Prediction intervals with controlled length in the heteroscedastic Gaussian regression

Christophe Denis, Mohamed Hebiri, Ahmed Zaoui

Prediction intervals with controlled length in the heteroscedastic Gaussian regression

Christophe Denis, Mohamed Hebiri, and Ahmed Zaoui
LAMA, UMR-CNRS 8050,
Université Gustave Eiffel

Abstract

We tackle the problem of building a prediction interval in heteroscedastic Gaussian regression. We focus on prediction intervals with constrained expected length in order to guarantee interpretability of the output. In this framework, we derive a closed form expression of the optimal prediction interval that allows for the development a data-driven prediction interval based on plug-in. The construction of the proposed algorithm is based on two samples, one labeled and another unlabeled. Under mild conditions, we show that our procedure is asymptotically as good as the optimal prediction interval both in terms of expected length and error rate. In particular, the control of the expected length is distribution-free. We also derive rates of convergence under smoothness and the Tsybakov noise conditions. We conduct a numerical analysis that exhibits the good performance of our method. It also indicates that even with a few amount of unlabeled data, our method is very effective in enforcing the length constraint.

1 Introduction

Prediction is one of the main goals in supervised learning, it consists in building, given historical data, a candidate output for a new observation. One common practice thereafter is to carry out inference on the output and then to ask for confidence in the predicted value, therefore, *prediction interval (PI)* appears as appropriate tools to handle this problem in the regression setting. A typical application is the prediction in the linear regression case when the data are assumed Gaussian with common variance. In this context, the notion of PI is well studied and well understood both from practice and theory.

However, in the general case, inference as a post-processing step may produce irrelevant conclusions due to the stochastic nature of the data-driven prediction procedure (see for instance [3]). Therefore, in order to guarantee the theoretical validity of the prediction intervals, it is suitable to process at once both aspects of the problem, that is, one might design a data-driven procedure directly devoted to the *prediction interval* purpose.

In a classical setting of PI, one often asks for a pre-specified level of confidence for the predicted range of values (says 95% or 99% according to the problem). This is for instance the approach that is considered in the *conformal prediction* literature [19, 18, 12, 11]. However, this strategy may suffer from interpretability issues for problems where prediction task is difficult or when classical assumptions on the noise are not satisfied. Specifically, for relatively restrictive values of the confidence level, the resulting output might be so large that it becomes useless.

In contrast, our purpose is to produce for future observation a prediction interval with a pre-determined expected length. This framework is completely different from the previous one since it does not ensure any coverage guarantee but rather ensures the interpretability of the predicted output. Indeed, since the length of the output interval is controlled, we do not expect for a given input instance $\mathbf{x} \in \mathbb{R}^d$ a too large set of candidate values.

Generally speaking, the range of values that we would output with PI has no reason to be an interval. However, in a Gaussian model, this range of values indeed forms an interval (or a union of it). In this paper, we investigate the problem of PI under expected length constraint in the Gaussian regression setup. We aim at providing a general device that outputs a PI for a new feature. Our procedure relies on the plug-in principle and we propose in the present contribution a statistical analysis of it in this setting.

Main contributions. Denote by $\Gamma : \mathbb{R}^d \to \mathcal{P}(\mathbb{R})$ a given prediction set, where $\mathcal{P}(\mathbb{R})$ is the set of subsets of $\mathbb{R}$. One of the main inputs of the present work is the introduction of a novel framework for PI in the regression setting, taking sides of controlling the expected size $\mathbb{E}\left[L(\Gamma(\mathbf{X}))\right]$ of the output predictor Γ while minimizing its error rate $\mathbb{P}\left(Y \notin \Gamma(\mathbf{X})\right)$, where $L(\Gamma(\mathbf{X})) = \int_{\mathbb{R}} \mathbb{1}_{y \in \Gamma(\mathbf{X})} dy$ stands for the Lebesgue measure of Γ. We derive the optimal rule for this problem which is defined as

$$\Gamma_\ell^* \in \underset{\Gamma : \mathbb{E}[L(\Gamma(\mathbf{X}))] \leq \ell}{\operatorname{argmin}} \mathbb{P}\left(Y \notin \Gamma(\mathbf{X})\right) \ ,$$

where $\ell > 0$ is a preset length chosen by the practitioner.

In the Gaussian framework, based on the plug-in principle, we then build a general procedure that estimate the optimum and prove that the resulting empirical predictor performs as well as Γ_ℓ^* both in term of expected length and error rate. Notably, the control on the expected length of the proposed estimator is distribution-free. Furthermore, our algorithm has two appealing properties. It can benefit from a semi-supervised setting and can be applied to any off-the-shelf machine learning algorithm.

On the other hand, we evaluate the performance of our estimator with respect to the symmetric difference distance and a risk measure which properly balances the expected length and the error rate. Specifically, we establish the consistency for our procedure under mild assumptions and provide rates of convergence under suitable assumptions on the distribution of the data.

We additionally conduct a numerical study that confirms our theoretical findings and shows how effective our method is in controlling the length, an important aspect to ensure the interpretability of the output. Finally, we provide a numerical comparison with the strategy which consists in building PI under expected coverage constraint. Our numerical experiment highlights that our proposed approach produced significantly more stable PI. In particular, our algorithm seems to be more adapted when the sample size of the training sample is moderate.

Related works. A first line of work related to PI is *confidence intervals*. This is one of the most popular tools in statistical inference and differs from PI by the fact that the purpose there is to output a range of values for a given parameter of the model such that the mean, while our goal is here the prediction. The spectrum of applications of confidence interval is extremely wide and from some perspective PI can be seen as part of the confidence interval literature where we focus on building a confidence interval for the output of a new observation.

Probably the closest direction of works to ours is *conformal prediction* [19, 12, 11]. The main difference relies on the way the expected length and the error rate of the prediction interval is considered. The goal there is to produce a PI with a pre-specified level of accuracy. The connection of PI with controlled expected size is important to figure out since, *at the population level*, each PI with controlled accuracy corresponds to a PI with controlled expected size. From practice however, the two approaches start to differ. We defer this discussion to Sections 4.2 and 5.2 where a complete comparison to *PI with controlled accuracy* is conducted.

Providing an output with a pre-defined length has rarely been considered. Probably the first reference that deals with such notion is [10]. There, the authors build confidence intervals for the mean and variance in a Gaussian problem that reach given confidence level while being of size L. In contrast to that work, we deal with prediction intervals, our control on the length is in expectation which offers more flexibility on "hard" points, we do not focus on a pre-specified level of confidence but rather minimize the error under a size constraint, and we derive a statistical and a numerical analysis of our method.

Finally, let us notify that constraining the expected length is not novel. It has already been considered in the multi-class classification setting [6, 5]. There, the control of the length is interpreted as the desired average number of output labels. Similar to the present work, the goal is to focus on a set of values for prediction while maintaining the interpretability of the output. The main difference with earlier work is that we deal here with real valued output which is more tricky. From this perspective the present paper is a generalization of these previous works to the Gaussian regression setting.

Outline of the paper. Section 2 provides the main notation and describes the framework of prediction intervals under expected length constraint in the Gaussian regression. In particular, the explicit form of the optimal rule is provided. Section 3 introduces our data-driven procedure as well as its statistical analysis. This theoretical analysis is complemented with a numerical study presented in Section 5.

Additional considerations beyond the Gaussian assumption and other frameworks of prediction intervals are considered in Section 4. A conclusion is provided in Section 6, while the proofs of our results are postponed to the Appendix.

2 General framework

In the present contribution we focus on the Gaussian model, that is, we assume that $(\mathbf{X}, Y) \in \mathbb{R}^d \times \mathbb{R}$ are such that

$$Y = f^*(\mathbf{X}) + \sigma(\mathbf{X})\,\varepsilon \ , \tag{1}$$

where $\varepsilon \sim \mathcal{N}(0,1)$ is independent of $\mathbf{X}$. In this expression, $f^* : \mathbb{R}^d \to \mathbb{R}$ is the regression function and $\sigma : \mathbb{R}^d \to \mathbb{R}_+^*$ is the conditional variance function, both of them assumed to be unknown. The main assumptions that we consider throughout the paper are presented in Section 2.1. The characterization of the optimal prediction interval under expected length constraint is provided in Section 2.2. Finally, we define the measure of performance dedicated to asses the quality of a prediction interval in Section 2.3.

2.1 Assumptions

Given an observation $\mathbf{X} \in \mathbb{R}^d$, our goal is to produce the most accurate, in a certain sense to be specified later, a range of predicted values where the corresponding label $Y \in \mathbb{R}$ lies. Such predictions will be describe a set of $\mathcal{P}(\mathbb{R})$ and denoted by $\Gamma(\mathbf{x})$ for each $\mathbf{x} \in \mathbb{R}^d$. In other words, the predictor Γ is a mapping from $\mathbb{R}^d$ onto $\mathcal{P}(\mathbb{R})$.

Throughout the paper we denote by $p(\cdot|\mathbf{x})$ the conditional density of Y given $\mathbf{x}$, that is, for all $y \in \mathbb{R}$

$$p(y|\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma(\mathbf{x})} \exp\left(-\frac{(y - f^*(\mathbf{x}))^2}{2\sigma^2(\mathbf{x})}\right) \ ,$$

that is, we focus on the heteroscedastic Gaussian regression model. In order to avoid pathological situations, we impose the following mild assumptions on the regression and conditional variance functions.

Assumption 1. *There exist* $0 < \sigma_0 < \sigma_1 < \infty$ *such that for all* $\mathbf{x} \in \mathbb{R}^d$

$$\sigma_0 \leq \sigma(\mathbf{x}) \leq \sigma_1 \ .$$

Assumption 2. *There exists* $C_1 > 0$ *such that*

$$\mathbb{E}\left[|f^*(\mathbf{X})|\right] \leq C_1 \ .$$

In addition, we consider an assumption which is PI context-specific. It ensures in particular the existence and uniqueness and the optimal PI. Note that similar assumption is considered in the set-valued classification framework [5].

Assumption 3 (Continuity)**.** *For all* $y \in \mathbb{R}$*, the mapping* $t \mapsto \mathbb{P}_{\mathbf{X}}(p(y|\mathbf{X}) \geq t)$ *is continuous on* $\mathbb{R}_+^*$*.*

In other word, we assume that $p(y|\mathbf{X})$ is atomless.

2.2 Prediction interval with expected length

For a given predictor Γ two features are of interest, its error rate $\mathbb{P}(Y \notin \Gamma(\mathbf{X}))$ and its expected Lebesgue measure defined as

$$\mathcal{L}(\Gamma) := \mathbb{E}\left[L(\Gamma(\mathbf{X})\right] = \mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\{y \in \Gamma(\mathbf{X})\}} \mathrm{d}y\right] \ .$$

Given $\ell > 0$, we focus on the following problem

$$\Gamma_\ell^* \in \arg\min\{\mathbb{P}\left(Y \notin \Gamma(\mathbf{X})\right) \ : \ \Gamma \text{ such that } \mathcal{L}(\Gamma) \leq \ell\} \ . \tag{2}$$

The next proposition provides the characterization of the optimal predictor under Assumption 3.

Proposition 1. *Let $\ell > 0$, under Assumption 3, the optimal predictor Γ^*_ℓ can be expressed as*

$$\Gamma^*_\ell(\mathbf{X}) = \{y \in \mathbb{R} : \ p(y|\mathbf{X}) \geq \lambda^*_\ell\} \ ,$$

where $\lambda^*_\ell = G^{-1}(\ell)$ with $G(t) := \int_{\mathbb{R}} \mathbb{P}(p(y|\mathbf{X}) \geq t)dy$ for all $t > 0$[1].

The parameter λ^*_ℓ, which corresponds to the value of the generalized inverse function G^{-1} at ℓ, plays a crucial role in our study since it fully determines the optimal predictor Γ^*. This being said, let us comment on Proposition 1. First, an important consequence of the above proposition is that the predictor Γ^*_ℓ is an interval of length ℓ, that is $\mathcal{L}(\Gamma^*_\ell) = \ell$ and we additionally can express Γ^*_ℓ as

$$\Gamma^*_\ell(\mathbf{X}) = \left[f^*(\mathbf{X}) - \sqrt{2\sigma^2(\mathbf{X}) \log\left(\frac{1}{\sqrt{2\pi}\lambda^*_\ell \sigma(\mathbf{X})}\right)} \ , \ f^*(\mathbf{X}) + \sqrt{2\sigma^2(\mathbf{X}) \log\left(\frac{1}{\sqrt{2\pi}\lambda^*_\ell \sigma(\mathbf{X})}\right)} \right] \ .$$

Second, the function G defined in Proposition 1 is the extension to the regression case of the function G defined in [6] in the multi-class setting. Note that the function G is always well-defined and continuous for $t > 0$, since by Markov Inequality and Fubini Theorem,

$$G(t) = \int_{\mathbb{R}} \mathbb{P}(p(y|\mathbf{X}) \geq t)dy \leq \frac{1}{t} \int_{\mathbb{R}} \mathbb{E}\left[p(y|\mathbf{X})\right] \mathrm{d}y \leq \frac{1}{t} \mathbb{E}\left[\int_{\mathbb{R}} p(y|\mathbf{X}) \mathrm{d}y\right] \leq \frac{1}{t} \ .$$

Finally, we highlight that parameter λ^*_ℓ is simply the Lagrange multiplier of the minimization problem defined by Equation (2). Therefore, Γ^*_ℓ can be expressed as the minimizer of the unconstrained problem

$$\Gamma^*_\ell \in \arg\min_{\Gamma} \mathbb{P}\left(Y \notin \Gamma(\mathbf{X})\right) + \lambda^*_\ell \mathbb{E}\left[L(\Gamma(\mathbf{X})\right] \ . \tag{3}$$

2.3 Measures of performance

In this paragraph we introduce two ways to quantify the quality of a given prediction interval Γ . The first one, suggested by Equation (3), balances the error rate and the expected length of the predictor

$$R_\ell(\Gamma) = \mathbb{P}\left(Y \notin \Gamma(\mathbf{X})\right) + \lambda^*_\ell \mathbb{E}\left[L(\Gamma(\mathbf{X})\right] \ ,$$

with $\lambda^*_\ell = G^{-1}(\ell)$. This risk is particularly important from our perspective since minimizing it over all predictors lead to the optimal predictor Γ^*_ℓ, which reaches the requested expected length. A natural "distance" to the optimal predictor is then evaluated through the excess risk

$$\mathcal{E}_\ell\left(\Gamma\right) = R_\ell(\Gamma) - R_\ell(\Gamma^*_\ell) \ .$$

The following proposition provides a closed formula for this term.

