Journal of Machine Learning Research 27 (2026) 1-54 Submitted 11/24; Revised 5/26; Published 5/26

Generalized Resubstitution for Regression Error Estimation

Diego Marcondes DIEGO.MARCONDES@ANU.EDU.AU
Mathematical Sciences Institute
France-Australia Mathematical Sciences and Interactions ANU-CNRS International Research Lab
The Australian National University
Canberra, Australia

Ulisses Braga-Neto ULISSES@TAMU.EDU
Department of Electrical and Computer Engineering
Texas A&M University
College Station, TX 77843, USA

Editor: Eric Laber

Abstract

We propose generalized resubstitution error estimators for regression. Each error estimator in this class corresponds to a choice of an empirical probability measure and a loss function. The standard empirical probability measure and the quadratic loss lead to the standard sum of squares error estimator. Other choices of empirical probability measure lead to more general estimators with superior bias and variance properties. We prove that these error estimators are consistent under broad assumptions. In addition, procedures for choosing the empirical measure based on the method of moments and maximum pseudo-likelihood are proposed and investigated. Detailed experimental results using polynomial regression demonstrate empirically the superior finite-sample bias and variance properties of the proposed estimators. The R code for the experiments is provided.

Keywords: Error Estimation, Regression, Resubstitution, Bolstering, Statistical Learning

1. Introduction

In supervised learning, the objective is to build a mapping $\psi : \mathcal{X} \to \mathcal{Y}$ to predict the value of a target $Y \in \mathcal{Y}$ from an input $X \in \mathcal{X}$, where $\mathcal{X}$ and $\mathcal{Y}$ are suitable spaces, using a training data set $S_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. Once ψ is built, it is necessary to evaluate its performance, and, in recent years, it has become the norm to use the empirical error on data not appear-

ing in S_n to benchmark predictors (Russakovsky et al., 2015; Jiang et al., 2019). A more rigorous alternative to performance evaluation is provided by a *statistical approach* to learning, where X and Y are assumed to be random variables taking values in $\mathcal{X}$ and $\mathcal{Y}$, respectively, and the training data S_n is generally assumed to be an independent and identically distributed sample from a joint probability measure $\nu(X, Y)$. This approach allows one to define rigorously the *generalization error* of the predictor on future data, given an appropriate loss criterion, as the expected loss with respect to ν.

A crucial problem in practice is how to estimate accurately the generalization error. A predictor is useful only if its generalization error can be stated with confidence. In the statistical approach to learning, the problem of error estimation has been extensively studied in the classification case, where $\mathcal{Y}$ consists of a finite set of labels, and the loss criterion is simply whether the predictor recovers the label or not; see Toussaint (1974); Hand (1986); McLachlan (1987); Schiavo and Hand (2000); Braga-Neto and Dougherty (2015) for comprehensive surveys. However, the problem of estimating the generalization error in the case of regression, where $\mathcal{Y}$ is a Euclidean space, is much less studied. In this paper, we provide a comprehensive study of a new class of error estimators for regression problems, namely, the family of *generalized regression resubstitution error estimators*.

The popular test-set error estimator is known to have excellent statistical properties; it is unbiased and consistent regardless of the sample size or data-generating distribution (Braga-Neto and Dougherty, 2015). However, this is only true if the test data is truly independent of training and used only once (Yousefi et al., 2011). In practice, this is rarely the case, with the test data being reused, in some cases heavily, to measure performance improvement, creating a situation known as "training to the test data" (Recht et al., 2019). In addition, if training and testing sample sizes are small, the test-set error estimator can display large variance and become unreliable, which means that test-set error estimation requires cheap access to plentiful labeled data.

Error estimators based on resampling, such as cross-validation (CV) (Lachenbruch and Mickey, 1968; Stone, 1974; Bates et al., 2024) and bootstrap (Efron, 1979, 1983; Efron and Tibshirani, 1997), are also popular choices. Despite their widespread use and strong theoretical foundations, these procedures have practical shortcomings that limit their applicability in many modern learning scenarios. Their primary limitation is computational, as they require refitting the model multiple times making it prohibitively expensive for models with long training times, such as deep neural networks or large-scale kernel methods. Moreover, CV estimators have high variance

when the sample size is small, and the choice of the number of folds introduces an additional hyperparameter that needs to be tuned. In settings where training is costly or data are scarce, these drawbacks render resampling methods impractical, motivating the need for more computationally efficient alternatives.

The simplest computationally cheap alternative that requires no separate test data is the empirical error on the training data; this is known as the *resubstitution* error estimator (Smith, 1947). The resubstitution estimator is, however, usually optimistically biased, the more so the more the prediction algorithm overfits the training data. Optimistic bias implies that the difference between the resubstitution estimate and the true error, which has been called the "generalization gap" (Keskar et al., 2016), is negative with a high probability. It is key, therefore, to investigate mechanisms to reduce the bias.

Bolstered resubstitution, introduced in Braga-Neto and Dougherty (2004), proposed a modification to resubstitution for classification, where the empirical measure, which produces the plain resubstitution estimator, is smoothed by kernels in order to reduce both the bias and variance in small-sample cases. In Ghane and Braga-Neto (2022), the family of generalized resubstitution error estimators for classification was proposed and studied. These estimators are defined in terms of arbitrary empirical probability measures and include as special cases both plain and bolstered resubstitution, as well as posterior-probability (Lugosi and Pawlak, 1994), Gaussian-process (Hefny and Atiya, 2010), and Bayesian (Dalton and Dougherty, 2011a,b) classification error estimators. It was shown in that paper that generalized resubstitution error estimators are consistent and asymptotically unbiased for the two-class problem if the corresponding empirical probability measure converges uniformly to the standard empirical probability measure and the hypothesis space has a finite VC dimension.

In this work, we extend the generalized resubstitution error estimators to the general statistical learning framework for regression when the loss function has a moment of order > 1 uniformly bounded in the hypothesis space. In particular, we consider regression problems under the quadratic loss function and extend the special cases studied in Ghane and Braga-Neto (2022) to them. The generalized error estimators can be defined as the expectation of the loss function under an arbitrary empirical probability measure or as the expectation of a generalized loss function under the standard empirical probability measure.

These representations allow us to establish several sufficient conditions for the consistency of these estimators that not only extend that of Ghane and

Braga-Neto (2022) but are also weaker than it. For example, we show that if the generalized loss function converges uniformly to the original one, or if the expectation of the generalized loss function under the data-generating distribution converges to that of the original loss function, then the generalized resubstitution error estimator is consistent. In particular, this last condition allows for the variance of the empirical measure to not converge to zero. Sufficient conditions for consistency based on the variance of the empirical measure are also established for the case of twice-differentiable loss functions.

