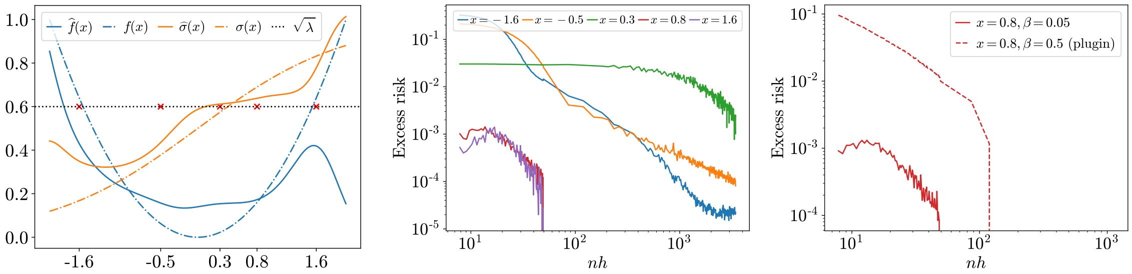
Additionally, we also estimate expected excess risk, since we know the values of $f(x)$ and $\sigma(x)$ for any x. For the first plot (see Figure 1(b)) we vary sample size n from 10 to 500. We compare the proposed approach with $\beta = 0.05$ with "plugin" baseline, corresponding to $\beta = 0.5$. For the testing-based method we see the very quick convergence for $\sigma^2(x) > \lambda$. For the points with $\sigma^2(x) < \lambda$ the convergence is slower especially for the smaller values of $\Delta(x)$. Thus, the observed behaviour well corresponds to the one predicted by the theory. For the plugin approach, the convergence is slower especially for the points with $\sigma^2(x) > \lambda$.

5.2.2. Dependence on β

In this experiment, we have studied the behavior of our method when changing its only hyperparameter β in the range between 0.01 and 0.5. Since $\beta = 0.5$ corresponds to "plugin" method described previously, we show it in red. For this we fixed the number of samples at $n = 100$, sampled 100 datasets and calculated the expected excess risk for each $x \in [-2, 2]$. With the increase of β the method becomes less conservative (more accepts), see Figure 2(a). It leads to the increased expected risk at points where prediction should be rejected and decreased risk at the points where predictions should be accepted, see Figure 2(b). Thus, in practice parameter β might be selected based on the trade-off between these two errors depending on the particular features of the considered applied problem.

5.2.3. Pointwise convergence

Finally, we sampled multiple datasets of increasing sizes n from 10 to 20000 and selected 5 diagnostic points: $x \in \{-1.6, -0.5, 0.3, 0.8, 1.6\}$, see Figure 3$(a)$. When sample size is



(a) True mean and standard deviation, and their estimates for $n = 100$ points.

(b) Expected excess risk.

(c) Comparison with plugin method at a single point.

Figure 3: We sample multiple datasets of each size and for selected points $x \in \{-1.6, -0.5, 0.3, 0.8, 1.6\}$ calculate the expected excess risk. In this experiment $X \sim \mathcal{U}(-2, 2)$, $\sigma(x) = \texttt{sigmoid}(x)$, abstention cost $\lambda = 0.36$, $\beta = 0.05$.

less than 100, we generate 20000 datasets of each size, while for larger sample size we only use 100. In order to perform a more straightforward averaging of the across datasets of the same size, we have used the same bandwidth $h \sim \frac{1}{n^5}$ that was selected to show the expected polynomial dependence in nh of the risk at points x with $\sigma^2(x) < \lambda$. The resulting dependencies of the risk on nh are depicted on Figure 3(b). We observe all the main outcomes predicted by the theory:

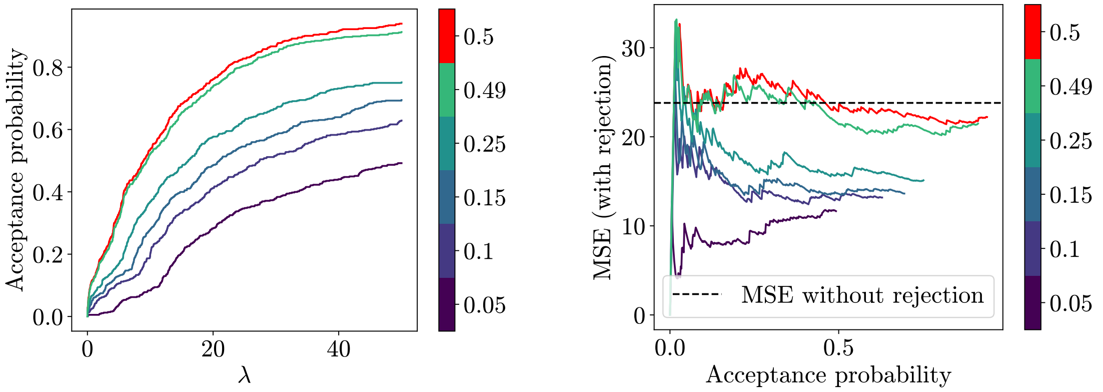
- rapid convergence of the risk for the points with $\sigma^2(x) > \lambda$ (points $x = 0.8$ and $x = 1.6$);
- polynomial convergence of the risk as function of nh for $\sigma^2(x) < \lambda$ with moderately large values of $\Delta(x)$, i.e., points $x = -1.6$ and $x = -0.5$;
- very slow convergence for the point with $\sigma^2(x) < \lambda$ and small value of $\Delta(x)$.