Proposition 2. *Let $\ell \geq 0$. For any predictor Γ*

$$\mathcal{E}_\ell(\Gamma) = \mathbb{E}\left[\int_{\Gamma(\mathbf{X}) \triangle \Gamma^*_\ell(\mathbf{X})} |p(y|\mathbf{X}) - \lambda^*_\ell| \ dy\right] \ .$$

Interestingly, the above result shows that the performance of a predictor Γ is directly linked to the behavior of the conditional density $p(y|\mathbf{x})$ around the threshold λ^*_ℓ on the symmetric difference $\{\Gamma(\mathbf{X}) \triangle \Gamma^*_\ell(\mathbf{X})\}$.

A second measure of performance arises naturally when we deal with predictors that are intervals. It is the expectation of symmetric difference between the considered predictor Γ and optimal predictor Γ^*_ℓ defined for all predictor Γ as

$$\mathcal{H}\left(\Gamma\right) = \mathbb{E}\left[L\left(\Gamma(\mathbf{X}) \triangle \Gamma^*_\ell(\mathbf{X})\right)\right] = \mathbb{E}\left[\int_{\Gamma(\mathbf{X}) \triangle \Gamma^*_\ell(\mathbf{X})} dy\right] \ .$$

In some sense, we note that the measure $\mathcal{H}$ provides a stronger guarantee than the excess risk since $\mathcal{E}_\ell(\Gamma) \leq C_2 \mathcal{H}(\Gamma)$ where C_2 is a positive constant which depends on σ_0. Besides, $\mathcal{H}(\Gamma) = 0$ implies that $\Gamma = \Gamma^*_\ell$ while this property does not necessarily hold for the excess risk.

[1] When $t = 0$, we have $G(t) = +\infty$ and then we will use the convention $G^{-1}(+\infty) = 0$.

3 Data-driven procedure

In this section, we provide a general data-driven procedure to estimate the optimal predictor Γ_ℓ^*. Two key features are expected from the resulting empirical prediction interval. The expected length should be of order ℓ while keeping its error rate close to one obtained by the oracle predictor. The estimation procedure is presented in the Section 3.1, and its main properties are provided in Section 3.2. Finally, Section 3.3 is dedicated to the study of rates of convergence.

3.1 Empirical prediction interval

The result provided in Proposition 1 suggests that an empirical prediction interval can be obtained through the plug-in principle by considering estimators of the conditional density p and the parameter $\lambda_\ell^* = G^{-1}(\ell)$. From a theoretical perspective, this learning task requires two independent samples.

First, in order to build an estimator of the conditional density p, we estimate the functions f^* and σ. Hence, we exploit a labeled sample $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ and build based on it estimators $\hat{f}$ and $\tilde{\sigma}$ of these two functions by the means of any machine learning algorithm. However, to establish theoretical guarantees, we require that the estimator $\tilde{\sigma}$ satisfies similar assumption as Assumption 1. To this end, we consider a thresholded version of the estimator $\tilde{\sigma}$ denoted by $\hat{\sigma}$ and define for $s > 0$ as

$$\hat{\sigma}^2(\mathbf{x}) = \tilde{\sigma}^2(\mathbf{x})\mathbb{1}_{\{s^{-1} \le \tilde{\sigma}^2(\mathbf{x}) \le s\}} + s^{-1}\mathbb{1}_{\{\tilde{\sigma}^2(\mathbf{x}) < s^{-1}\}} + s\mathbb{1}_{\{\tilde{\sigma}^2(\mathbf{x}) > s\}} \ .$$

A straightforward consequence of the definition of $\hat{\sigma}$ is that $\frac{1}{s} \le \hat{\sigma}^2(\mathbf{x}) \le s$. Furthermore, if s satisfies $\frac{1}{s} \le \sigma_0^2 \le \sigma_1^2 \le s$, we have for all $\mathbf{x}$

$$\left|\hat{\sigma}^2(\mathbf{x}) - \sigma^2(\mathbf{x})\right| \le \left|\tilde{\sigma}^2(\mathbf{x}) - \sigma^2(\mathbf{x})\right| \ ,$$

Hence consistency of $\tilde{\sigma}^2$ would imply the consistency of $\hat{\sigma}^2$.

Based on $\hat{f}$ and $\hat{\sigma}$, an estimator $\tilde{p}$ of the conditional density p naturally derives and can be written for all $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ as

$$\tilde{p}(y|\mathbf{x}) = \frac{1}{\sqrt{2\pi}\hat{\sigma}(\mathbf{x})} \exp\left(-\frac{\left(y - \hat{f}(\mathbf{x})\right)^2}{2\hat{\sigma}^2(\mathbf{x})}\right) \ .$$

The second step is devoted to the estimation of the parameter λ_ℓ^* and requires an *unlabeled* sample $\mathcal{D}_N = \{\mathbf{X}_{n+1}, \ldots, \mathbf{X}_{n+N}\}$ which consists of i.i.d. observations of $\mathbf{X}$ and is independent of $\mathcal{D}_n$. Since λ_ℓ^* depends on the function G, it is suitable to consider the empirical counterpart of the function G, that we build based on $\hat{p}$ and define for all $t \in [0, 1]$ as

$$\tilde{G}(t) = \int_{\mathbb{R}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{\hat{p}(y|\mathbf{X}_{n+i}) > t\}} \mathrm{d}y \ .$$

As a result, the empirical prediction interval is defined[2] point-wise as

$$\tilde{\Gamma}(\mathbf{x}) = \{y \in \mathbb{R} : \ \tilde{p}(y|\mathbf{x}) \ge \tilde{G}^{-1}(\ell)\} \ .$$

The predictor $\tilde{\Gamma}$ is very natural but has a few limitations: i) because Y is unbounded, the study of the theoretical properties of the estimator $\tilde{\Gamma}$ might be difficult; ii) in addition, establishing a theoretical analysis on $\tilde{\Gamma}$ involves similar assumption to Assumption 3 for $\tilde{G}$. More precisely, it requires that conditional on $\mathcal{D}_n$ the cumulative distribution of $\tilde{p}(y|\mathbf{X})$ is atomless; iii) furthermore, the above expression of $\tilde{\Gamma}(\mathbf{x})$ is explicit but relies on computing an integral in order to evaluate the function $\tilde{G}$. This integral should be approximated. To circumvent all these issues, we consider the following modifications of the initial estimator $\tilde{\Gamma}$.

For i) − Thresholding. Let $s > 0$, we consider a thresholded version of p given by

$$\hat{p}(y|\mathbf{x}) = \frac{1}{\sqrt{2\pi}\hat{\sigma}(\mathbf{x})} \exp\left(-\frac{\left(y - \hat{f}(\mathbf{x})\right)^2}{2\hat{\sigma}^2(\mathbf{x})}\right) \mathbb{1}_{\{|y| \le s\}} \ . \tag{4}$$

[2]Here again, we use the convention $\tilde{G}^{-1}(+\infty) = 0$.

For ii) – Randomization. To ensure the continuity of the conditional C.D.F. of $\hat{p}(y|\mathbf{X})$ for $y \in [-s, s]$, we introduce a random perturbation ζ distributed according to a Uniform distribution on $[0, u]$, for $u > 0$ and independent of $(\mathbf{X}, Y)$. We then define the randomized version of $\hat{p}$ as

$$\hat{p}(y|\mathbf{X}, \zeta) = \hat{p}(y|\mathbf{X}) + \zeta \mathbb{1}_{\{|y| \leq s\}} \ . \tag{5}$$

For iii) – Discretization. To approximate $\tilde{G}$, we simply consider the Riemann sum based on the regular grid $\mathcal{G} = \{y_1, \ldots, y_M\}$ of $[-s, s]$ for some $M \geq 1$. To this end, we introduce $(\zeta_1, \ldots, \zeta_N)$ i.i.d. copies of ζ and then define

$$\hat{G}(t) = \frac{2s}{MN} \sum_{k=1}^{M} \sum_{i=1}^{N} \mathbb{1}_{\{\hat{p}(y_k|\mathbf{X}_{n+i}, \zeta_i) > t\}} \ .$$

Finally, the resulting empirical prediction interval writes as

$$\hat{\Gamma}(\mathbf{X}, \zeta) = \{y \in \mathbb{R} : \ \hat{p}(y|\mathbf{X}, \zeta) \geq \hat{G}^{-1}(\ell)\} \ . \tag{6}$$

3.2 Theoretical guarantees

In this section, we provide the main properties of the empirical prediction interval $\hat{\Gamma}$. We first illustrate that the prediction interval $\hat{\Gamma}$ has an expected length equal to the requested value ℓ. This is one of the main striking feature of our data-driven procedure.

Proposition 3. *Assume that $M > 4\sqrt{N}$, then*

$$\mathbb{E}\left[\left|\mathcal{L}(\hat{\Gamma}) - \ell\right|\right] \leq C \frac{s}{\sqrt{N}} \ ,$$

where $C > 0$ is an absolute constant.

The above result states that our methodology is able to produce a prediction interval with an expected length ℓ, irrespectively of the distribution of the data and of whether or not we have build accurate estimates for f^* and σ. Importantly, Proposition 3 holds even if $(\mathbf{X}, Y)$ does not satisfy Equation (1). From this perspective the control on the expected length of the produced prediction interval is *distribution-free*. Notice in particular that the stated bound depends only on the parameter s which should be specified by the practitioner (this choice is discussed later) and on the number N of unlabeled data. In some semi-supervised applications, the amount of these data can be very large so that we can get a good approximation of the marginal distribution $\mathbb{P}_{\mathbf{X}}$ and then we can expect a good control of the expected length almost for free. Let us also add that Proposition 3 is a fundamental step to show the following bound on the excess risk:

Proposition 4. *Let Assumption 3 be satisfied. For $M > 4\sqrt{N}$, we have*

$$\mathbb{E}\left[\mathcal{E}_\ell\left(\hat{\Gamma}\right)\right] \leq C\left(\mathbb{E}\left[\int_{\mathbb{R}} |\hat{p}(y|\mathbf{X}) - p(y|\mathbf{X})| \, \mathrm{d}y\right] + su + \frac{s}{\sqrt{N}}\right) \ ,$$

where $C > 0$ is an absolute constant.

The above result shows that the excess-risk of $\hat{\Gamma}$ is mainly controlled by the L_1-risk of the estimator of the conditional density. The residual terms are related to the randomization on the one hand and to the control of the expected length of $\hat{\Gamma}$, given in Proposition 3, on the other hand. Proposition 4 is an intermediate step to establish consistency of the proposed prediction interval as well as to build explicit rates of convergence for the excess-risk of $\hat{\Gamma}$. This is the purpose of the next paragraph and Section 3.3 respectively.

Consistency result. Proposition 4 shows that the consistency of $\hat{\Gamma}$ with respect to the excess-risk relies to the consistency of the estimator $\hat{p}(y|\mathbf{x})$. In view of Equation (4), it is clear that the performance of $\hat{p}$ is directly linked to the statistical properties of $\hat{f}$ and $\hat{\sigma}$. More precisely, we obtain the following result.

Theorem 1. *Let Assumptions 1, 2, and 3. Consider* $s = \log(\min(n, N))$, $M > 4\sqrt{N}$, *and* $u = u_n \to 0$. *Assume that*

$$\sqrt{s}\mathbb{E}\left[(\hat{f}(X) - f^*(X))^2\right] \to 0, \text{ and } s^{5/2}\mathbb{E}\left[|\hat{\sigma}^2(X) - \sigma(X)|\right] \to 0 \ ,$$

then the following holds

$$\mathbb{E}\left[\mathcal{E}_\ell\left(\hat{\Gamma}\right)\right] \leq C_2\mathbb{E}\left[\mathcal{H}\left(\hat{\Gamma}\right)\right] \to 0 \ .$$

Let us make several comments on this theorem. First, under suitable assumptions, both excess-risk and expected symmetric difference of $\hat{\Gamma}$ converge to 0. Notably, since $\mathbb{E}\left[\mathcal{E}_\ell\left(\hat{\Gamma}\right)\right] \leq C_2\mathbb{E}\left[\mathcal{H}\left(\hat{\Gamma}\right)\right]$, consistency *w.r.t.* the expected symmetric difference implies consistency *w.r.t.* the excess-risk. From this perspective, symmetric difference control is a more difficult problem that excess-risk control. In particular, $\mathbb{E}\left[\mathcal{H}\left(\hat{\Gamma}\right)\right] \to 0$ indicates that $\hat{\Gamma} = \Gamma_\ell^*$ asymptotically. Another aspect that need to be discussed is the assumptions that are requested for the proof of Theorem 1. More specifically, consistency of $\hat{f}$, and $\hat{\sigma}^2$ are naturally required to ensure that $\hat{p}$ is a consistent estimator of p. In particular, convergence of $\hat{f}$ and $\hat{\sigma}$ can be made possible by several learning algorithms such as kernel methods, local polynomials, regularized least-squares among many others.

3.3 Rates of convergence

Theorem 1 establishes the consistency of the prediction interval $\hat{\Gamma}$ under mild assumptions. In this section, we focus on rates of convergence. More structural assumptions are then required. We borrow conditions from [7] introduced in the framework of regression with abstention. We assume that $\mathbf{X}$ belongs to a compact $\mathcal{C}$, and we consider the following assumptions.

Assumption 4 (Regularity)**.** *The functions* f^* *and* σ^2 *are Lipschitz.*

Assumption 5 (Strong density assumption)**.** *The marginal distribution* $\mathbb{P}_\mathbf{X}$ *satisfies the strong density assumption*

- $\mathbb{P}_\mathbf{X}$ *is supported on a compact regular set* $\mathcal{C} \subset \mathbb{R}^d$,
- $\mathbb{P}_\mathbf{X}$ *admits a density* μ *w.r.t. to the Lebesgue measure such that* $0 < \mu_{\min} \leq \mu(\mathbf{x}) \leq \mu_{\max} < \infty$, *for all* $\mathbf{x} \in \mathcal{C}$.