Finally, since consistency is attained under many conditions, there is plenty of room to choose the empirical measure, and we propose methods to estimate the parameters of the empirical measure from the data. We focus on bolstered error estimators and propose method of moments and maximum pseudo-likelihood estimators of the covariance matrix of the bolstered empirical measure, which formalizes the heuristic approach proposed in Braga-Neto and Dougherty (2004) for that purpose.

We summarize our contributions as follows:

- We propose generalized resubstitution error estimators for regression, a general framework for resubstitution-like regression error estimators. Like the classification error estimators in Ghane and Braga-Neto (2022), our estimators are based on arbitrary empirical probability measures.

- We establish several sufficient conditions for the consistency of these estimators that not only extend that of Ghane and Braga-Neto (2022) but are also weaker than it.

- Additional sufficient conditions for consistency based on the variance of the empirical measure are established for the case of twice-differentiable loss functions.

- To address the issue of choosing the proper amount of smoothing to cancel estimator bias, we propose method of moments and maximum pseudo-likelihood estimators of the covariance matrix of the Gaussian empirical probability measure for Gaussian bolstering error estimation.

Generalized resubstitution estimators offer a practical compromise between computational efficiency and predictive accuracy. Unlike CV, these estimators avoid repeated model fitting, making them particularly suitable for modern machine learning models whose training costs can be prohibitive.

At the same time, they mitigate the bias of plain resubstitution while keeping its statistical consistency. Consequently, generalized resubstitution is a viable alternative in applications where (a) no sufficiently large test sample is available and (b) CV is too expensive to compute.

Our paper is organized as follows:

- In Sections 2 and 3 we define the problem of error estimation in statistical learning and give an informal presentation of the generalized resubstitution error estimators.
- In Section 4, we formally define the generalized resubstitution error estimators and the consistency of error estimators, and present sufficient conditions for the consistency of generalized resubstitution error estimators. In particular, in Section 4.2 we present sufficient conditions that hold in general cases, in Section 4.3 we focus on estimators in which the associated empirical measure depends only on the sample size, and in Section 4.4 we study the case of twice-differentiable loss functions.
- The sufficient conditions established give room to choose the generalized empirical measure in various ways, and in Section 5 we propose two methods for learning the covariance matrix for Gaussian bolstering from the training data.
- In Section 6, we empirically assess the performance of the proposed error estimators in polynomial regression experiments, and in Section 7 we provide concluding remarks.

Proofs for the results are presented in Appendix C. In Appendices B, D and E we discuss the VC dimension of hypothesis spaces under general loss functions and prove some auxiliary results that characterize the loss functions and hypothesis spaces for which some of the established sufficient conditions hold. Detailed results of the experiments are presented in Appendix A.

2. Background

Let the pair of random vectors (X, Y) take values in a product space $\mathcal{X} \times \mathcal{Y}$. In most cases of interest, $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}^m$, where usually $m = 1$ (though this is not necessary). The pair (X, Y) is jointly distributed with an unknown probability law ν on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_{\mathcal{X}\mathcal{Y}})$, where $\mathcal{B}_{\mathcal{X}\mathcal{Y}}$ denotes the Borel σ-algebra in $\mathcal{X} \times \mathcal{Y}$. Random variables X and Y are the *feature vector* and *target*, respectively, in a statistical learning problem. The *hypothesis*

space $\mathcal{H}$ is a set of Borel-measurable functions $\psi : \mathcal{X} \to \mathcal{Y}$. For example, in a classification problem, $\mathcal{Y} = \{0, 1, \ldots, c-1\}$ and ψ is called a *classifier*, whereas in a standard regression problem, $\mathcal{Y} = \mathbb{R}$ and ψ is called a *regression function*.

Each hypothesis ψ in $\mathcal{H}$ has an error defined as

$$L(\psi) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, \psi(x))\ d\nu(x, y)$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ is an appropriate *loss function*. For example, with $\mathcal{Y} = \mathbb{R}$, the *quadratic loss* $\ell(y, \psi(x)) = (y - \psi(x))^2$ yields the regression mean squared error, whereas with $\mathcal{Y} = \{0, 1, \ldots, c-1\}$, the *misclassification loss* $\ell(y, \psi(x)) = \mathbb{1}_{y\neq\psi(x)}$ yields the probability of classification error. This error is unknown, and in order to estimate it in practice, one collects an i.i.d. *training sample* of (X, Y)

$$S_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

and considers the empirical error

$$L_n(\psi) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, \psi(X_i))$$

as an estimator of $L(\psi)$.

Solving a learning problem in this instance means picking a hypothesis ψ_n in $\mathcal{H}$ according to a given empirical criterion based on a sample S_n. This task can be described as a prediction rule $\Psi_n : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$ which associates each sample $S_n \in (\mathcal{X} \times \mathcal{Y})^n$ to a hypothesis $\psi_n := \Psi_n(S_n) \in \mathcal{H}$. Observe that ψ_n is random since it is a function of the sample S_n.

The quantity of interest in statistical learning is how well one can expect ψ_n to perform on data not in the sample, but generated by the same law as (X, Y), that is the prediction error of ψ_n, defined as

$$\varepsilon_n \ = \ L(\psi_n) \ = \ \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, \psi_n(x))\, d\nu(x, y)\,. \tag{1}$$

3. Generalized Resubstitution Error Estimators

A family of estimators of the prediction error ε_n in (1) is obtained by replacing the generally unknown probability measure ν in (1) with an *empirical measure* ν_n, which is a probability measure defined on $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_{\mathcal{X}\mathcal{Y}})$ that depends on the sample S_n:

$$\hat{\varepsilon}_n^{\nu_n} \ = \ \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, \psi_n(x))\, d\nu_n(x, y)\,. \tag{2}$$

Following Ghane and Braga-Neto (2022), we call these *generalized resubstitution error estimators*.

This general family of error estimators includes the estimator generated by the standard empirical measure, which puts discrete mass $1/n$ on each data point:

$$\nu_n \;=\; \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i,Y_i}\,, \tag{3}$$

where δ_{X_i,Y_i} is the (random) point measure located at (X_i, Y_i). Substituting (3) into (2) yields the empirical error $L_n(\psi_n)$ of ψ_n on S_n, which is known as the *resubstitution error estimator*:

$$\hat{\varepsilon}_n^{\,r} \;=\; L_n(\psi_n) \;=\; \frac{1}{n}\sum_{i=1}^{n}\ell(Y_i,\psi_n(X_i))\,. \tag{4}$$

For example, with $\mathcal{Y}=\mathbb{R}$ and the quadratic loss, this is the sum of squared errors in regression, while with $\mathcal{Y}=\{0,1,\ldots,c-1\}$ and the misclassification loss, this is the training error of a classifier.