Additionally, on Figure 3(c) we experimentally confirm that plugin method has slower convergence than testing-based method for $\sigma^2(x) > \lambda$.

5.3. Airfoil Self-Noise Data Set

We have tested our method on the Airfoil dataset from the UCI collection (Dua and Graff, 2017). We do not perform any special preprocessing of the data or feature engineering, only standard scaling of features. We prepare train and test sets in two steps. First, we select a pivot feature and put 70% of the data with the lowest values of this feature to part A and the rest of the data becomes part B. For the second step we select 20% of each part (sampled uniformly) and put it in the other part. First part becomes the train set and the second part the test set. In this way we guarantee that test set will have data with low values of $\widehat{p}(\mathbf{x})$ as well as data distributed similar to train data.

In our experiments we select different features as pivots for the split and then vary $\lambda \in [0, 50]$, calculating acceptance (retention) fraction and mean squared error. We present



(a) Fraction of accepted points for each threshold λ.

(b) MSE with rejection.

Figure 4: Airfoil data, split 70/30 by the second feature. For $\lambda \in [0, 50]$ we calculate acceptance (retention) probability and MSE at accepted points. x-values of acceptance probabilities are inferred as fraction of accepted points for each λ.

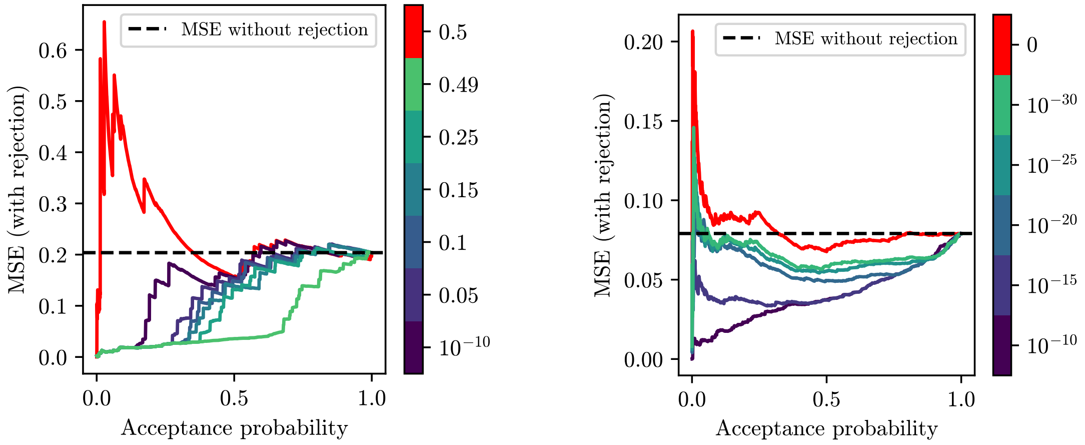
results for splitting by the second feature: "Angle of attack". Other configurations can be found in the Supplementary Material, Section D.2. On Figure 4(a) we show how acceptance probability varies as a function of λ. Figure 4(b) illustrates the dependence of the mean squared error of estimation as a function of the fraction of points accepted for prediction. The curves show the expected trend to increase when accepting more points. Using more conservative estimates one obtains higher accuracy for the given acceptance rate. However, by construction, high acceptance rates are not achievable for the proposed method due to the limitations on the values of the estimated density $\widehat{p}_n(\mathbf{x})$.

5.4. CPU-small Data Set

Another dataset from UCI collection that we used is CPU-small. This dataset has 8192 instances and 12 features. Data splitting is done in the same manner as the Airfoil dataset. During the preprocessing we standardize the training data to have zero mean and unit variance and then apply the same scaling to the test set. Splitting was done based on the first feature "lread".

In this experiment we have tried two scenarios: first is to use the data as it is and the second is to use higher dimensional version (embeddings) of the data, obtained with a neural network. First we trained a two layer neural network with 50 neurons in each layer and ReLU activations and then used the values from the last layer as input for our method. On Figures 5(a) and 5(b) we present the partial MSE scores obtained with rejection in these two setups.

Using embeddings provides much lower MSE without rejection as one would expect. It also shows that using our method we can significantly outperform the baseline plug-in method. For this dataset our algorithm is less sensitive to the choice of β than for the previous one. We opted to vary $z_{1-\beta}$ directly in the embedding case since for a higher



(a) Raw data.

(b) Embeddings.