Assumption 6 (α-Margin assumption)**.** *We say that* $p(\cdot|X)$ *satisfies Margin assumption with parameter* $\alpha \geq 0$ *at level* λ_ℓ *with respect to* $\mathbb{P}_X$ *if there exist constants* $c_0 > 0$ *and* $t_0 > 0$ *such that for all* $0 < t \leq t_0$,

$$\int_\mathbb{R} \mathbb{P}_X\left(|p(y|\mathbf{X}) - \lambda_\ell| \leq t\right) dy \leq c_0 t^\alpha \ .$$

The above first two assumptions are rather classical when we deal with rates of convergence in non-parametric statistic. We refer the reader to the book [8] for a more detail discussion. In addition, Assumption 6, also known as Tsybakov noise condition [15], has been introduced in the binary classification setting to get fast rates of convergence [1]. In our setting, we notice that the Tsybakov noise condition is required around the threshold λ_ℓ. Moreover, since we extend this assumption to the case of regression, we need to ingrate it *w.r.t.* $y \in \mathbb{R}$. Based on the above conditions, we can establish the following result.

Proposition 5. *Let Assumptions 1, 4, 5, and 6 be satisfied. For* $s = \log(\min(n, N))$, *and* $M > 4\sqrt{N}$, *we have that*

$$\mathbb{E}\left[\mathcal{E}_\ell\left(\hat{\Gamma}\right)\right] \leq C\left(\mathbb{E}\left[\left(\sup_{(x,y)\in\mathcal{C}\times[-s,s]} |\hat{p}(y|x) - p(y|x)|\right)^{1+\alpha}\right] + \frac{1}{\min(n, N)^{1+\alpha}} + u^{1+\alpha} + \frac{\log(N)}{\sqrt{N}}\right),$$

where $C > 0$ *is a constant which depends on* f^*, σ^2, c_0, α, *and* $\mathcal{C}$.

As compared to the upper-bound that we get in Proposition 4, the bound here is better because of the exponent $1 + \alpha$ against 1. However, it is obtained under stronger assumptions.

Estimators of regression and variance function. The framework that we have described so far is quite general and allows to use any off-the-shelf machine learning algorithms to estimate the regression and the variance functions. In what follows, we propose a more concrete illustration of our approach by considering empirical prediction intervals $\hat{\Gamma}$ where both regression and variance functions are estimated with the kNN algorithm. Hereafter, we briefly recall the definition of the estimators that are based on the $l$abeled sample $\mathcal{D}_n$. For any $\mathbf{x} \in \mathbb{R}^d$, we denote by $(\mathbf{X}_{(i,n)}(\mathbf{x}), Y_{(i,n)}(\mathbf{x})), i = 1, \dots n$ the reordered data according to the ℓ_2 distance in $\mathbb{R}^d$, meaning that

$$\|\mathbf{X}_{(i,n)}(\mathbf{x}) - \mathbf{x}\| < \|\mathbf{X}_{(j,n)}(\mathbf{x}) - \mathbf{x}\| \ ,$$

for all $i < j$ in $\{1, \dots, n\}$. For simplicity, we assume that ties occur with probability 0. Let $k = k_n$ be an integer. The kNN estimator of f^* and σ^2 are then defined, for all $\mathbf{x} \in \mathbb{R}^d$, as follows

$$\hat{f}(\mathbf{x}) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{(i,n)}(\mathbf{x}) \ \text{ and } \ \tilde{\sigma}^2(\mathbf{x}) = \frac{1}{k_n} \sum_{i=1}^{k_n} \left(Y_{(i,n)}(\mathbf{x}) - \hat{f}(X_{(i,n)}(\mathbf{x})) \right)^2 \ . \tag{7}$$

The properties of these estimator are provided in [8] for the regression function and in [7] for the variance function. In particular, the authors in [7] establish rates of convergence *w.r.t.* the sup-norm for the estimator $\hat{\sigma}$.

Rates of convergence. The next result, which is an adaption of Proposition 3.1 in [5], is useful to derive upper-bound on the measure of risk $\mathcal{H}$ of $\hat{\Gamma}$ thanks to a control on the excess-risk.

Proposition 6. *Let Assumptions 6 be satisfied. There exists an absolute constant $C_3 > 0$ such that*

$$\mathbb{E}\left[\mathcal{H}(\hat{\Gamma})\right] \leq C_3 \left(\mathbb{E}\left[\mathcal{E}_\ell\left(\hat{\Gamma}\right)\right] \right)^{\alpha/\alpha+1} \ .$$

Importantly, this proposition, together with the inequality $\mathcal{E}_\ell(\Gamma) \leq C_2 \mathcal{H}(\Gamma)$ for all Γ, shows that under appropriate regularity condition consistency of $\hat{\Gamma}$ *w.r.t.* the distance $\mathcal{H}$ and the excess-risk are equivalent. The only difference is in the rates of convergence. The above result highlights the link between them under Assumption 6. In particular, we only have to establish rates of convergence *w.r.t.* $\mathcal{E}$. Let us introduce the following notation. When $a \propto b$, it means that the quantities a and b are equal up to a constant. Moreover $\lesssim_{\log(n)}$ says that the inequality holds up to some constants and logarithmic factors. Now, we state the main result of this section.

Theorem 2. *Let Assumptions 1 and 4-6 be satisfied. Let $k_n \propto n^{-2/d+2}$, $s = \log(\min(n, N))$, $M > 4\sqrt{N}$, and $u_n = \frac{1}{n}$. The following holds*

$$\mathbb{E}\left[\mathcal{E}_\ell(\hat{\Gamma})\right] \lesssim_{\log(n)} n^{-(1+\alpha)/(d+2)} + \min(n, N)^{-(1+\alpha)} + N^{-1/2} \ .$$

Several comments can be made from the above result. The first term is the classical nonparametric fast rate of convergence for the excess-risk under the Margin assumption and the Lipschitzness of the regression function. The last two terms that are related to the problem of PI estimation have different behavior according to the interplay between n and N. In particular, as soon as $N \leq n$, the limiting term is $N^{-1/2}$ and the rate becomes slow if $n^{-(1+\alpha)/(d+2)}$ goes faster to 0. On the other hand, if the number of unlabeled data N is large with $N \gg n^{1+\alpha}$ we recover the fast rate of convergence $n^{-(1+\alpha)/(d+2)}$. Between these two extremes, $N^{-1/2}$ can still be the limiting term. However, we hope that in our semi-supervised setting, enough data are available to make this term negligible as compared to the others.

4 Extension and other approach

In this section, we discuss some points beyond the considered framework in this paper. The extension of our results to other regression models is presented in Section 4.1. Another approach to build prediction interval based on the control of the expected error rate [12] is described in Section 4.2. In particular, we exhibit the main differences with our considered procedure.

4.1 Beyond Gaussian setting

In the present work, we study prediction intervals under expected length constraint in the heteroscedastic Gaussian regression setup. The appealing aspect of this framework lies in the form of the optimal predictor

$$\Gamma_\ell^*(\mathbf{X}) = \{y \in \mathbb{R} :\ p(y|\mathbf{X}) \geq \lambda_\ell^*\} \ , \tag{8}$$

with $\lambda_\ell^* = G^{-1}(\ell)$ and $G(t) := \int_{\mathbb{R}} \mathbb{P}(p(y|\mathbf{X}) \geq t)dy$. Furthermore, the density $p(y|\mathbf{X})$ has an explicit expression that exclusively depends on the regression and the conditional variance functions f and σ. Therefore, our proposed algorithm only involves estimators of f and σ to estimate the conditional density p. In particular, we do not consider any general procedure for density estimation.

In this paragraph, we discuss possible extensions outside the Gaussian framework but still considering the regression framework $Y = f^*(\mathbf{X}) + \sigma(\mathbf{X})\,\varepsilon$. In order to make sure that the optimal predictor is well defined, we require the following assumption.

Assumption 7. *We assume that the variable Y given $\mathbf{X}$ has density $p(\cdot|\mathbf{X})$.*

If we do not assume that $Y|\mathbf{X}$ belongs to a given family of distribution, the characterization of the prediction interval (8) still holds but the expression of the conditional density can not be simplify. Therefore, a data-driven predictor, based on the plug-in principle, must rely on estimates $\hat{p}(\cdot|\mathbf{x})$ of the conditional density $p(\cdot|\mathbf{x})$. The way to build the estimator $\hat{\Gamma}$ does not differ from the Gaussian case ones $\hat{p}$ is obtained (see Section 3). From the theoretical perspective, general properties such as Propositions 1 and 2 still hold and the question here is to investigate consistency results of the algorithm $\hat{\Gamma}$. The control on the expected length of the prediction interval $\mathbb{E}\left[\left|\mathcal{L}(\hat{\Gamma}) - \ell\right|\right] \leq C\frac{s}{\sqrt{N}}$ given in Proposition 3 is also still valid since this result is distribution-free. On the other hand, consistency for the excess-risk requires conditions. In the case where Y is bounded, if the estimator of the conditional probabilities is such that $\mathbb{E}\left[\int_{\mathbb{R}} |\hat{p}(y|\mathbf{X}) - p(y|\mathbf{X})|dy\right] \underset{n\to+\infty}{\longrightarrow} 0$, we can establish under Assumptions 1, 2, 3, and 7 that

$$\mathbb{E}\left[\mathcal{H}\left(\hat{\Gamma}\right)\right] \underset{n,N\to+\infty}{\longrightarrow} 0 \ .$$

Essentially, this result says that the estimation procedure that we study in this paper extends beyond the Gaussian setting. In particular, we still manage to get consistency for bounded random variable. It is worth mentioning that consistency might also be obtained as soon as $Y|\mathbf{X}$ is sub-Gaussian. Then our method is statistically valid for general settings.

4.2 Prediction interval under expected coverage constraint

In this section, we present the approach which focuses on the construction of prediction interval under expected coverage. This method consists in minimizing the length of the prediction interval under a constraint on its expected error rate. This approach is for instance studied in [12].

More precisely, let $\beta > 0$. We consider the following problem

$$\Gamma_\beta^* = \arg\min_{\mathbb{P}(Y\notin\Gamma(\mathbf{X})\leq\beta} \mathbb{E}\left[L(\Gamma(\mathbf{X}))\right] \ .$$

Under Assumptions 3 and 7 we can derive an expression of Γ_β^* based on thresholding of the conditional densities:

$$\Gamma_\beta^* = \{y \in \mathbb{R},\ \ p(y|\mathbf{x}) \geq t_\beta\} \ ,$$

with t_β defined as solution of

$$\mathbb{E}\left[\mathbb{1}_{\{p(Y|\mathbf{X})\geq t_\beta\}}\right] = \int_{\mathbb{R}} \mathbb{1}_{\{p(y|\mathbf{x})\geq t_\beta\}} p(y|\mathbf{x})\,\mathrm{d}y = 1-\beta \ .$$

Therefore, from the above equation, we deduce that

$$H^{-1}(t_\beta) = 1-\beta \ ,$$

where $H(t) = \mathbb{E}\left[\mathbb{1}_{\{p(Y|\mathbf{X})\geq t\}}\right]$. Similarly to the procedure described in Section 3.1, we are able to provide a randomized prediction interval $\hat{\Gamma}_\beta$ based on the estimator $\hat{p}$. We point out that an important difference between the construction of estimators $\hat{\Gamma}_\beta$ and $\hat{\Gamma}$ is the estimation of the function H. Indeed, this step require a *labeled* and not an *unlabeled* dataset, but do not request the discretization step. More formally, considering a *labeled* dataset $\mathcal{D}_K = \{(\mathbf{X}_i, Y_i),\ i = 1, \ldots, K\}$, and $(\zeta_1, \ldots, \zeta_K)$ the vector of perturbation, the estimator $\hat{H}$ of the function H is defined for each $t > 0$, as follows

$$\hat{H}(t) = \frac{1}{K}\sum_{i=1}^{K} \mathbb{1}_{\{p(Y_i|\mathbf{X}_i,\zeta_i)\geq t\}} \ .$$

Although a theoretical comparison with our proposed method is not our purpose, using similar arguments as in [12], we can establish the consistency of $\hat{\Gamma}_\beta$ under same assumptions as in Theorem 1.

$$\mathbb{E}\left[\mathcal{H}\left(\hat{\Gamma}_\beta\right)\right] \to 0 \ .$$

In Section 5, we focus on a comparison between our method and the expected coverage approach from a numerical perspective.

5 Numerical experiments

This section is devoted to a numerical study of the performance of our procedure. More precisely, we analyze our approach on synthetic data in Section 5.1 and provide a comparison with the expected coverage approach described in Section 5.2.

5.1 Simulation study

We illustrate the performance of our procedure on the following model

$$Y = \exp(-\|\mathbf{X}\|_2) + \frac{d\varepsilon}{2 + 4\|\mathbf{X}\|_2}, \ \ \mathbf{X} \in \mathbb{R}^d \ , \tag{9}$$

where $\mathbf{X} = (X^1, \ldots, X^d)$ is such that for $j = 1, \ldots, d$, the X^j are i.i.d. simulated according to a Uniform distribution on $[0, 1]$ and are independent from $\varepsilon \sim \mathcal{N}(0, 1)$. Note that the considered model satisfies Equation 1, and that Assumptions 1, and 2 are fulfilled.

For our numerical experiments, we choose reasonable dimensions of the features space $d \in \{1, 5\}$. Before going further in our investigations, we display the boxplots of the output variable Y in Figure 1. We see that the range of values of Y is much larger for $d = 5$ and is included in $[-5, 5]$ for both $d = 1, 5$. Besides, we chose to focus on $\ell \in \{0.1, 0.5, 1, 2\}$ which seems to be relevant values according to Figure 1 in order to still get interpretation of the output. For $\ell \in \{0.1, 0.5, 1, 2\}$, we provide the evaluation of the expected length and the error rate for the oracle prediction set Γ^*_ℓ. To this end, we repeat 100 times the following scheme.

i) estimate λ^*_ℓ from an *unlabeled* dataset of size $N = 1000$ on a regular grid of size $M = 1000$ of the interval $[-5, 5]$;

ii) derive the resulting prediction interval on the same grid over a test set of size $T = 1000$;

iii) based on the test set, compute the expected length and the error rate.

From these repetitions, we compute the mean and standard deviation of the estimates. The obtained results are provided in Table 1.