We restrict our attention in this paper to a specific, yet broad, class of generalized resubstitution estimators, which are based on the family of empirical measures of the form:

$$\nu_n \;=\; \frac{1}{n}\sum_{i=1}^{n}\beta_{n,i}, \tag{5}$$

where $\beta_{n,i}$ is a measure on $(\mathcal{X}\times\mathcal{Y},\mathcal{B}_{\mathcal{X}\mathcal{Y}})$, for $i=1,\ldots,n$. Although $\beta_{n,i}$ may depend on the entire sample S_n, we assume that it has a special dependence on X_i, Y_i and ψ_n, and we may use the explicit notation $\beta_{X_i,Y_i,\psi_n,S_n}$ to make this clear. When $\beta_{n,i}=\delta_{X_i,Y_i}$ we recover the standard resubstitution error estimator.

A special case of interest is when $\beta_{n,i}$ is a product measure,

$$\beta_{n,i} = \mu_{n,i}\,\pi_{n,i}\,, \tag{6}$$

where $\mu_{n,i}$ and $\pi_{n,i}$ are empirical measures on $(\mathcal{X},\mathcal{B}_{\mathcal{X}})$ and $(\mathcal{Y},\mathcal{B}_{\mathcal{Y}})$, respectively. In this instance, substituting (5) into (2) and using Fubini's Theorem yields the following generalized resubstitution error estimator:

$$\begin{aligned}\hat{\varepsilon}_n^{\,\nu_n} \;&=\; \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{X}\times\mathcal{Y}}\ell(y,\psi_n(x))\,d\beta_{n,i}(x,y)\\ &=\; \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{Y}}\left(\int_{\mathcal{X}}\ell(y,\psi_n(x))\,d\mu_{n,i}(x)\right)d\pi_{n,i}(y)\end{aligned} \tag{7}$$

The standard resubstitution error estimator in (4) is a special case of (7), with $\mu_{n,i} = \delta_{X_i}$, $\pi_{n,i} = \delta_{Y_i}$, and $\beta_{n,i} = \delta_{X_i}\delta_{Y_i} = \delta_{X_i,Y_i}$. One can interpret the empirical measures $\mu_{n,i}$, $\pi_{n,i}$, and $\beta_{n,i}$ as *smoothed* versions of the point measures δ_{X_i}, δ_{Y_i}, and δ_{X_i,Y_i}, respectively.

Specific examples of generalized resubstitution estimators are given in the next few subsections.

3.1 Bolstered Resubstitution (smoothing only in the "X direction")

In this case, $\beta_{n,i} = \mu_{n,i}\,\pi_{n,i}$ is the product measure (6), where $\pi_{n,i} = \delta_{Y_i}$ and $\mu_{n,i}$ is a general empirical measure, in which case (7) becomes:

$$\hat{\varepsilon}_n^{br} = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{X}} \ell(Y_i, \psi_n(x))\, d\mu_{n,i}(x)\,. \tag{8}$$

This is the *bolstered resubstitution* error estimator, which was proposed in Braga-Neto and Dougherty (2004) for the case of binary classification, here extended to the general statistical learning case. Notice that this estimator performs smoothing in the "X direction".

Usually, $\mu_{n,i}$ is assumed to be absolutely continuous, with a probability density function $p_{n,i}(x)$. If in addition we assume the quadratic loss, (8) becomes

$$\hat{\varepsilon}_n^{br} = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}^d} (\psi_n(x) - Y_i)^2\; p_{n,i}(x)\; dx \tag{9}$$

In practice, the density $p_{n,i}(x)$ is centered on the data point X_i and it is called a *bolstering kernel*. Then the contribution of each point (X_i, Y_i) to $\hat{\varepsilon}_n^{br}$ is given by assessing how well $\psi_n(x)$ approximates y when x is a random perturbation of X_i, given by $p_{n,i}(x)$, and $y = Y_i$. This is illustrated in Figure 1, with a Gaussian density $p_{n,i}$ centered at each X_i.

3.2 Posterior-Probability Error Estimator (smoothing only in the "Y direction")

In this case, $\mu_{n,i} = \delta_{X_i}$ and $\pi_{n,i}$ is a general empirical measure so (7) becomes:

$$\hat{\varepsilon}_n^{\nu_n} = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{Y}} \ell(y, \psi_n(X_i))\, d\pi_{n,i}(y)\,. \tag{10}$$

A special case in classification problems is the *posterior-probability generalized resubstitution error estimator* in which $\pi_{n,i}$ is a posterior probability of