Figure 5: CPU-small data, split 70/30 by the first feature. For $\lambda \in [0, 2]$ we calculate acceptance (retention) probability and MSE at accepted points. For the raw data we vary β, in case of embeddings we vary $z_{1-\beta}$ directly due to higher dimensionality of the data.

dimension of the data the values of $\widehat{p}$ span a larger strip in the logarithmic scale. In order to show dependence on β we would need to choose values very close to 0.5. Choosing the same set of β values as for raw data case yields curves similar to $z_{1-\beta} = 10^{-10}$.

6. Conclusion

In this work, propose a new method for selective prediction in heteroskedastic regression tasks under the Chow risk model. The method is based on the natural idea of testing the values of conditional variance at a given point. Our theoretical analysis show the existence of exponential and polynomial convergence regimes that depend on the relative values of the variance and abstention cost. The proposed method compares favorably to the plugin baseline both in theory and in the conducted experimental valuation.

Acknowledgments

The research was supported by the Russian Science Foundation grant 21-11-00373. The work of F. Noskov was prepared within the framework of the HSE University Basic Research Program.

References

Peter L. Bartlett and Marten H. Wegkamp. Classification with a Reject Option using a Hinge Loss. *Journal of Machine Learning Research*, 9(59):1823–1840, 2008. URL http://jmlr.org/papers/v9/bartlett08a.html.

C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, January 1970. Conference Name: IEEE Transactions on Information Theory.

C. K. Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, EC-6(4):247–254, December 1957. Conference Name: IRE Transactions on Electronic Computers.

Corinna Cortes, Giulia DeSalvo, and M. Mohri. Learning with Rejection. In *ALT*, 2016.

Christophe Denis and Mohamed Hebiri. Consistency of plug-in confidence sets for classification in semi-supervised learning. *Journal of Nonparametric Statistics*, 32(1):42–72, 2020.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Ran El-Yaniv and Yair Wiener. On the Foundations of Noise-free Selective Classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL http://jmlr.org/papers/v11/el-yaniv10a.html.

Jianqing Fan and Qiwei Yao. Efficient Estimation of Conditional Variance Functions in Stochastic Regression. *Biometrika*, 85(3):645–660, 1998. URL http://www.jstor.org/stable/2337393. Publisher: [Oxford University Press, Biometrika Trust].

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2151–2159. PMLR, May 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Yves Grandvalet, Alain Rakotomamonjy, Joseph Keshet, and Stéphane Canu. Support Vector Machines with a Reject Option. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2009. URL https://papers.nips.cc/paper/2008/hash/3df1d4b96d8976ff5986393e8767f5b2-Abstract.html.

Radu Herbei and Marten H. Wegkamp. Classification with Reject Option. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 34(4):709–721, 2006. URL https://www.jstor.org/stable/20445230. Publisher: [Statistical Society of Canada, Wiley].

Wenming Jiang, Ying Zhao, and Zehan Wang. Risk-controlled selective prediction for regression deep neural network models. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020.

Ulf Johansson, Tuwe Löfström, Cecilia Sönströd, and Helena Löfström. Conformal prediction for accuracy guarantees in classification with reject option. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 133–145. Springer, 2023.

Jing Lei. Classification with confidence. *Biometrika*, 101(4):755–769, 2014. URL https://www.jstor.org/stable/43304686. Publisher: [Oxford University Press, Biometrika Trust].

Henrik Linusson, Ulf Johansson, Henrik Boström, and Tuve Löfström. Classification with reject option using conformal prediction. In *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part I 22*, pages 94–105. Springer, 2018.

Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. Accuracy-Rejection Curves (ARCs) for Comparing Classification Methods with a Reject Option. In *Proceedings of the third International Workshop on Machine Learning in Systems Biology*, pages 65–81. PMLR, March 2009. URL https://proceedings.mlr.press/v8/nadeem10a.html.

Gergely Neu and Nikita Zhivotovskiy. Fast rates for online prediction with abstention. In *Conference on Learning Theory*, pages 3030–3048. PMLR, 2020.

Nikita Puchkin and Nikita Zhivotovskiy. Exponential savings in agnostic active learning through abstention. In *Conference on Learning Theory*, pages 3806–3832. PMLR, 2021.

Mahmoud Salem, Mohamed Osama Ahmed, Frederick Tung, and Gabriel Oliveira. Gumbel-softmax selective networks, 2022. URL http://arxiv.org/abs/2211.10564.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Abhin Shah, Yuheng Bu, Joshua K. Lee, Subhro Das, Rameswar Panda, Prasanna Sattigeri, and Gregory W. Wornell. Selective regression under fairness criteria. In *Proceedings of the 39th International Conference on Machine Learning*, pages 19598–19615. PMLR, 2022. URL https://proceedings.mlr.press/v162/shah22a.html.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009. URL http://link.springer.com/10.1007/b13794.

Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 444–453, 1999.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. URL https://www.cambridge.org/core/books/highdimensional-statistics/8A91ECEEC38F46DAB53E9FF8757C7A4E.

Ahmed Zaoui, Christophe Denis, and Mohamed Hebiri. Regression with reject option and application to knn. *Advances in Neural Information Processing Systems*, 33:20073–20082, 2020.