Simulation scheme. To assess the performance of our procedure, we consider the following scheme. For $d \in \{1, 5\}$ and $\ell \in \{0.1, 0.5, 1, 2\}$, we repeat 100 the following steps.

i) estimate f^* and σ^2 from a training test of size $n = 500$. We consider the residual-based method [9]. The estimation of f^* and σ^2 relies on the random forests algorithm from `python` library `sklearn`. We also choose $u = 10^{-5}$ for the parameter of the perturbation ζ (see Eq. (5));

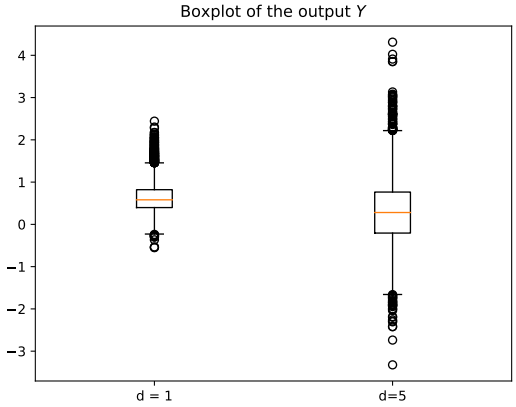

Figure 1: Boxplot of the output Y for $d = 1, 5$

	Expected length		Error rate	
ℓ	$d = 1$	$d = 5$	$d = 1$	$d = 5$
0.1	0.1 (0.01)	0.1 (0.01)	0.81 (0.01)	0.94 (0.01)
0.5	0.49 (0.01)	0.49 (0.01)	0.34 (0.01)	0.71 (0.01)
1	0.99 (0.01)	0.99 (0.01)	0.07 (0.01)	0.48 (0.01)
2	1.99 (0.03)	1.99 (0.01)	0.00 (0.00)	0.17 (0.01)

Table 1: Performance of the Oracle PI for $\ell \in \{0.1, 0.5, 1, 2\}$.

ii) compute $\hat{G}^{-1}(\ell)$ using an *unlabeled* dataset of size $N = 100$ on a regular grid of size $M = 100$ of the interval $[-s, s]$, where $s = \max(-\min(Y_{train}); \max(Y_{train})$,

iii) derive the resulting prediction interval on a regular grid of size 1000 of $[-s, s]$ over a test set of size $T = 1000$;

iv) based on the test set, compute the expected length and the error rate.

From these experiments, we compute the empirical means and standard deviations expected length and the error rate. The results are provided in Table 2. A visual description of the behavior of our PI is also given in Figure 2.

Notice that the value of s that we consider here is different from the one suggested by the theory in Theorem 2. This is a minor point. The parameter s in the theory is set such that most of the labels lie in $[-s, s]$ with high probability. This happens when n and M grow since $s = \log(\min(n, N))$. Our choice in practice ensures that this property holds regardless the values of n and N.

Results. Two conclusions can be made from this first numerical study. First Tables 1 and 2 highlight how effective our method is in producing PI with (almost) exactly the right length. This is an important point and suggests that our strategy succeeds to enforce the constraint on the length prescribed by the optimization problem. Second, let us focus on a comparison between Γ^*_ℓ, the oracle PI, and its empirical counterpart $\hat{\Gamma}$. Table 1 and Table 2 show how close are the performance of these two PI both in terms of

	Expected length		Error rate	
ℓ	$d = 1$	$d = 5$	$d = 1$	$d = 5$
0.1	0.1 (0.01)	0.1 (0.02)	0.81 (0.02)	0.94 (0.01)
0.5	0.50 (0.01)	0.50 (0.02)	0.34 (0.02)	0.72 (0.02)
1	1.00 (0.02)	1.00 (0.02)	0.07 (0.01)	0.48 (0.01)
2	2.01 (0.06)	2.01 (0.01)	0.00 (0.00)	0.17 (0.01)

Table 2: Performance of $\hat{\Gamma}$ for $\ell \in \{0.1, 0.5, 1, 2\}$.

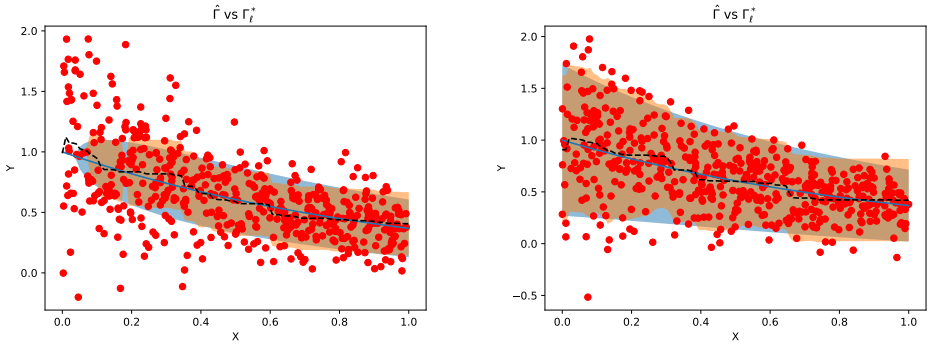

Figure 2: Visual description of the empirical PI $\hat{\Gamma}$ and its oracle counterpart Γ^*_ℓ, with $\ell = 0.5$ on the left and $\ell = 1$ on the right for $d = 1$. The scatter plot of data is displayed and the graph of both regression function f^* and estimator $\hat{f}$ is represented (solid line for f^*, dashed line for $\hat{f}$. The oracle PI Γ^*_ℓ (empirical PI $\hat{\Gamma}$, respectively) is given in blue (orange, respectively).

expected length and of error rate. Interestingly, the performance of $\hat{\Gamma}$ is obtained with a moderate size N of the unlabeled sample that is used to estimate the threshold. These results also suggest that $n = 500$ is enough to have good estimations of the regression and variance functions. The closeness between Γ^*_ℓ and $\hat{\Gamma}$ is also illustrated in Figure 2.

5.2 Numerical comparison with expected coverage approach

In this section, we numerically compare our procedure to the approach that constraint the expected coverage described in Section 4.2. We consider the model defined in Equation 9 with $d = 5$ and focus on the estimation of Γ^*_ℓ for $\ell = 2$. With this expected length, the oracle predictor Γ^*_ℓ reaches an error rate of $\beta = 0.17$. Therefore, for this learning task, we are able to provide empirical PI for both approaches. That is to say, we compute $\hat{\Gamma}$ with $\ell = 2$ as expected length and $\hat{\Gamma}_\beta$ with $\beta = 0.17$ as expected error. In order to get a fair comparison of the methods, we repeat 20 times the following steps. For both approaches, we use a training set of size $n = 500$ to estimate the density p and we estimate the threshold of the considered procedure with a dataset of size $N \in \{10, 30, 50, 70, 100, 150, 200, 500, 1000\}$. Finally, we compute the expected length and error rate of both empirical PI over a test set of size $T = 1000$. From these repetitions, we compute empirical means and standard deviations. The results are displayed in Figure 3.

As expected, in average, both methods behaves similarly. However there are important differences in favor of our approach. First, the convergence of our method is much faster to the mean value both for the expected length and the error rate. We notice that $N = 10$ is already enough for our method while more than 500 samples are needed for the method that focus on the coverage as constraint. Second, it seems that our construction is much more stable, in particular for length calibration. It illustrates the efficiency of our procedure to build prediction interval with the right expected length.

The two approaches are definitively not comparable in terms of objectives. Indeed, if we are really focused on constraining the error rate, then the length constraint appears (at first sight) sub-optimal and vice versa if we ask for interpretable outputs. However, our numerical analysis clearly suggests that our methodology is more stable: it induces a procedure with a lower variance.

6 Conclusion

In this paper, we provide a general methodology to build *prediction intervals with controlled expected length* in the Gaussian regression. Our proposed algorithm is very effective in controlling the expected length of the output and then ensure the interpretability of the outcome. The theoretical analysis indicates that our method mimics the optimal rule *w.r.t.* the expected length and, under appropriate properties on the base estimators of the regression function, it is also efficient *w.r.t.* the symmetric difference distance and the excess-risk. Furthermore, a numerical study supports our theoretical results.

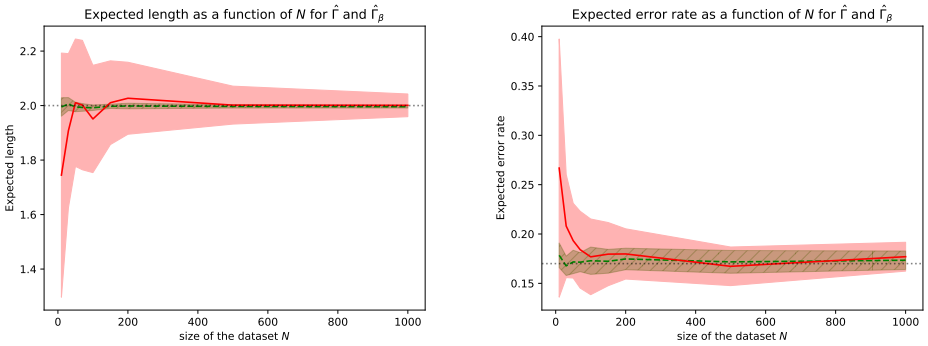

Figure 3: Comparison between $\hat{\Gamma}$ and $\hat{\Gamma}_\beta$. We plot the expected length (on the left) and the expected coverage (on the right) as a function of N over 20 repetitions for $\hat{\Gamma}$ (dashed) and $\hat{\Gamma}_\beta$ (solid line in red). The true value of the parameter is given by the dotted line.

Notably, it highlights good stability properties as compared to prediction intervals that focus on expected coverage constraints.

Our numerical comparison to PI under expected coverage constraint additionally opens a very significant door to the use of our method. Because of the stability of our method, one may think to the following two-stage procedure to produce a PI with error rate β.

- *Step 1.* Build the PI with error rate β and evaluate its length $\tilde{\ell}$;
- *Step 2.* Build our PI with average length $\tilde{\ell}$.

While we do not expect a significant improvement in average, the resulting prediction interval might be more stable. This will be the purpose of future investigation.

On the other hand, inference in the high-dimensional setting is a crucial challenge with modern data. Several successful studies consider the Gaussian *homoscedastic* linear regression [13, 16, 14, 2]. An important direction for future research is to carry out PI i) for non Gaussian models; ii) and that can handle heteroscedastic model. Both of these questions have their applications in the high dimensional setting.

References

[1] J.-Y. Audibert and A. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35(2):608–633, 2007.

[2] A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.*, 81(2):608–650, 2014.

[3] R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *Ann. Statist.*, 41(2):802–837, 2013.

[4] S. Bobkov and M. Ledoux. One-dimensional empirical measures, order statistics and Kantorovich transport distances. *Memoirs of the Amer. Math. Soc.*, 2016.

[5] E. Chzhen, C. Denis, and M. Hebiri. Minimax semi-supervised set-valued approach to multi-class classification. *Bernoulli*, 2021.

[6] C. Denis and M. Hebiri. Confidence sets with expected sizes for multiclass classification. *J. Mach. Learn. Res.*, 18(102):1–28, 2017.

[7] C. Denis, M. Hebiri, and A. Zaoui. Regression with reject option and application to $k$nn. *NeurIPS*, 2020.

[8] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Ser. Statist. Springer-Verlag, New York, 2002.

[9] P. Hall and R.J. Carroll. Variance function estimation in regression: The mean effect of estimating the mean. *Journal of the Royal Statistical Society: Series B (Methodological)*, 51(1):3–14, 1989.

[10] L.H. Koopmans and Clifford Qualls. Fixed length confidence intervals for parameters of the normal distribution based on two-stage sampling procedures. *Rocky Mountain J. Math.*, 1(4):587–602, 1971.

[11] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[12] J. Lei and L. Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society, Series B*, 76(1):71–96, 2013.

[13] S. Lu, Y. Liu, L. Yin, and K. Zhang. Confidence intervals and regions for the lasso by using stochastic variational inequality techniques in optimization. *Journal of the Royal Statistical Society Series B*, 79(2):589–611, 2017.

[14] J. Minnier, L. Tian, and T. Cai. A perturbation method for inference on regularized regression estimates. *J. Amer. Statist. Assoc.*, 106(496):1371–1382, 2011.

[15] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[16] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 2014.

[17] A.W. van der Vaart. *Asymptotic statistics*. volume 3 of Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998.

[18] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer, New York, 2005.

[19] V. Vovk, I. Nouretdinov, and A. Gammerman. On-line preditive linear regression. *The Annals of Statistics*, 37:1566–1590, 2009.

Appendix

This appendix is devoted to the proof of our main results. The proofs related to Section 2 are provided in Section B, while Section C is devoted to the proofs of Section 3. Finally, Section A gathers useful results. In particular, we give rates of convergence for KNN estimates for both regression and variance function. Notice that in the whole appendix, C is a positive constant that may change from one line to another.

A Technical results

In this section, we provide some useful properties that are used for the proof of our main results

A.1 Technical lemmas

The first tool we introduce is a generalization of the classical inverse transform theorem [17, Lemma 21.1] to the continuous case. Let $a > 0$. We consider a random process $(Z_y)_{y\in[-a,a]}$ such that the function H defined by

$$H(t) = \frac{1}{2a}\int_{-a}^{a} \mathbb{P}(Z_y \geq t)dy \ ,$$

is continuous on $\mathbb{R}_+$.