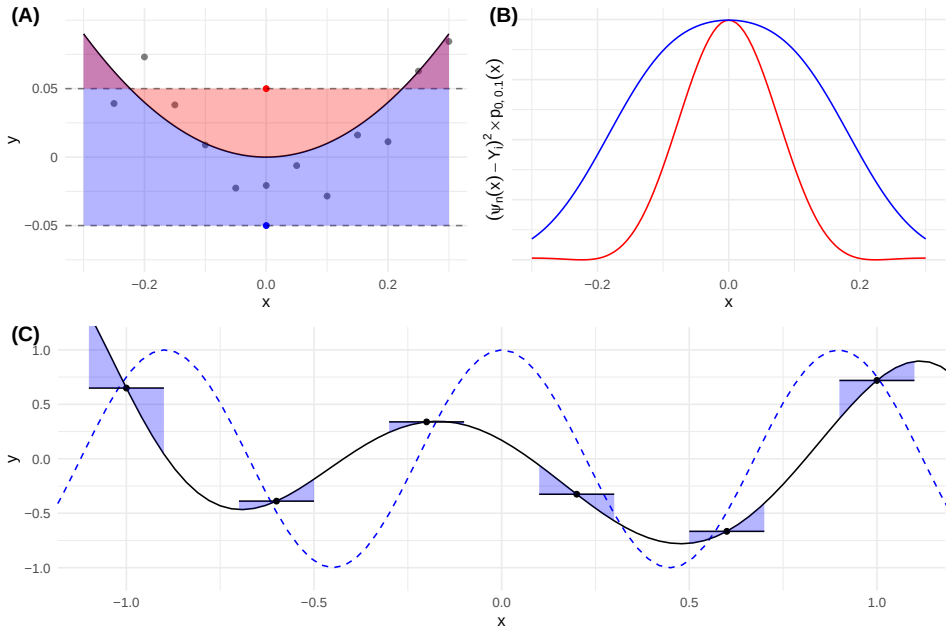

Figure 1: An illustration of Gaussian bolstering for regression. **(A)** An example in which $\psi_n(x) = x^2$ with the points $(0, 0.05)$ and $(0, -0.05)$ outlined in red and blue, respectively. The predictor ψ_n has the same quadratic loss at both points. The shaded area represents the distance from $\psi_n(x)$ to 0.05 (red) and -0.05 (blue). **(B)** The squared distance from $\psi_n(x)$ to 0.05 (red) and -0.05 (blue) in **(A)** weighted by a Gaussian density with mean zero and standard deviation 0.1. The contribution of each point to the Gaussian bolstering estimator (9) is the area under the respective curve, and hence that of the blue point is greater. **(C)** Points (x, y) generated by $y = \psi^\star(x) + \epsilon$, in which $\psi^\star$ is the dashed curve in blue and ϵ is a Gaussian noise. The shaded regions represent $X_i \pm 3\sigma$ in which σ is the standard deviation of the Gaussian bolstering distribution. The polynomial ψ_n (black) interpolates the data, so its resubstitution error is zero. Nevertheless, the Gaussian bolstering error estimator is not zero and equals essentially the mean distance from $\psi_n(x)$ to Y_i for x in the shaded neighborhood of X_i when x is distributed as a Gaussian distribution with mean X_i and standard deviation σ.

Y conditioned on $X = X_i$:

$$\pi_{n,i}(y) = \hat{P}_n\left(Y = y | X = X_i\right)$$

where $\hat{P}_n$ is a posterior probability in $\mathcal{Y} = \{0, \ldots, c-1\}$, so (10) reduces to

$$\hat{\varepsilon}_n^{ppr} = \frac{1}{n}\sum_{i=1}^{n} \hat{E}_n\left(\ell(Y, \psi_n(X_i)) | X = X_i\right)$$

in which $\hat{E}_n$ is the respective posterior expectation. In the same manner, one could consider $\hat{\varepsilon}_n^{ppr}$ in regression by considering a posterior probability density function as the density of the absolutely continuous measure $\pi_{n,i}$. In Figure 2 we illustrate posterior-probability resubstitution for Bayesian regression.

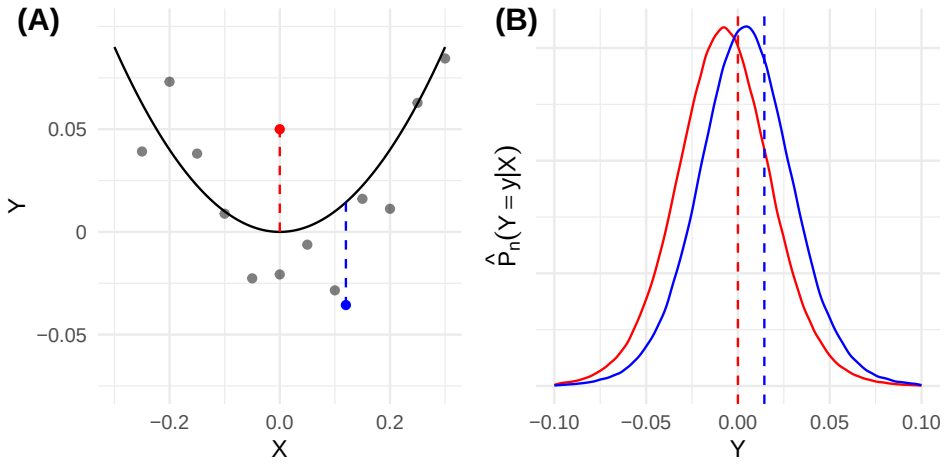

Figure 2: An illustration of posterior-probability resubstitution for Bayesian regression. **(A)** An example in which $\psi_n(x) = x^2$ with the points $(0, 0.05)$ and $(0.12, -0.0356)$ outlined in red and blue, respectively. The predictor ψ_n has the same quadratic loss at both of these points. **(B)** The posterior probability of Y given $X_i = 0$ (red) and $X_i = 0.12$ (blue). The vertical dashed lines represent the respective value of $\psi_n(X_i)$. The contribution of each point to $\hat{\varepsilon}_n^{ppr}$ is the expected value of $(\psi_n(X_i) - Y)^2$ under the respective posterior probability, that is, the expected squared distance from Y to the respective dashed line. The contribution of the blue point is greater than that of the red one.

3.3 Smoothing in "both directions"

This is the general case, where neither $\mu_{n,i}$ nor $\pi_{n,i}$ are point measures. For example, one could combine bolstered and posterior-probability resubstitution by considering $\pi_{n,i}$ as a posterior distribution of Y conditioned on $X = X_i$ and $\mu_{n,i}$ with general bolstering kernel $p_{n,i}$.

A more general setting is when $\beta_{n,i}$ is not a product measure. An important case is when $\beta_{n,i}$ is a Gaussian distribution with mean vector (X_i, Y_i) and a possibly non-spherical covariance matrix Σ_i, as exemplified in Figure 3. We call this special case the *XY-Gaussian bolstering* error estimator.

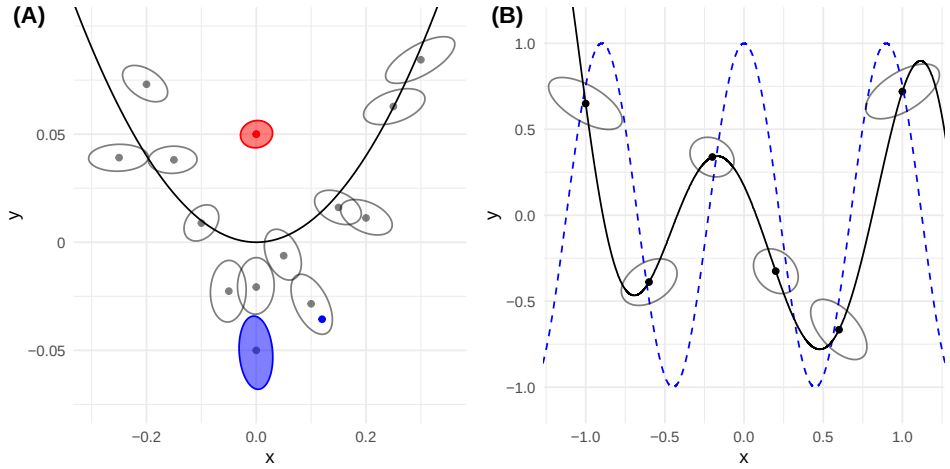

Figure 3: An illustration of the XY-Gaussian bolstering error estimator with the same data of Figure 1. The contribution of each data point (X_i, Y_i) to the bolstered error estimator is the expected squared distance of Y to $\psi_n(X)$ for (X, Y) distributed as a Gaussian distribution with mean (X_i, Y_i) and a covariance matrix Σ_i. The ellipses in **(A)** and **(B)** are level curves of the respective Gaussian distribution and illustrate the form of Σ_i.

3.4 Notation and assumptions

In order to avoid heavy notation, we denote $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and define $Z = (X, Y)$ as the random vector taking values in $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}})$ with probability law ν. Following this notation, the sample becomes $S_n = \{Z_1, \ldots, Z_n\}$. Furthermore, for $z = (x, y) \in \mathcal{Z}$ and $\psi \in \mathcal{H}$ we denote by $\ell(z, \psi)$ or $\ell((x, y), \psi)$ the loss previously defined as $\ell(y, \psi(x))$.

Let $C_\ell > 0$ be defined as

$$C_\ell := \sup_{z \in \mathcal{Z}, \psi \in \mathcal{H}} \ell(z, \psi).$$

If $C_\ell < \infty$ then ℓ is a bounded loss function, and if $C_\ell = \infty$ then it is an unbounded loss function. We assume that ℓ has a finite moment of an order $\alpha > 1$ under ν, for all $\psi \in \mathcal{H}$. Formally, denoting

$$L^\alpha(\psi) := \int_{\mathcal{Z}} \ell^\alpha(z, \psi) \, d\nu(z),$$

we assume there exists an $\alpha > 1$ such that

$$\sup_{\psi \in \mathcal{H}} L^\alpha(\psi) < \infty. \tag{11}$$

If ℓ is bounded, then (11) holds for all $\alpha > 1$, hence in this case, any measure ν satisfies (11). When the loss function is unbounded, condition (11) defines a constrained class of measures.

In the next section, we state the main results of this paper about the consistency of resubstitution generalized error estimators.

4. Consistency of generalized resubstitution error estimators

We formally define the generalized resubstitution error estimators, extending the concept proposed by Ghane and Braga-Neto (2022) for classification to the general statistical learning framework.

Definition 1 *Fix a hypothesis space $\mathcal{H}$, a loss function ℓ and a prediction rule Ψ_n. For each $n \geqslant 1$ and sample $S_n \in \mathcal{Z}^n$, let*

$$\mathcal{B}(S_n) = \{\beta_{z,\psi,S_n} : z \in \mathcal{Z}, \psi \in \mathcal{H}\}$$

be a collection of smoothing measures defined on $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}})$ and consider the loss function $\ell_{\mathcal{B}}$ given by

$$\ell_{\mathcal{B}}(z, \psi) = \int_{\mathcal{Z}} \ell(z', \psi)\ d\beta_{z,\psi,S_n}(z')$$

for $z \in \mathcal{Z}$ and $\psi \in \mathcal{H}$. We define the $\mathcal{B}$-generalized resubstitution error of ψ_n as

$$\hat{\varepsilon}_n^{\mathcal{B}} = \frac{1}{n}\sum_{i=1}^{n} \ell_{\mathcal{B}}(Z_i, \psi_n) = \frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{Z}} \ell(z', \psi_n)\ d\beta_{Z_i,\psi_n,S_n}(z').$$

When $\mathcal{B}(S_n)$ depends on S_n only through its size n we denote it by $\mathcal{B}_n$, its measures by $\beta_{z,\psi,n}$, the generated loss by $\ell_{\mathcal{B}_n}$ and the estimator by $\hat{\varepsilon}_n^{\mathcal{B}_n}$.

The estimator $\hat{\varepsilon}_n^{\mathcal{B}}$ is the standard resubstitution error of ψ_n under the *generalized loss function* $\ell_{\mathcal{B}}$ which is a smoothed version of the original loss ℓ. For each sample S_n, the collection $\mathcal{B}(S_n)$ generates the empirical measure

$$\nu_n = \frac{1}{n}\sum_{i=1}^{n} \beta_{Z_i,\psi_n,S_n} \tag{12}$$

so $\hat{\varepsilon}_n^{\mathcal{B}}$ coincides with definition (2) for ν_n of form (12). Analogous to the examples in Section 2, the loss function $\ell_{\mathcal{B}}$ can be a smoothing of ℓ on the X, Y or both directions by considering $\beta_{z,\psi_n,S_n} = \mu_{z,\psi_n,S_n}\pi_{z,\psi_n,S_n}$ as the product of suitable measures on $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}})$, respectively. Generalized error

estimators can be defined equivalently by smoothing the empirical measure or the loss function, but, even though definition (2) might be more intuitive, the main results of this paper are better stated and proved considering Definition 1.

A special case of generalized resubstitution error is when the collection $\mathcal{B}_n$ depends on S_n only through its size. This is the case, for example, of Gaussian bolstering when each Gaussian distribution has a covariance matrix $K_{i,n}$ that depends only on the data points X_i and Y_i, and on the sample size n, e.g., by converging to zero when n increases. On the other hand, this is not the case for the posterior-probability error estimator when the posterior distribution is generated by the sample S_n or for Gaussian bolstering when the kernel is estimated from S_n.

The importance of this dependence lies in the fact that, when the smoothing measures depend on S_n only through its size, the loss $\ell_{\mathcal{B}_n}(z, \psi)$ is not random and classical deviation-bounds of statistical learning theory can be applied to its expectation under the data-generating distribution. Indeed, for each $\psi \in \mathcal{H}$, let

$$L^{\mathcal{B}}(\psi) = \int_{\mathcal{Z}} \ell_{\mathcal{B}}(z, \psi)\, d\nu(z) \qquad \text{and} \qquad L_n^{\mathcal{B}}(\psi) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\mathcal{B}}(Z_i, \psi)$$

be the expected loss $\ell_{\mathcal{B}}$ under the data-generating distribution and on the sample S_n, respectively. If the collection of measures depends on S_n only through its size, then $L^{\mathcal{B}_n}(\psi)$ is not a random quantity, while $L_n^{\mathcal{B}}(\psi)$ is always random.