Lemma 1. *Let T uniformly distributed on $[-a,a]$ and independent of $(Z_y)_{y\in[-a,a]}$. We consider the random variable Z_T and let U be distributed according to the uniform distribution on $[0,1]$. Then*

$$H(Z_T) \stackrel{\mathcal{L}}{=} U \ \text{ and } \ H^{-1}(U) \stackrel{\mathcal{L}}{=} Z_T \ .$$

Proof. For every $t \geq 0$, we have $\mathbb{P}(H(Z_T) \leq t) = \mathbb{P}(Z_T \geq H^{-1}(t))$. Denote by $d\mathbb{P}_T$ the marginal distribution of T. Since the variable T is independent of $(Z_y)_{y\in[-a,a]}$ and H is continuous, one gets

$$\begin{aligned}
\mathbb{P}(H(Z_T) \leq t) &= \int \mathbb{P}(Z_T \geq H^{-1}(t)|T=y))\ d\mathbb{P}_T(y) \\
&= \frac{1}{2a}\int_{-a}^{a} \mathbb{P}(Z_y \geq H^{-1}(t)|T=y))\ dy \\
&= \frac{1}{2a}\int_{-a}^{a} \mathbb{P}(Z_y \geq H^{-1}(t))\ dy = H(H^{-1}(t)) = t \ ,
\end{aligned}$$

and we deduce that $H(Z_T) \stackrel{\mathcal{L}}{=} U$. For the second point of the Lemma, we observe that

$$\mathbb{P}(H^{-1}(U) \leq t) \ = \ \mathbb{P}(U \geq H(t)) = \frac{1}{2a}\int_{-a}^{a} \mathbb{P}(Z_y \leq t)dy = \frac{1}{2a}\int_{-a}^{a} \mathbb{P}(Z_y \leq t|T=y))dy = \mathbb{P}(Z_T \leq t) \ .$$

□

A.2 Rates of convergence for K-NN estimators

In this section, we gather the results we use for K-NN estimators of both regression and variance function. The proof of this result is provided in [7].

Theorem 3. *Grants Assumptions 4, 5, for $k_n \propto n^{-2/d+2}$, and all $\alpha > 0$, the K-NN estimators defined in Equation (7) satisfy*

$$\begin{aligned}
\mathbb{E}\left[\left(\sup_{\mathbf{x}\in\mathcal{C}} |\hat{f}(\mathbf{x}) - f^*(\mathbf{x})|\right)^{1+\alpha}\right] &\leq C\log(n)^{1+\alpha} n^{-(1+\alpha)/(2+d)} \ , \\
\mathbb{E}\left[\left(\sup_{\mathbf{x}\in\mathcal{C}} |\hat{\sigma}^2(\mathbf{x}) - \sigma(\mathbf{x})|\right)^{1+\alpha}\right] &\leq C\log(n)^{1+\alpha} n^{-(1+\alpha)/(2+d)} \ .
\end{aligned}$$

B Proof of Section 2

In this section, we provide proofs related to the optimal confidence and to the excess-risk formula

Proof of Proposition 1. First, let us consider the Lagrangian of the optimization problem 2. It can be written as

$$H(\Gamma, \lambda) = \mathbb{P}\left(Y \notin \Gamma(\mathbf{X})\right) + \lambda \left(\mathbb{E}_{\mathbf{X}}[L(\Gamma(\mathbf{X})] - \ell\right) \ ,$$

where $\lambda \geq 0$ is a dual variable of the problem. Since,

$$\mathbb{P}\left(Y \in \Gamma(\mathbf{X})\right) = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}\left[\mathbb{1}_{\{Y \in \Gamma(\mathbf{X})\}} | \mathbf{X}\right]\right] = \mathbb{E}_{\mathbf{X}}\left[\int_{\mathbb{R}} p(y|\mathbf{X}) \mathbb{1}_{\{y \in \Gamma(\mathbf{X})\}} dy\right] \ ,$$

the Lagrangian reads as

$$H(\Gamma, \lambda) = 1 - \lambda\ell - \mathbb{E}_{\mathbf{X}}\left[\int_{\mathbb{R}} (p(y|\mathbf{X}) - \lambda) \mathbb{1}_{\{y \in \Gamma(\mathbf{X})\}} dy\right] \ . \tag{10}$$

Minimizing *w.r.t.* Γ leads to an optimal solution that can be written for all $\lambda \geq 0$ and all $\mathbf{x} \in \mathbb{R}^d$ as

$$\Gamma^*(\lambda, \mathbf{x}) = \{y \in \mathbb{R} : p(y|\mathbf{X}) \geq \lambda\} \ .$$

Injecting this value into 10 gives

$$H(\Gamma^*(\lambda, \mathbf{X}), \lambda) = 1 - \lambda\ell - \mathbb{E}_{\mathbf{X}}\left[\int_{\mathbb{R}} (p(y|\mathbf{X}) - \lambda)_+ \mathbb{1}_{\{y \in \Gamma^*(\lambda, \mathbf{X})\}} dy\right] \ ,$$

where $(\cdot)_+$ stands for the positive part. First order optimality conditions for convex non-smooth minimization problems implies $0 \in \partial H(\Gamma^*(\lambda_\ell^*, \mathbf{X}), \lambda_\ell^*)$ where ∂H is the sub-differential of H. Therefore, using the Fundamental Theorem of Calculus, we get $\mathbb{E}_{\mathbf{X}}\left[\int_{\mathbb{R}} \mathbb{1}_{\{y \in \Gamma^*(\lambda_\ell^*, \mathbf{X})\}} dy\right] = \ell$. But, using the above definition of Γ^* we can write by Fubini's theorem the left hand side term as $\mathbb{E}_{\mathbf{X}}\left[\int_{\mathbb{R}} \mathbb{1}_{\{y \in \Gamma^*(\lambda_\ell^*, \mathbf{X})\}} dy\right] = \int_{\mathbb{R}} \mathbb{P}\left((p(y|\mathbf{X}) \geq \lambda_\ell^*\right) dy = G(\lambda_\ell^*)$. We then conclude that $\lambda_\ell^* = G^{-1}(\ell)$. Notice that for this value, we have

$$\mathcal{L}(\Gamma^*) = \mathbb{E}_{\mathbf{X}}[L(\Gamma^*(\lambda_\ell^*, \mathbf{X})] = \mathbb{E}_{\mathbf{X}}\left[\int \mathbb{1}_{\{y \in \Gamma^*(\lambda_\ell^*, \mathbf{X})\}} dy\right] = G(\lambda_\ell^*) = \ell \ .$$

□

Proof of Proposition 2. Let $\ell \geq 0$. Considering a similar decomposition as in the proof of Proposition 1, we can write the error rate of a predictor Γ as

$$R_\ell(\Gamma) = 1 - \mathbb{E}_{\mathbf{X}}\left[\int_{\mathbb{R}} (p(y|\mathbf{X}) - \lambda_\ell^*) \mathbb{1}_{\{y \in \Gamma(\mathbf{X})\}} dy\right] \ . \tag{11}$$

Therefore, we deduce

$$\mathcal{E}_\ell\left(\Gamma\right) = \mathbb{E}_{\mathbf{X}}\left[\int_{\mathbb{R}} (p(y|\mathbf{X}) - \lambda_\ell^*) \left(\mathbb{1}_{\{y \in \Gamma_\ell^*(\mathbf{X})\}} - \mathbb{1}_{\{y \in \Gamma(\mathbf{X})\}}\right) dy\right] \ ,$$

and the result follows from the fact that $\mathbb{1}_{\{y \in \Gamma_\ell^*(\mathbf{X})\}} - \mathbb{1}_{\{y \in \Gamma(\mathbf{X})\}} = \operatorname{sgn}(p(y|\mathbf{X}) - \lambda_\ell^*)$ since we have the equality between events $\{y \in \Gamma_\ell^*(\mathbf{X})\} = \{p(y|\mathbf{X}) - \lambda_\ell^* \geq 0\}$, where $\operatorname{sgn} : \mathbb{R} \to \{-1, 1\}$ stands for the sign. □

C Proof of Section 3

We now consider the theoretical properties of the prediction interval $\hat{\Gamma}$. We first consider its expected length and then derive a finite sample bound on its excess-risk.

C.1 Length control

Proof of Proposition 3. To show this result, we need to introduce some pseudo-oracle predictor that has expected length ℓ. Let us then define the randomized predictor

$$\bar{\Gamma}(\mathbf{X},\zeta) = \{y \in \mathbb{R} :\ \hat{p}(y|\mathbf{X},\zeta) \geq \bar{G}^{-1}(\ell)\}\ , \tag{12}$$

where $\bar{G}(t) := \int_{\mathbb{R}} \mathbb{P}_{\mathbf{X},\zeta}(\hat{p}(y|\mathbf{X},\zeta) \geq t)dy$ for all $t > 0$. Here again, the property $\mathcal{L}(\bar{\Gamma}) := \mathbb{E}_{\mathbf{X},\zeta}\left[L\left(\bar{\Gamma}(\mathbf{X},\zeta)\right)\right] = \ell$ is due to the fact that the conditional on the data $\mathcal{D}_n$ the r.v. $\hat{p}(y|\mathbf{X},\zeta)$ has no atoms since it is randomized.

Let us now consider the purpose of the proposition. We need to bound $\mathbb{E}\left[\left|\mathcal{L}(\hat{\Gamma}) - \ell\right|\right]$. We can write

$$\begin{aligned}
\left|\mathcal{L}(\hat{\Gamma}) - \ell\right| = \left|\mathcal{L}(\hat{\Gamma}) - \mathcal{L}(\bar{\Gamma})\right| &= \left|\mathbb{E}\left[\int_{\mathbb{R}} \left(\mathbb{1}_{\{\hat{G}(\hat{p}(y|\mathbf{X},\zeta))\leq\ell\}} - \mathbb{1}_{\{\bar{G}(\hat{p}(y|\mathbf{X},\zeta))\leq\ell\}}\right) dy\right]\right| \\
&\leq \mathbb{E}\left[\int_{\mathbb{R}} \left|\mathbb{1}_{\{\hat{G}(\hat{p}(y|\mathbf{X},\zeta))\leq\ell\}} - \mathbb{1}_{\{\bar{G}(\hat{p}(y|\mathbf{X},\zeta))\leq\ell\}}\right| dy\right] \\
&\leq \mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\{|\hat{G}(\hat{p}(y|\mathbf{X},\zeta))-\bar{G}(\hat{p}(y|\mathbf{X},\zeta))|\geq|\bar{G}(\hat{p}(y|\mathbf{X},\zeta))-\ell|\}} dy\right] \\
&= \int_{\mathbb{R}} \mathbb{P}\left(|\hat{G}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq |\bar{G}(\hat{p}(y|\mathbf{X},\zeta)) - \ell|\right) dy\ ,
\end{aligned} \tag{13}$$

where we use Fubini's theorem at last. Now notice that the above integral is limited to the compact $[-s,s]$ since, this is the support of the function $\hat{p}(\cdot|\mathbf{x},z)$ for all $(\mathbf{x},z) \in \mathbb{R}^d \times [0,u]$. To bound this integral, we make use of the peeling technique of [1]. That is, we consider for $\delta > 0$ and $y \in [-s,s]$

$$\begin{aligned}
A_0(y) &= \left\{0 \leq |\bar{G}(\hat{p}(y|\mathbf{X},\zeta)) - \ell| \leq \delta\right\} \\
A_j(y) &= \left\{2^{j-1}\delta \leq |\bar{G}(\hat{p}(y|\mathbf{X},\zeta)) - \ell| \leq 2^j\delta\right\}, \qquad \text{for} \quad j \geq 1\ .
\end{aligned}$$

Since for $y \in [-s,s]$, the events $(A_j(y))_{j\geq 0}$ are mutually exclusive, we deduce

$$\begin{aligned}
&\int_{-s}^{s} \mathbb{P}\left(|\hat{G}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq |\bar{G}(\hat{p}(y|\mathbf{X},\zeta)) - \ell|\right)\ dy = \\
&\qquad \int_{-s}^{s} \sum_{j\geq 0} \mathbb{P}\left(|\hat{G}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq |\bar{G}(\hat{p}(y|\mathbf{X},\zeta)) - \ell|\ ,\ A_j(y)\right)\ dy\ .
\end{aligned} \tag{14}$$

Controlling this term relies on a bound on $\int_{-s}^{s} \mathbb{P}(A_j(y))dy$. It is clear that $0 \leq \bar{G}(t) = \int_{-s}^{s} \mathbb{P}_X(\hat{p}(y|\mathbf{X},\zeta) \geq t|\mathcal{D}_n)dy \leq 2s$ for all $t \in [0,1]$. We can apply Lemma 1 to say that $\bar{G}(Z_T)$ is uniformly distributed on $[0,2s]$ and then, for all $j \geq 0$ and $\delta > 0$, we deduce that

$$\begin{aligned}
\int_{-s}^{s} \mathbb{P}(A_j(y))dy &= 2s\frac{1}{2s}\int_{-s}^{s} \mathbb{P}\left(|\bar{G}(\hat{p}(y|\mathbf{X},\zeta)) - \ell| \leq 2^j\delta \mid \mathcal{D}_n\right) dy \\
&= 2s \times \mathbb{P}\left(|\bar{G}(Z_T) - \ell| \leq 2^j\delta|\mathcal{D}_n\right) \leq 2s\frac{2^{j+1}\delta}{2s} = 2^{j+1}\delta\ .
\end{aligned} \tag{15}$$

Next, let us consider (14). We observe that for all $j \geq 1$

$$\begin{aligned}
&\int_{-s}^{s} \mathbb{P}\left(|\hat{G}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq |\bar{G}(\hat{p}(y|\mathbf{X},\zeta)) - \ell|\ ,\ A_j(y)\right) dy \\
&\qquad \leq \int_{-s}^{s} \mathbb{P}\left(|\hat{G}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq 2^{j-1}\delta\ ,\ A_j(y)\right) dy \\
&\qquad \leq \int_{-s}^{s} \mathbb{E}_{(\mathcal{D}_n,\mathbf{X},\zeta)}\left[\mathbb{P}_{\mathcal{D}_N}\left(|\hat{G}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq 2^{j-1}\delta\right)\mathbb{1}_{A_j(y)}\right] dy.
\end{aligned} \tag{16}$$

In Section 3.1, we have presented the predictor $\hat{\Gamma}$ that relies on the function $\hat{G}$ which is discretized. On the other hand, $\bar{G}$ is not discretized. Because of this difference, it is convenient, in order to control (16),

to provide some additional notation. Let us define

$$\hat{\bar{G}}(t) := \frac{1}{N}\sum_{i=1}^{N}\int_{-s}^{s} \mathbb{1}_{\{\hat{p}(y|\mathbf{X}_{n+i},\zeta_i)\geq t\}} dy \ .$$

Then for all $y \in [-s,s]$, conditional on $(\mathcal{D}_n, \mathbf{X}, \zeta)$, the probability in Eq. (16) is bounded as follows

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D}_N}\left(|\hat{G}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq 2^{j-1}\delta\right) \leq \\
&\mathbb{P}_{\mathcal{D}_N}\left(|\hat{\bar{G}}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq 2^{j-1}\frac{\delta}{2}\right) + \quad \mathbb{P}_{\mathcal{D}_N}\left(|\hat{\bar{G}}(\hat{p}(y|\mathbf{X},\zeta)) - \hat{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq 2^{j-1}\tfrac{\delta}{2}\right) \quad (17)
\end{aligned}$$

These two last terms are treated in different ways. For the first one, we observe that for all $t \in [0,1]$

$$\begin{aligned}
|\hat{G}(t) - \hat{\bar{G}}(t)| &= \left|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{M}\left(\int_{y_k}^{y_{k+1}} \mathbb{1}_{\{\hat{p}(y|\mathbf{X}_{n+i},\zeta_i)\geq t\}} - \mathbb{1}_{\{\hat{p}(y_k|\mathbf{X}_{n+i},\zeta_i)\geq t\}}\right) dy\right| \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{M}\left(\int_{y_k}^{y_{k+1}} \left|\mathbb{1}_{\{\hat{p}(y|\mathbf{X}_{n+i},\zeta_i)\geq t\}} - \mathbb{1}_{\{\hat{p}(y_k|\mathbf{X}_{n+i},\zeta_i)\geq t\}}\right|\right) dy \ .
\end{aligned}$$

We recall that for all $|y| \leq s$, we have $\hat{p}(y|\mathbf{x},\zeta) = \hat{p}(y|\mathbf{x}) + \zeta$. Because, conditional on $\mathcal{D}_n$, the function $\hat{p}(\cdot|\mathbf{x})$ is a Gaussian density and since the perturbation ζ acts on each y in the same way, it turns out that the function $\hat{p}(.|\mathbf{x},\zeta)$ is continuously increasing and then decreasing with a maximum at $y = \hat{f}(\mathbf{x})$. Therefore, for any fixed t the indicators $\mathbb{1}_{\{\hat{p}(y|\mathbf{X}_{n+i},\zeta_i)\geq t\}}$ and $\mathbb{1}_{\{\hat{p}(y_k|\mathbf{X}_{n+i},\zeta_i)\geq t\}}$ differ at most in 2 intervals of the form $[y_k, y_{k+1}]$. Then we deduce that

$$|\hat{G}(t) - \hat{\bar{G}}(t)| \leq 2 \times \frac{2s}{M} \ .$$

Injecting this inequality to (17) gives

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D}_N}\left(|\hat{G}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq 2^{j-1}\delta\right) \leq \\
&\qquad\qquad \mathbb{P}_{\mathcal{D}_N}\left(|\hat{\bar{G}}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq 2^{j-1}\frac{\delta}{2}\right) + \mathbb{1}_{\{4s/M \geq 2^{j-2}\delta\}} \ . \quad (18)
\end{aligned}$$

Let us now consider the second term. Conditional on $(\mathcal{D}_n, \mathbf{X}, \zeta)$, the random variable $\hat{\bar{G}}(\hat{p}(y|\mathbf{X},\zeta))$ is an empirical mean of i.i.d. random variables of common mean $\bar{G}(\hat{p}(y|X,\zeta)) \in [0,2s]$, we deduce from Hoeffding's inequality that

$$\mathbb{P}_{\mathcal{D}_N}\left(|\hat{\bar{G}}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq 2^{j-2}\delta | \mathcal{D}_n, \mathbf{X}\right) \leq 2\exp\left(\frac{-N\delta^2 2^{2j-1}}{16s^2}\right) \ .$$

Therefore, from Inequalities (14), (15), (16), and (18) one gets for $\delta = \frac{4s}{\sqrt{N}}$ and $M > 4\sqrt{N}$

$$\begin{aligned}
&\int_{-s}^{s} \mathbb{P}\left(|\hat{G}(\hat{p}(y|\mathbf{X},\zeta)) - \bar{G}(\hat{p}(y|\mathbf{X},\zeta))| \geq |\bar{G}(\hat{p}(y|\mathbf{X},\zeta)) - \ell|\right) dy \\
&\qquad \leq \int_{-s}^{s} \mathbb{P}(A_0(y))dy + \sum_{j\geq 1} 2\exp\left(\frac{-N\delta^2 2^{2j-1}}{16s^2}\right)\int_{-s}^{s}\mathbb{P}(A_j(y))dy \\
&\qquad\qquad \leq 2\delta + \delta\sum_{j\geq 1} 2^{j+2}\exp\left(\frac{-N\delta^2 2^{2j-1}}{16s^2}\right) \leq \frac{Cs}{\sqrt{N}} \ . \quad (19)
\end{aligned}$$

□

C.2 Excess-risk control

Proof of Proposition 4. Throughout the proof, we denote $\bar{\lambda}_\ell := \bar{G}^{-1}(\ell)$, where $\bar{G}$ is defined in Equation (12). We start with the following decomposition.

$$\mathcal{E}_\ell\left(\hat{\Gamma}\right) = \mathcal{E}\left(\bar{\Gamma}\right) + \left(R_\ell(\hat{\Gamma}) - R_\ell(\bar{\Gamma})\right) \ . \tag{20}$$

For the second term of the *r.h.s.* in the above equation, thanks to Equation (11), we have that

$$R_\ell(\hat{\Gamma}) - R_\ell(\bar{\Gamma}) = \mathbb{E}_{\mathbf{X},\zeta}\left[\int_{\mathbb{R}} \left(p(y|\mathbf{X}) - \lambda_\ell\right)\left(\mathbb{1}_{\{y\in\bar{\Gamma}(\mathbf{X},\zeta)\}} - \mathbb{1}_{\{y\in\hat{\Gamma}(\mathbf{X},\zeta)\}}\right) dy\right] \ .$$

From Assumption 1, we have that $|p(y|\mathbf{X}) - \lambda_\ell|$ is bounded by $C_1 > 0$ which depends on σ_0. Hence, we deduce that

$$\mathbb{E}\left[\left|R_\ell(\hat{\Gamma}) - R_\ell(\bar{\Gamma})\right|\right] \leq C_1 \mathbb{E}\left[\int_{\mathbb{R}} \left|\mathbb{1}_{\{y\in\bar{\Gamma}(\mathbf{X},\zeta)\}} - \mathbb{1}_{\{y\in\hat{\Gamma}(\mathbf{X},\zeta)\}}\right| dy\right] \ .$$

This last inequality can be rewritten as

$$\mathbb{E}\left[\left|R_\ell(\hat{\Gamma}) - R_\ell(\bar{\Gamma})\right|\right] \leq C_1 \mathbb{E}\left[\int_{\mathbb{R}} \left|\mathbb{1}_{\{\hat{G}(\hat{p}(y|\mathbf{X},\zeta))\leq\ell\}} - \mathbb{1}_{\{\bar{G}(\hat{p}(y|\mathbf{X},\zeta))\leq\ell\}}\right| dy\right] \ .$$

Therefore, from Equation 13, and (19), we deduce

$$\mathbb{E}\left[\left|R_l(\hat{\Gamma}) - R_l(\bar{\Gamma})\right|\right] \leq C\frac{s}{\sqrt{N}} \ . \tag{21}$$

Now we bound the first term in the *r.h.s.* in Equation (20). Thanks to Proposition 2, we have that

$$\mathcal{E}_\ell(\bar{\Gamma}) = \mathbb{E}_{\mathbf{X},\zeta}\left[\int_{\bar{\Gamma}(\mathbf{X},\zeta)\triangle\Gamma^*_\ell(\mathbf{X})} |p(y|\mathbf{X}) - \lambda^*_\ell| \ dy\right] \ .$$

Now, we consider the following cases

- If $y \in \bar{\Gamma}(\mathbf{X},\zeta) \setminus \Gamma^*_\ell(\mathbf{X})$, we have that $p(y|\mathbf{X}) < \lambda^*_\ell$ and $\hat{p}(y|\mathbf{X},\zeta) \geq \bar{\lambda}_\ell$. Therefore,
$$|p(y|\mathbf{X}) - \lambda^*_\ell| = (\lambda^*_\ell - \bar{\lambda}_\ell) + (\bar{\lambda}_\ell - \hat{p}(y|\mathbf{X},\zeta)) + (\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})) \ .$$
Using the fact that $\bar{\lambda}_\ell - \hat{p}(y|\mathbf{X},\zeta) \leq 0$, we get
$$\int |p(y|\mathbf{X}) - \lambda^*_\ell| \mathbb{1}_{\{y\in\bar{\Gamma}(\mathbf{X},\zeta)\setminus\Gamma^*_\ell(\mathbf{X})\}} dy \leq \int \left((\lambda^*_\ell - \bar{\lambda}_\ell) + \left|\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})\right|\right) \mathbb{1}_{\{y\in\bar{\Gamma}(\mathbf{X},\zeta)\setminus\Gamma^*_\ell(\mathbf{X})\}} dy \ .$$
- If $y \in \Gamma^*_\ell(\mathbf{X}) \setminus \bar{\Gamma}(\mathbf{X},\zeta)$, we have that $p(y|\mathbf{X}) \geq \lambda^*_\ell$ and $\hat{p}(y|\mathbf{X},\zeta) < \bar{\lambda}_\ell$. Therefore,
$$|p(y|\mathbf{X}) - \lambda^*_\ell| = (p(y|\mathbf{X}) - \hat{p}(y|\mathbf{X},\zeta)) + (\hat{p}(y|\mathbf{X},\zeta) - \bar{\lambda}_\ell) + (\bar{\lambda}_\ell - \lambda^*_\ell) \ .$$
Using the fact that $\hat{p}(y|\mathbf{X},\zeta) - \bar{\lambda}_\ell < 0$, we get
$$\int |p(y|\mathbf{X}) - \lambda_\ell| \mathbb{1}_{\{y\in\Gamma^*_\ell(\mathbf{X})\setminus\bar{\Gamma}(\mathbf{X},\zeta)\}} dy \leq \int \left((\bar{\lambda}_\ell - \lambda^*_\ell) + \left|\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})\right|\right) \mathbb{1}_{\{y\in\Gamma^*_\ell(\mathbf{X})\setminus\bar{\Gamma}(\mathbf{X},\zeta)\}} dy \ .$$

From the above considerations, we deduce the following inequality

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X},\zeta}\left[\int_{\bar{\Gamma}(\mathbf{X},\zeta)\triangle\Gamma^*_\ell(\mathbf{X})} |p(y|\mathbf{X}) - \lambda^*_\ell| \ dy\right] \\
&\qquad \leq |\bar{\lambda}_\ell - \lambda^*_\ell| \ \mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\{y\in\bar{\Gamma}(\mathbf{X},\zeta)\triangle\Gamma^*_\ell(\mathbf{X})\}} \mathrm{d}y\right] + \mathbb{E}\left[\int_{\mathbb{R}} |\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \, \mathrm{d}y\right] \\
&\qquad\qquad \leq |\bar{\lambda}_\ell - \lambda^*_\ell| \times (\mathcal{L}(\bar{\Gamma}) - \mathcal{L}(\Gamma^*)) + \mathbb{E}\left[\int_{\mathbb{R}} |\hat{p}(y|\mathbf{X}) - p(y|\mathbf{X})| \, \mathrm{d}y\right] + 2su \ ,
\end{aligned}$$

where the last inequality is due to the fact that $\hat{p}(y|\mathbf{X},\zeta) = \hat{p}(y|\mathbf{X}) + \zeta\mathbb{1}_{y\in[-s,s]}$ with $|\zeta| \leq u$. But $\mathcal{L}(\bar{\Gamma}) = \mathcal{L}(\Gamma^*) = \ell$ by construction. Then,

$$\mathbb{E}\left[\mathcal{E}_\ell(\bar{\Gamma})\right] \leq \mathbb{E}\left[\int_{\mathbb{R}} |\hat{p}(y|\mathbf{X}) - p(y|\mathbf{X})| \, \mathrm{d}y\right] + 2su \ .$$

Injecting this last inequality and (21) into (20) gives the announced result. □

C.3 Consistency Result

This section is devoted to the proof of Theorem 1. We first provide a result on the L_1-integrated estimation error of $\hat{p}$.

Proposition 7. *Under Assumption 1, we have that*

$$\mathbb{E}\left[\int_{\mathbb{R}} |\hat{p}(y|\mathbf{X}) - p(y|\mathbf{X})| \, \mathrm{d}y\right] \leq C\left(\sqrt{s}\mathbb{E}\left[(\hat{f}(\mathbf{X}) - f^*(\mathbf{X}))^2\right] + \mathbb{E}\left[\left|\hat{f}(\mathbf{X}) - f^*(\mathbf{X})\right|\right]\right) + Cs^{5/2}\mathbb{E}\left[|\hat{\sigma}^2(\mathbf{X}) - \sigma^2(\mathbf{X})|\right] \ ,$$

where $C > 0$ is a constant which depends on σ_0 and σ_1 in Assumption 1.