4.1 Consistency of generalized resubstitution error estimators

In this paper, we are concerned with the consistency of generalized resubstitution error estimators, which is characterized by the convergence of the estimated error to $\varepsilon_n = L(\psi_n)$ with probability one as the sample size increases. Observe that ε_n is a random variable since it depends on ψ_n, which is random.

Definition 2 *Fix a hypothesis space $\mathcal{H}$, a loss function ℓ, a prediction rule Ψ_n and a collection $\mathcal{B}(S_n)$ of probability measures for each sample $S_n \in \mathcal{Z}^n$. The generalized resubstitution error $\hat{\varepsilon}_n^{\mathcal{B}}$ is consistent if*

$$\left|\hat{\varepsilon}_n^{\mathcal{B}} - \varepsilon_n\right| \to 0$$

with probability one as $n \to \infty$.

Assume that the random vectors in the sample S_n are defined on a probability space $(\Omega, \mathcal{S}, \mathbb{P})$. Denote expectation in this probability space by $\mathbb{E}$. Define $\varepsilon_n^{\mathcal{B}} := L^{\mathcal{B}}(\psi_n)$ and observe that $\hat{\varepsilon}_n^{\mathcal{B}} = L_n^{\mathcal{B}}(\psi_n)$. We assume there exists a constant C such that, for all $n \geqslant 1$,

$$\mathbb{P}\left(L_n^{\mathcal{B}}(\psi_n) < C\right) = 1. \tag{13}$$

Under assumption (13), if $\hat{\varepsilon}_n^{\mathcal{B}}$ is consistent, then it is asymptotically unbiased.

Proposition 3 *Fix a loss function ℓ, a prediction rule Ψ_n and a collection $\mathcal{B}(S_n)$ of probability measures for each sample $S_n \in \mathcal{Z}^n$. If $\hat{\varepsilon}_n^{\mathcal{B}}$ is consistent and* (13) *holds, then*

$$\lim_{n\to\infty} |\mathbb{E}[\hat{\varepsilon}_n^{\mathcal{B}}] - \mathbb{E}[\varepsilon_n]| = 0. \tag{14}$$

The next theorem states that the resubstitution error $\hat{\varepsilon}_n^r$ is consistent if $\mathcal{H}$ has finite VC dimension under the loss function ℓ. See Appendix B for the formal definition of this VC dimension, which is also known as the pseudo-dimension (Pollard, 1989; Vapnik, 1998).

Theorem 4 *Fix a loss function ℓ and let $\mathcal{H}$ be a hypothesis space with $d_{VC}(\mathcal{H}, \ell) < \infty$. For any prediction rule Ψ_n, the resubstitution error $\hat{\varepsilon}_n^r$ is consistent.*

In Section 4.2, we present general sufficient conditions for the consistency of generalized resubstitution errors, in Section 4.3 we study the consistency when $\mathcal{B}_n$ depends only on the sample size, and in Section 4.4 we study the consistency when the loss function is twice differentiable in z.

4.2 Sufficient conditions for consistency: general case

The first sufficient condition for the consistency of $\hat{\varepsilon}_n^{\mathcal{B}}$ is the convergence of $\hat{\varepsilon}_n^{\mathcal{B}}$ to $\hat{\varepsilon}_n^r$, which is a direct consequence of Theorem 4.

Proposition 5 *Fix a loss function ℓ, a prediction rule Ψ_n and a collection $\mathcal{B}(S_n)$ of probability measures for each sample $S_n \in \mathcal{Z}^n$. Let $\mathcal{H}$ be a hypothesis space with $d_{VC}(\mathcal{H}, \ell) < \infty$. If it holds*

$$|\hat{\varepsilon}_n^{\mathcal{B}} - \hat{\varepsilon}_n^r| \to 0$$

with probability one as $n \to \infty$, then $\hat{\varepsilon}_n^{\mathcal{B}}$ is consistent.

For a fixed loss function ℓ, $\psi \in \mathcal{H}$ and $b \in (0, C_\ell)$ define

$$A_{\ell,\psi,b} = \{z \in \mathcal{Z} : \ell(z, \psi) > b\} \in \mathcal{B}_{\mathcal{Z}}$$

as the points in $\mathcal{Z}$ where $\ell(z, \psi) > b$ and let

$$\mathcal{A}^{\star}_{\mathcal{H},\ell} = \{A_{\ell,\psi,b} : \psi \in \mathcal{H}, b \in (0, C_\ell)\} \subset \mathcal{B}_{\mathcal{Z}}$$

be the collection of such sets for $\psi \in \mathcal{H}$ and $b \in (0, C_\ell)$. An extension of Theorem 1 in Ghane and Braga-Neto (2022) follows from Proposition 5 for bounded loss functions.

Corollary 6 *Fix a bounded loss function ℓ, a prediction rule Ψ_n and a collection $\mathcal{B}(S_n)$ of probability measures for each sample $S_n \in \mathcal{Z}^n$. Let $\mathcal{H}$ be a hypothesis space with $d_{VC}(\mathcal{H}, \ell) < \infty$. If*

$$\sup_{A \in \mathcal{A}^{\star}_{\mathcal{H},\ell}} \left| \frac{1}{n} \sum_{i=1}^{n} \beta_{Z_i,\psi_n,S_n}(A) - \frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i}(A) \right| \to 0$$

with probability one as $n \to \infty$, then $\hat{\varepsilon}_n^{\mathcal{B}}$ is consistent. In particular, if

$$\sup_{\substack{A \in \mathcal{A}^{\star}_{\mathcal{H},\ell} \\ z \in \mathcal{Z}, \psi \in \mathcal{H}}} |\beta_{z,\psi,S_n}(A) - \delta_z(A)| \to 0$$

with probability one as $n \to \infty$, then $\hat{\varepsilon}_n^{\mathcal{B}}$ is consistent.

If $\ell_{\mathcal{B}}(z, \psi)$ converges to $\ell(z, \psi)$ uniformly on $\mathcal{Z}$ and $\mathcal{H}$, then $\hat{\varepsilon}_n^{\mathcal{B}}$ is consistent. This is a sufficient condition that also holds for unbounded loss functions.

Corollary 7 *Fix a loss function ℓ, a prediction rule Ψ_n, a collection $\mathcal{B}(S_n)$ of probability measures for each sample $S_n \in \mathcal{Z}^n$, and let $\mathcal{H}$ be a hypothesis space with $d_{VC}(\mathcal{H}, \ell) < \infty$. If*

$$\lim_{n \to \infty} \sup_{z \in \mathcal{Z}, \psi \in \mathcal{H}} |\ell_{\mathcal{B}}(z, \psi) - \ell(z, \psi)| = 0,$$

with probability one, then $\hat{\varepsilon}_n^{\mathcal{B}}$ is consistent.