Proof. To build this proof, we use the triangle inequality to split the term $|\hat{p}(y|\mathbf{X}) - p(y|\mathbf{X})|$ into 3. We then have to consider each of these terms consecutively. The first of these terms can be bounded as follows:

$$\begin{aligned}
&\left|\frac{1}{\sqrt{2\pi\hat{\sigma}^2(\mathbf{X})}}\exp\left(-\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right) - \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\right| \\
&\qquad \leq \frac{1}{\sqrt{2\pi\hat{\sigma}^2(\mathbf{X})}}\exp\left(-\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\left|1 - \frac{\hat{\sigma}(\mathbf{X})}{\sigma(\mathbf{X})}\right| \\
&\qquad = \frac{1}{\sqrt{2\pi\hat{\sigma}^2(\mathbf{X})}}\exp\left(-\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\left|\frac{\sigma(\mathbf{X}) - \hat{\sigma}(\mathbf{X})}{\sigma(\mathbf{X})}\right| \ . \qquad (22)
\end{aligned}$$

This upper-bound consists of two parts. One part which is the density of a Gaussian random variable (whose integral *w.r.t.* y is 1) and a second term which is independent of y. Observe that this second term $|\sigma(\mathbf{X}) - \hat{\sigma}(\mathbf{X})|$ is of the same order as $|\sigma^2(\mathbf{X}) - \hat{\sigma}^2(\mathbf{X})|$. Indeed, notice that when $\sigma(\mathbf{X}) > \hat{\sigma}(\mathbf{X})$

$$\sigma^2(\mathbf{X}) - \hat{\sigma}^2(\mathbf{X}) = (\sigma(\mathbf{X}) - \hat{\sigma}(\mathbf{X}))(\sigma(\mathbf{X}) + \hat{\sigma}(\mathbf{X})) \geq \left(\sigma_0 + \frac{1}{\sqrt{s}}\right)(\sigma(\mathbf{X}) - \hat{\sigma}(\mathbf{X})) \ ,$$

where in the last inequality, we use Assumption 1 and the fact that $\hat{\sigma}(X) \geq 1/\sqrt{s}$. Written differently, this means that

$$|\sigma(\mathbf{X}) - \hat{\sigma}(\mathbf{X})| \leq \frac{\sqrt{s}}{1+\sigma_0\sqrt{s}}|\sigma^2(\mathbf{X}) - \hat{\sigma}^2(\mathbf{X})| \leq \frac{\sigma_0\sqrt{s}}{1+\sigma_0\sqrt{s}}\frac{|\sigma^2(\mathbf{X}) - \hat{\sigma}^2(\mathbf{X})|}{\sigma_0} \leq C|\sigma^2(\mathbf{X}) - \hat{\sigma}^2(\mathbf{X})| \ ,$$

since $1/\sigma_0 \leq C$. The same reasoning holds in the case where $\sigma(\mathbf{X}) < \hat{\sigma}(\mathbf{X})$ and then we conclude that

$$|\sigma(\mathbf{X}) - \hat{\sigma}(\mathbf{X})| \leq C|\sigma^2(\mathbf{X}) - \hat{\sigma}^2(\mathbf{X})| \ .$$

Injecting this bound into (22) and using again Assumption 1, we deduce that

$$\begin{aligned}
&\int_{\mathbb{R}}\left|\frac{1}{\sqrt{2\pi\hat{\sigma}^2(\mathbf{X})}}\exp\left(-\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right) - \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\right| dy \\
&\qquad \leq C\left|\sigma(\mathbf{X}) - \hat{\sigma}(\mathbf{X})\right| \leq C\left|\sigma^2(\mathbf{X}) - \hat{\sigma}^2(\mathbf{X})\right| \ . \qquad (23)
\end{aligned}$$

Let us now consider the second term in the decomposition of $|\hat{p}(y|\mathbf{X}) - p(y|\mathbf{X})|$. Since $x \mapsto \exp(-x)$ is

1-Lipschitz on $\mathbb{R}_+$, from Assumption 1 we have that in the case where $(y-\hat{f}(\mathbf{X}))^2 \geq (y-f^*(\mathbf{X}))^2$

$$\begin{aligned}
&\left|\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)-\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\right| \\
&\quad=\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\left|\exp\left(-\left(\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\right)-1\right| \\
&\quad\leq\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})\times 2\hat{\sigma}^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\left|(y-\hat{f}(\mathbf{X}))^2-(y-f^*(\mathbf{X}))^2\right| \\
&\quad\leq\frac{C}{\sqrt{2\pi\hat{\sigma}^2(\mathbf{X})}\hat{\sigma}(\mathbf{X})}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\left|(y-\hat{f}(\mathbf{X}))^2-(y-f^*(\mathbf{X}))^2\right| \ .
\end{aligned}$$

Using the following decomposition

$$(y-\hat{f}(\mathbf{X}))^2-(y-f^*(\mathbf{X}))^2=(\hat{f}(\mathbf{X})-f^*(\mathbf{X}))^2+2(y-f^*(\mathbf{X})(f^*(\mathbf{X})-\hat{f}(\mathbf{X})) \ ,$$

we deduce

$$\begin{aligned}
&\left|\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)-\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\right| \\
&\leq\frac{C}{\sqrt{2\pi\hat{\sigma}^2(\mathbf{X})}\hat{\sigma}(\mathbf{X})}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\left(\left(\hat{f}(\mathbf{X})-f^*(\mathbf{X})\right)^2+|y-f^*(\mathbf{X})||\hat{f}(\mathbf{X})-f^*(\mathbf{X})|\right) \ . \quad (24)
\end{aligned}$$

In the case where $(y-\hat{f}(\mathbf{X}))^2 \leq (y-f^*(\mathbf{X}))^2$, we obtain similar bound as in the above equation by switching the role of $\hat{f}$ by f^*. Notice that $\int_{\mathbb{R}}\frac{1}{\sqrt{2\pi\hat{\sigma}^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\times|y-f^*(\mathbf{X})|dy$ is the expectation of the r.v. $|Y-f^*(\mathbf{X})|$ where Y is Gaussian with expectation $f^*(\mathbf{X})$ and variance $\hat{\sigma}^2(\mathbf{X})$. Therefore, using the fact that $\mathbb{E}\left[|Z-\mathbb{E}\left[Z\right]|\right]\leq\sqrt{\mathrm{Var}(Z)}$ for any real valued random variable Z, we get

$$\begin{aligned}
&\int_{\mathbb{R}}\left|\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)-\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\right|dy \\
&\qquad\leq\frac{C}{\hat{\sigma}(\mathbf{X})}\left(\left(\hat{f}(\mathbf{X})-f^*(\mathbf{X})\right)^2+\hat{\sigma}(\mathbf{X})|\hat{f}(\mathbf{X})-f^*(\mathbf{X})|\right) \ .
\end{aligned}$$

Finally, using that $\hat{\sigma}(\mathbf{X})\geq 1/\sqrt{s}$, we deduce

$$\begin{aligned}
&\int_{\mathbb{R}}\left|\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-\hat{f}(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)-\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)\right|dy \\
&\qquad\leq C\left(\sqrt{s}\left(\hat{f}(\mathbf{X})-f^*(\mathbf{X})\right)^2+|\hat{f}(\mathbf{X})-f^*(\mathbf{X})|\right) \ . \quad (25)
\end{aligned}$$

The remaining term in the decomposition of $|\hat{p}(y|\mathbf{X})-p(y|\mathbf{X})|$ is

$$\begin{aligned}
&\left|\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)-\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\sigma^2(\mathbf{X})}\right)\right| \\
&\quad=\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\sigma^2(\mathbf{X})}\right)\left|\exp\left(-\left(\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}-\frac{(y-f^*(\mathbf{X}))^2}{2\sigma^2(\mathbf{X})}\right)\right)-1\right| \ .
\end{aligned}$$

Hence, if $\sigma^2(\mathbf{X})\geq\hat{\sigma}^2(\mathbf{X})$, since $x\mapsto\exp(-x)$ is 1-Lipschitz on $\mathbb{R}_+$, we deduce from the above inequality that

$$\begin{aligned}
&\left|\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)-\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\sigma^2(\mathbf{X})}\right)\right| \\
&\qquad\leq\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\sigma^2(\mathbf{X})}\right)\left|\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}-\frac{(y-f^*(\mathbf{X}))^2}{2\sigma^2(\mathbf{X})}\right| \ .
\end{aligned}$$

Therefore, from Assumption 1, and since $\hat{\sigma}^2(\mathbf{X}) \geq 1/s$, we get in the case where $\sigma^2(\mathbf{X}) \geq \hat{\sigma}^2(\mathbf{X})$

$$\left| \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right) - \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\sigma^2(\mathbf{X})}\right) \right| \\ \leq \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\sigma^2(\mathbf{X})}\right) Cs(y-f^*(\mathbf{X}))^2 \left|\hat{\sigma}^2(\mathbf{X}) - \sigma^2(\mathbf{X})\right| \ . \quad (26)$$

In the case where $\sigma^2(\mathbf{X}) \leq \hat{\sigma}^2(\mathbf{X})$, using same arguments and additionally the fact that $\hat{\sigma}^2(\mathbf{X}) \leq s$, we can obtain

$$\left| \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right) - \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\sigma^2(\mathbf{X})}\right) \right| \\ \leq \frac{\sqrt{s}}{\sigma_0\sqrt{2\pi\hat{\sigma}^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right) Cs(y-f^*(\mathbf{X}))^2 \left|\hat{\sigma}^2(\mathbf{X}) - \sigma^2(\mathbf{X})\right| \ . \quad (27)$$

Therefore, from Equation (26), and (27), we get

$$\int_{\mathbb{R}} \left| \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right) - \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\sigma^2(\mathbf{X})}\right) \right| dy \\ \leq Cs^{5/2} \left|\hat{\sigma}^2(\mathbf{X}) - \sigma^2(\mathbf{X})\right| \ , \quad (28)$$

where we used the fact that the integral *w.r.t.* y is the variance of Gaussian r.v. with variance $\hat{\sigma}^2(\mathbf{X})$ and is then such that $\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\hat{\sigma}^2(\mathbf{X})}} \exp\left(-\frac{(y-f^*(\mathbf{X}))^2}{2\hat{\sigma}^2(\mathbf{X})}\right)(y-f^*(\mathbf{X}))^2 dy = \hat{\sigma}^2(\mathbf{X}) \leq s$. The combination of Equations (23), (25), and (28) yields the result. □

Now, we provide the proof of Theorem 1.

Proof of Theorem 1. We prove the consistency $\hat{\Gamma}$ *w.r.t.* the symmetric difference distance $\mathcal{H}$. We have that

$$\mathcal{H}(\hat{\Gamma}) \leq \mathbb{E}\left[\int_{\hat{\Gamma}(\mathbf{X},\zeta)\triangle\bar{\Gamma}(\mathbf{X},\zeta)} dy\right] + \mathbb{E}\left[\int_{\bar{\Gamma}(\mathbf{X},\zeta)\triangle\Gamma^*(\mathbf{X})} dy\right] \ . \quad (29)$$

We bound the first term in the *r.h.s.* in the above inequality.

$$\mathbb{E}\left[\int_{\hat{\Gamma}(\mathbf{X},\zeta)\triangle\bar{\Gamma}(\mathbf{X},\zeta)} dy\right] = \mathbb{E}\left[\int_{\mathbb{R}} \left|\mathbb{1}_{\{y\in\hat{\Gamma}(\mathbf{X},\zeta)\}} - \mathbb{1}_{\{y\in\bar{\Gamma}(\mathbf{X},\zeta)\}}\right|\right] \\ = \mathbb{E}\left[\int_{\mathbb{R}} \left|\mathbb{1}_{\{\hat{G}(\hat{p}(y|\mathbf{X},\zeta))\leq\ell\}} - \mathbb{1}_{\{\bar{G}(\hat{p}(y|\mathbf{X},\zeta))\leq\ell\}}\right| dy\right] \ .$$

Therefore, from Equations (13) and (19), we deduce

$$\mathbb{E}\left[\int_{\hat{\Gamma}(\mathbf{X},\zeta)\triangle\bar{\Gamma}(\mathbf{X},\zeta)} dy\right] \leq \frac{Cs}{\sqrt{N}} \ . \quad (30)$$

Now, we study the second term in the *r.h.s.* of Equation (29). We observe that if $y \in \bar{\Gamma}(\mathbf{X},\zeta) \setminus \Gamma^*_\ell(\mathbf{X})$ the following holds

- on the event $\{\bar{G}^{-1}(\ell) \geq G^{-1}(\ell)\}$, $|\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \geq \left|p(y|\mathbf{X}) - G^{-1}(\ell)\right|$,
- on the event $\{\bar{G}^{-1}(\ell) < G^{-1}(\ell)\}$,

$$\text{either } \ \hat{p}(y|\mathbf{X},\zeta) \in (\bar{G}^{-1}(\ell), G^{-1}(\ell)) \ \text{ or } \ |\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \geq \left|p(y|\mathbf{X}) - G^{-1}(\ell)\right| \ .$$

Note that similar reasoning holds if $y \in \Gamma^*_\ell(\mathbf{X}) \setminus \bar{\Gamma}(\mathbf{X}, \zeta)$. Therefore, we deduce that conditional on $\mathcal{D}_n$,

$$\begin{aligned}\mathbb{E}\left[\int_{\bar{\Gamma}(\mathbf{X},\zeta)\triangle\Gamma^*(\mathbf{X})} dy\right] \leq \mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\{|\hat{p}(y|\mathbf{X},\zeta)-p(y|\mathbf{X})|\geq|p(y|\mathbf{X})-G^{-1}(\ell)|\}} dy\right] \\ + \mathbb{1}_{\{\bar{G}^{-1}(\ell)<G^{-1}(\ell)\}}\mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\{\hat{p}(y|\mathbf{X},\zeta)\in(\bar{G}^{-1}(\ell),G^{-1}(\ell))\}} dy\right] \\ + \mathbb{1}_{\{\bar{G}^{-1}(\ell)\geq G^{-1}(\ell)\}}\mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\{\hat{p}(y|\mathbf{X},\zeta)\in(G^{-1}(\ell),\bar{G}^{-1}(\ell))\}} dy\right] \ .\end{aligned}$$

Using first the definition of $\bar{G}$ and then the fact that $\bar{G}(\bar{G}^{-1}(\ell)) = G(G^{-1}(\ell)) = \ell$ in this last inequality, we deduce the following

$$\begin{aligned}\mathbb{E}\left[\int_{\bar{\Gamma}(\mathbf{X},\zeta)\triangle\Gamma^*(\mathbf{X})} dy\right] \leq \mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\{|\hat{p}(y|\mathbf{X},\zeta)-p(y|\mathbf{X})|\geq|p(y|\mathbf{X})-G^{-1}(\ell)|\}} dy\right] \\ + \mathbb{E}\left[\left|G(G^{-1}(\ell)) - \bar{G}(G^{-1}(\ell))\right|\right] \ .\end{aligned}$$

Now, we observe that

$$\begin{aligned}\mathbb{E}\left[\left|G(G^{-1}(\ell)) - \bar{G}(G^{-1}(\ell))\right|\right] &\leq \mathbb{E}\left[\int_{\mathbb{R}} \left|\mathbb{1}_{\{p(y|\mathbf{X})\geq G^{-1}(\ell)\}} - \mathbb{1}_{\{\hat{p}(y|\mathbf{X},\zeta)\geq G^{-1}(\ell)\}}\right| dy\right] \\ &\leq \mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\{|\hat{p}(y|\mathbf{X},\zeta)-p(y|\mathbf{X})|\geq|p(y|\mathbf{X})-G^{-1}(\ell)|\}} dy\right] \ .\end{aligned}$$