4.3 Sufficient conditions for consistency: dependence only on sample size

When the collection $\mathcal{B}_n$ depends on the sample S_n only through its size n, then

$$\hat{\varepsilon}_n^{\mathcal{B}_n} = \frac{1}{n}\sum_{i=1}^{n} \ell_{\mathcal{B}_n}(Z_i, \psi_n)$$

is the resubstitution error of ψ_n under the loss function $\ell_{\mathcal{B}_n}$. It follows from Theorem 4 that $\hat{\varepsilon}_n^{\mathcal{B}_n}$ converges with probability one to $\varepsilon_n^{\mathcal{B}_n}$ if $\limsup d_{VC}(\mathcal{H}, \ell_{\mathcal{B}_n}) < \infty$. Therefore, if $\varepsilon_n^{\mathcal{B}_n}$ converges to ε_n, then $\hat{\varepsilon}_n^{\mathcal{B}_n}$ converges to ε_n, so it is consistent. This result is stated in Proposition 8. In Appendix E.2 we present sufficient conditions for $\limsup d_{VC}(\mathcal{H}, \ell_{\mathcal{B}_n})$ to be finite when $d_{VC}(\mathcal{H}, \ell) < \infty$.

Proposition 8 *Fix a loss function ℓ, a prediction rule Ψ_n and a collection $\mathcal{B}_n$ of probability measures for each $n \geqslant 1$. Let $\mathcal{H}$ be a hypothesis space with $\limsup_{n\to\infty} d_{VC}(\mathcal{H}, \ell_{\mathcal{B}_n}) < \infty$. If*

$$\left|\varepsilon_n^{\mathcal{B}_n} - \varepsilon_n\right| \to 0$$

with probability one as $n \to \infty$, then $\hat{\varepsilon}_n^{\mathcal{B}_n}$ is consistent.

As a consequence of Proposition 8, the next proposition gives a sufficient condition when $\beta_{z,\psi,n}$ is absolutely continuous wrt the data-generating measure ν.

Proposition 9 *Fix a loss function ℓ, a prediction rule Ψ_n, a collection $\mathcal{B}_n$ of probability measures for each $n \geqslant 1$, and let $\mathcal{H}$ be a hypothesis space with $\limsup_{n\to\infty} d_{VC}(\mathcal{H}, \ell_{\mathcal{B}_n}) < \infty$. If*

(a) For all $z \in \mathcal{Z}, \psi \in \mathcal{H}$ and $n \geqslant 1$, $\beta_{z,\psi,n}$ is absolutely continuous with respect to the data-generating measure ν. Denote by $\rho_{z,\psi,n} = \frac{d\beta_{z,\psi,n}}{d\nu}$ a version of the respective Radon-Nikodym derivative;

(b) For all $\psi \in \mathcal{H}$, $n \geqslant 1$ and $z' \in \mathcal{Z}$ it holds

$$\int_{\mathcal{Z}} \rho_{z,\psi,n}(z')\, d\nu(z) = 1;$$

then $\hat{\varepsilon}_n^{\mathcal{B}_n}$ is consistent.

It follows from Proposition 9 that if $\rho_{z,\psi,n}(z')$ is a symmetric function of (z, z') with ν-probability one, then $\hat{\varepsilon}_n^{\mathcal{B}_n}$ is consistent.

Corollary 10 *Fix a loss function ℓ, a prediction rule Ψ_n, a collection $\mathcal{B}_n$ of probability measures for each $n \geqslant 1$, and let $\mathcal{H}$ be a hypothesis space with $\limsup_{n\to\infty} d_{VC}(\mathcal{H}, \ell_{\mathcal{B}_n}) < \infty$. If (a) in Proposition 9 holds and*

$$(\nu \times \nu)\left(\{(z, z') \in \mathcal{Z}^2 : \rho_{z,\psi,n}(z') = \rho_{z',\psi,n}(z)\}\right) = 1 \tag{15}$$

for all $\psi \in \mathcal{H}$ and all $n \geqslant 1$, then $\hat{\varepsilon}_n^{\mathcal{B}_n}$ is consistent.

Corollary 10 can be applied to show the consistency of Gaussian bolstering when the covariance matrix of the Gaussian distributions is the same in all sample points.

Example 1 (Gaussian bolstering) Assume that X is absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}^d$ and consider the Gaussian bolstering error estimator

$$\hat{\varepsilon}_n^{gbr} := \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}^d} \ell((x, Y_i), \psi)\; p_{X_i,\psi_n,n}(x)\; dx \tag{16}$$

in which $p_{X_i,\psi_n,n}$ is the density function of a Gaussian distribution with mean X_i and covariance matrix K_{n,ψ_n}, which may depend on n and ψ_n. In this case, $\mathcal{B}_n = \{\beta_{x,y,\psi,n} : (x, y) \in \mathcal{Z}, \psi \in \mathcal{H}\}$ is a collection of measures $\beta_{x,y,\psi,n} = \mu_{x,\psi,n}\delta_y$ in which $\mu_{x,\psi,n}$ is a Gaussian measure with mean x and covariance matrix $K_{n,\psi}$.

The Gaussian bolstering error estimator (16) is consistent if

$$\limsup_n d_{VC}(\mathcal{H}, \ell_{\mathcal{B}_n}) < \infty,$$

a condition which holds, for example, for linear regression under the quadratic loss and in classification problems under binary loss functions if

$$\limsup_{n\to\infty} |\ell_{\mathcal{B}_n}(z, \psi) - \ell(z, \psi)| < 1/2$$

for all $z \in \mathcal{Z}$ and $\psi \in \mathcal{H}$ (cf. Appendix E.2).

Proposition 11 *If $\limsup_{n\to\infty} d_{VC}(\mathcal{H}, \ell_{\mathcal{B}_n}) < \infty$, then the Gaussian bolstering error estimator $\hat{\varepsilon}_n^{gbr}$ is consistent.*

Proposition 11 establishes the consistency of the Gaussian bolstering error estimator in classification and regression scenarios in which the kernel does not necessarily converge to zero. This is an improvement from the results in Ghane and Braga-Neto (2022) which required the kernel to converge to zero. This proposition also holds for XY-Gaussian bolstering.