Therefore, we have obtained

$$\begin{aligned}\mathbb{E}\left[\int_{\bar{\Gamma}(\mathbf{X},\zeta)\triangle\Gamma^*(\mathbf{X})} dy\right] &\leq 2\mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\{|\hat{p}(y|\mathbf{X},\zeta)-p(y|\mathbf{X})|\geq|p(y|\mathbf{X})-G^{-1}(\ell)|\}} dy\right] \qquad (31) \\ &\leq 2\int_{\mathbb{R}} \mathbb{P}\left(|\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \geq \left|p(y|\mathbf{X}) - G^{-1}(\ell)\right|\right) dy \ .\end{aligned}$$

Let us consider the term in the *r.h.s* of Equation (31). Let $\delta > 0$, we have that

$$\begin{aligned}\int_{\mathbb{R}} \mathbb{P}\left(|\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \geq \left|p(y|\mathbf{X}) - G^{-1}(\ell)\right|\right) dy \leq \int_{\mathbb{R}} \mathbb{P}\left(|\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \geq \delta\right) dy \\ + \int_{\mathbb{R}} \mathbb{P}\left(\left|p(y|\mathbf{X}) - G^{-1}(\ell)\right| \leq \delta\right) dy \ .\end{aligned}$$

From Markov's inequality, we deduce

$$\begin{aligned}\int_{\mathbb{R}} \mathbb{P}\left(|\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \geq \left|p(y|\mathbf{X}) - G^{-1}(\ell)\right|\right) dy \leq \frac{1}{\delta}\mathbb{E}\left[\int_{\mathbb{R}} |\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \, dy\right] \\ + G(G^{-1}(\ell) - \delta) - G(G^{-1}(\ell) + \delta) \ . \qquad (32)\end{aligned}$$

Since $\hat{p}$ is supported on $[-s, s]$, we observe that

$$\begin{aligned}\mathbb{E}\left[\int_{\mathbb{R}} |\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \, \mathrm{d}y\right] &= \mathbb{E}\left[\int_{[-s,s]} |\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \, \mathrm{d}y\right] + \mathbb{E}\left[\int_{|y|\geq s}^{+\infty} p(y|\mathbf{X}) \mathrm{d}y\right] \\ &\leq \mathbb{E}\left[\int_{[-s,s]} |\hat{p}(y|\mathbf{X}) - p(y|\mathbf{X})| \, \mathrm{d}y\right] + 2su + \mathbb{E}\left[\int_{|y|\geq s}^{+\infty} p(y|\mathbf{X}) \mathrm{d}y\right] \ . \qquad (33)\end{aligned}$$

Now, we observe that

$$\mathbb{E}\left[\int_{s}^{+\infty} p(y|\mathbf{X}) \mathrm{d}y\right] = \mathbb{E}\left[\mathbb{1}_{\{|f^*(X)|\leq\frac{s}{2}\}} \int_{s}^{+\infty} p(y|\mathbf{X}) \mathrm{d}y\right] + \mathbb{E}\left[\mathbb{1}_{\{|f^*(X)|>\frac{s}{2}\}} \int_{s}^{+\infty} p(y|\mathbf{X}) \mathrm{d}y\right] \ .$$

From Markov inequality and Assumption 2, the second term of the r.h.s. in the above inequality is bounded by

$$\mathbb{E}\left[\mathbb{1}_{\{|f^*(\mathbf{X})|>\frac{s}{2}\}}\int_s^{+\infty} p(y|\mathbf{X})\mathrm{d}y\right] \leq \mathbb{E}\left[\mathbb{1}_{\{|f^*(\mathbf{X})|>\frac{s}{2}\}}\right] \leq \frac{2C_1}{s} \ .$$

On the other hand, we observe that

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{\{|f^*(\mathbf{X})|\leq\frac{s}{2}\}}\int_s^{+\infty} p(y|\mathbf{X})\mathrm{d}y\right] &\leq \mathbb{E}\left[\int_s^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{(y-s/2)^2}{2\sigma^2(\mathbf{X})}\right)\mathrm{d}y\right] \\ &= \mathbb{E}\left[\int_{s/2}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2(\mathbf{X})}}\exp\left(-\frac{y^2}{2\sigma^2(\mathbf{X})}\right)\mathrm{d}y\right] \ .\end{aligned}$$

Therefore, Assumption 1 and standard result on Gaussian tails yields for $s \geq 1$

$$\mathbb{E}\left[\mathbb{1}_{\{|f^*(\mathbf{X})|\leq\frac{s}{2}\}}\int_s^{+\infty} p(y|\mathbf{X})\mathrm{d}y\right] \leq \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{s^2}{8\sigma_1^2}\right) \ .$$

Hence combining the above inequalities, we get for $s \geq 1$

$$\mathbb{E}\left[\int_s^{+\infty} p(y|\mathbf{X})\mathrm{d}y\right] \leq C'\left[\exp\left(-\frac{s^2}{8\sigma_1^2}\right)+\frac{C}{s}\right] \ ,$$

where C and C' are two positive constants. Note that similar arguments yields

$$\mathbb{E}\left[\int_{-\infty}^{-s} p(y|\mathbf{X})\mathrm{d}y\right] \leq C'\left[\exp\left(-\frac{s^2}{8\sigma_1^2}\right)+\frac{C}{s}\right] \ .$$

Therefore considering Equation (33) and Proposition 7 and defining $s = \log(\min(n, N))$, we get

$$\lim_n \frac{1}{\delta}\mathbb{E}\left[\int_{\mathbb{R}} |\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})|\, dy\right] = 0 \ .$$

Hence we obtain from Equation (32) that for all $\delta > 0$

$$\limsup_n \int_{\mathbb{R}} \mathbb{P}\left(|\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \geq \left|p(y|\mathbf{X}) - G^{-1}(\ell)\right|\right) dy \leq G(G^{-1}(\ell)-\delta) - G(G^{-1}(\ell)+\delta) \ .$$

Since G is continuous, with $\delta \to 0$, we get

$$\lim_n \int_{\mathbb{R}} \mathbb{P}\left(|\hat{p}(y|\mathbf{X},\zeta) - p(y|\mathbf{X})| \geq \left|p(y|\mathbf{X}) - G^{-1}(\ell)\right|\right) dy = 0 \ .$$

The above equation together with Equation (29), (30), (31) yields the desired result. □

C.4 Rates of convergence

We start this section with a result on the estimation error of $\hat{p}$ *w.r.t.* the sup-norm.

Proposition 8. *Let $s = \log(\min(n, N))$. Under Assumptions 1, 4, and 5, we have that*

$$\begin{aligned}&\sup_{(\mathbf{x},y)\in\mathcal{C}\times[-s,s]} |\hat{p}(y|\mathbf{x}) - p(y|\mathbf{x})| \leq \\ &\qquad C\left(s\sup_{\mathbf{x}\in\mathcal{C}}\left(\hat{f}(\mathbf{x}) - f^*(\mathbf{x})\right)^2 + s^2\sup_{\mathbf{x}\in\mathcal{C}}\left|\hat{f}(\mathbf{x}) - f^*(\mathbf{x})\right| + s^3\sup_{\mathbf{x}\in\mathcal{C}}\left|\hat{\sigma}^2(\mathbf{x}) - \sigma^2(\mathbf{x})\right|\right) \ ,\end{aligned}$$

where $C > 0$ is a constant which depends on f^*, σ^2, and on the set $\mathcal{C}$.

Proof. We consider the same decomposition into 3 that we used in the proof of Proposition 7. Using the fact that $\hat{\sigma}(\mathbf{x}) \geq \frac{1}{\sqrt{s}}$ and Assumption 1, we get for all $\mathbf{x} \in \mathcal{C}$, and $y \in [-s, s]$ (*c.f.,* Eq. (22)), the first term is controlled as follows:

$$\left| \frac{1}{\sqrt{2\pi\hat{\sigma}^2(\mathbf{x})}} \exp\left(-\frac{(y - \hat{f}(\mathbf{x}))^2}{2\hat{\sigma}^2(\mathbf{x})} \right) - \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{x})}} \exp\left(-\frac{(y - \hat{f}(\mathbf{x}))^2}{2\hat{\sigma}^2(\mathbf{x})} \right) \right| \leq Cs \sup_{x \in \mathcal{C}} \left| \hat{\sigma}^2(\mathbf{x}) - \sigma^2(\mathbf{x}) \right| \ . \quad (34)$$

According to the second term, from Assumptions 4 and 5, and since f^* is a Lipschitz function on the compact $\mathcal{C}$, we have that $|f^*(\mathbf{x})| \leq s$ for n, N large enough. Therefore, using the fact that $x \mapsto \exp(-x)$ is 1-Lipschitz on $\mathbb{R}_+$ and that $\frac{1}{s} \leq \hat{\sigma}^2(\mathbf{x})$, we get (*c.f.,* Eq. (24))

$$\left| \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{x})}} \exp\left(-\frac{(y - \hat{f}(\mathbf{x}))^2}{2\hat{\sigma}^2(\mathbf{x})} \right) - \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{x})}} \exp\left(-\frac{(y - f^*(\mathbf{x}))^2}{2\hat{\sigma}^2(\mathbf{x})} \right) \right| \leq$$
$$C \left(s \sup_{\mathbf{x} \in \mathcal{C}} \left(\hat{f}(\mathbf{x}) - f^*(\mathbf{x}) \right)^2 + s^2 \sup_{\mathbf{x} \in \mathcal{C}} \left| \hat{f}(\mathbf{x}) - f^*(\mathbf{x}) \right| \right) \ . \quad (35)$$

Finally, considering the last term, we deduce from (26) and (27) that

$$\left| \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{x})}} \exp\left(-\frac{(y - f^*(\mathbf{x}))^2}{2\hat{\sigma}^2(\mathbf{x})} \right) - \frac{1}{\sqrt{2\pi\sigma^2(\mathbf{x})}} \exp\left(-\frac{(y - f^*(\mathbf{x}))^2}{2\sigma^2(\mathbf{x})} \right) \right| \leq Cs^3 \sup_{\mathbf{x} \in \mathcal{C}} \left| \hat{\sigma}^2(\mathbf{x}) - \sigma^2(\mathbf{x}) \right| \ . \quad (36)$$

The combination of Equations (34), (35), and (36) gives the proposition. □

Proof of Proposition 5. We recall that

$$\mathcal{E}_\ell\left(\bar{\Gamma}\right) = \mathbb{E}\left[\int_{\bar{\Gamma}(\mathbf{X},\zeta) \triangle \Gamma_\ell^*(\mathbf{X})} |p(y|\mathbf{X}) - \lambda_\ell^*| \, dy \right] \ .$$

Now, we observe that for $y \in \bar{\Gamma}(\mathbf{X}, \zeta) \triangle \Gamma_\ell^*(\mathbf{X})$

$$|p(y|\mathbf{X}) - \lambda_\ell^*| \leq |\hat{p}(y|\mathbf{X}, \zeta) - p(y|\mathbf{X}) \ | + \left| \bar{\lambda}_\ell - \lambda_\ell^* \right| \ ,$$

where we recall that $\bar{\lambda}_\ell := \bar{G}^{-1}(\ell)$, with $\bar{G}$ defined in Eq. (12). Using similar arguments as those used in the proof of Theorem 4.4 in [7] that is inspired by Theorem 2.12 in [4], it is not difficult to see that conditional on $\mathcal{D}_n$,

$$\left| \bar{\lambda}_\ell - \lambda_\ell^* \right| \leq \sup_{(\mathbf{x},y) \in \mathcal{C} \times \mathbb{R}} |\hat{p}(y|\mathbf{x}) - p(y|\mathbf{x})| + u := \hat{m}(u) \ .$$

Therefore, we deduce that

$$\mathbb{E}\left[\int_{\bar{\Gamma}(\mathbf{X},\zeta) \triangle \Gamma_\ell^*(\mathbf{X})} |p(y|\mathbf{X}) - \lambda_\ell^*| \, dy \right] \leq 2\hat{m}(u) \mathbb{E}\left[\int_{\mathbb{R}} \mathbb{1}_{\{|p(y|\mathbf{X}) - \lambda_\ell^*| \leq 2\hat{m}(u)\}} \mathrm{d}y \right] \ .$$

Hence from the above inequality, and Assumption 6 we get

$$\mathbb{E}\left[\mathcal{E}_\ell\left(\bar{\Gamma}\right) \right] \leq 2^{1+\alpha} c_0 \mathbb{E}\left[\hat{m}(u)^{1+\alpha} \right] \ .$$

Therefore, from Equations (20) and (21), we obtain the following with $s = \log(\min(n, N))$

$$\mathbb{E}\left[\mathcal{E}_\ell\left(\hat{\Gamma}\right) \right] \leq C \left(\mathbb{E}\left[\left(\sup_{(x,y) \in \mathbb{R} \times \mathcal{C}} |\hat{p}(y|\mathbf{x}) - p(y|\mathbf{x})| \right)^{1+\alpha} \right] + u^{1+\alpha} + \frac{\log(N)}{N} \right) \ .$$

Finally, since $\hat{p}$ is supported on $[-s, s]$, we have

$$\sup_{(\mathbf{x},y) \in \mathcal{C} \times \mathbb{R}} |\hat{p}(y|\mathbf{x}) - p(y|\mathbf{x})| \leq \sup_{(\mathbf{x},y) \in \mathcal{C} \times [-s,s]} |\hat{p}(y|\mathbf{x}) - p(y|\mathbf{x})| + \sup_{(\mathbf{x},y) \in \mathcal{C} \times \mathbb{R} \setminus [-s,s]} p(y|\mathbf{x}) \ .$$

For n, N large enough, we can assume, since f^* is bounded, that $|f^*(\mathbf{X})| \leq s/2$. From Assumption 1, we have for n, N large enough

$$\sup_{(\mathbf{x},y)\in\mathcal{C}\times\mathbb{R}\setminus[-s,s]} p(y|\mathbf{x}) \leq C\exp\left(-\frac{s^2}{8\sigma_1^2}\right) \leq \exp(-s) \leq \frac{C}{\min(n,N)} \ ,$$

which yields the desired result.

□

Proof of Theorem 2. The proof is a straightforward application of Proposition 5, 8, and Theorem 3, where we also use the fact that for n, N large enough

$$\left|\hat{\sigma}^2(\mathbf{x}) - \sigma^2(\mathbf{x})\right| \leq \left|\tilde{\sigma}^2(\mathbf{x}) - \sigma^2(\mathbf{x})\right| \ .$$

□