■

4.4 Sufficient conditions for consistency: twice differentiable loss functions

More specific sufficient conditions for the consistency of generalized resubstitution error estimators may be obtained under the assumption that $\ell((x,y),\psi)$ is twice differentiable in (x,y) and its second derivatives are uniformly bounded on $\mathcal{X}\times\mathcal{Y}$ and $\mathcal{H}$. Important examples of this case consider ℓ as the quadratic loss function and $\psi(x)$ as twice differentiable functions.

Denote by Z_{z,ψ,S_n} a random variable with a probability law β_{z,ψ,S_n} for $z\in\mathcal{Z}$, $\psi\in\mathcal{H}$ and $S_n\in\mathcal{Z}^n$. If $\ell(z,\psi)$ is twice differentiable in z and its second derivatives are uniformly bounded in $\mathcal{Z}$ and $\mathcal{H}$, then $\hat{\varepsilon}_n^{\mathcal{B}}$ is consistent if the expectation of Z_{z,ψ,S_n} is z and if the variance of its coordinates converges to zero as n increases with probability one over the possible samples. This result is a consequence of Corollary 7. Recall that $Z=(X,Y)$ is a random vector with $d+m$ coordinates.

Proposition 12 *Fix a loss function ℓ, a prediction rule Ψ_n, a collection $\mathcal{B}(S_n)$ of probability measures for each sample $S_n\in\mathcal{Z}^n$, and let $\mathcal{H}$ be a hypothesis space with $d_{VC}(\mathcal{H},\ell)<\infty$. If*

(a) $\ell(\cdot,\psi)\in C^2(\mathcal{Z})$ for all $\psi\in\mathcal{H}$ and there exists a constant C_2 such that

$$\sup_{z\in\mathcal{Z},\psi\in\mathcal{H}}\left|\frac{\partial^2\ell}{\partial z_i\partial z_j}(z,\psi)\right|<C_2 \tag{17}$$

for all $i,j=1,\ldots,d+m$;

(b) For all $z\in\mathcal{Z},\psi\in\mathcal{H}$ and $S_n\in\mathcal{Z}^n$ it holds $\mathbb{E}(Z_{z,\psi,S_n})=z$ and

$$\lim_{n\to\infty}\sup_{z\in\mathcal{Z},\psi\in\mathcal{H}} Var\left([Z_{z,\psi,S_n}]_j\right)=0$$

for all $j=1,\ldots,d+m$ with probability one;

then $\hat{\varepsilon}_n^{\mathcal{B}}$ is consistent.

We extend Theorem 2 of Ghane and Braga-Neto (2022) for Gaussian bolstering for regression under the quadratic loss function when ψ is twice differentiable and has derivatives uniformly bounded on $\mathcal{X}$ and $\mathcal{H}$. In this case, the Gaussian bolstering error estimator reduces to

$$\hat{\varepsilon}_n^{gbr} := \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}^d}\left[\psi_n(x) - Y_i\right]^2 p_{X_i,Y_i,\psi_n,S_n}(x)\; dx, \tag{18}$$

in which p_{X_i,Y_i,ψ_n,S_n} is the density function of a Gaussian distribution with mean X_i and covariance matrix K_{X_i,Y_i,ψ_n,S_n}. The next result presents a condition for the Gaussian bolstering error estimator to be consistent when ψ is twice differentiable. The result also holds for XY-Gaussian bolstering.

Proposition 13 *Fix a prediction rule Ψ_n and a collection $\mathcal{B}(S_n)$ of probability measures for each sample $S_n \in \mathcal{Z}^n$. Let ℓ be the quadratic loss function, let $\mathcal{Y}$ be a compact set of $\mathbb{R}$ and let $\mathcal{H} \subset C^2(\mathcal{X},\mathcal{Y})$ be a set of twice differentiable functions with $d_{VC}(\mathcal{H},\ell) < \infty$. If*

$$\lim_{n\to\infty}\sup_{x\in\mathcal{X},y\in\mathcal{Y},\psi\in\mathcal{H}}[K_{x,y,\psi,S_n}]_{j,j} = 0, \tag{19}$$

with probability one, and

$$\sup_{x\in\mathcal{X},\psi\in\mathcal{H}}\left|\frac{\partial\psi}{\partial x_i}(x)\right| < \infty \qquad \textit{and} \qquad \sup_{x\in\mathcal{X},\psi\in\mathcal{H}}\left|\frac{\partial^2\psi}{\partial x_i\partial x_j}(x)\right| < \infty$$

for all $i,j = 1,\ldots,d$, then the Gaussian bolstering error estimator $\hat{\varepsilon}_n^{gbr}$ defined in (18) *is consistent.*

Although Proposition 13 was stated for Gaussian bolstering, it holds for any bolstered error estimator by considering a probability density p_{x,y,ψ,S_n} with mean x and covariance matrix K_{x,y,ψ,S_n} that converges to zero with probability one when n increases.

5. Estimation of the kernel in Gaussian bolstering

The results in Section 4 outline a range of instances in which the Gaussian bolstering error estimator is consistent, allowing room to choose the covariance matrix (kernel) based on the training data S_n. In this section, we propose estimators for the Gaussian bolstering kernel based on the method of moments and on the maximization of a pseudo-likelihood function. The proposed estimators can also be used to estimate the kernel for Gaussian bolstering classification error estimation and in XY-Gaussian bolstering by applying them to the vector (X,Y) instead of X. Throughout this section, we consider that the sample S_n is fixed and that $\mathcal{X} = \mathbb{R}^d$.

5.1 Method of moments

We assume that the kernel has the form $K_{x,y,\psi,S_n} = \sigma^2_{S_n}\Sigma$ for a known fixed matrix Σ, and propose a method of moments estimator for σ_{S_n}. We also consider the case in which $K_{x,y,\psi,S_n} = \sigma^2_{y,S_n}\Sigma$ and estimate σ_{y,S_n} for each class $y \in \mathcal{Y}$ in a classification problem. The method of moments estimator is obtained following ideas analogous to that of Braga-Neto and Dougherty (2004).

For $x, x' \in \mathcal{X}$ denote by

$$\delta(x,x') = \sqrt{(x-x')^T\ \Sigma^{-1}\ (x-x')}$$

the *Mahalanobis distance* from x to x' and let

$$\delta(x) = \min_i \{\delta(x, X_i)\} \qquad\qquad \delta_y(x) = \min_{i:Y_i=y} \{\delta(x, X_i)\}$$

be the minimum distance from x to any X_i and to an X_i such that $Y_i = y \in \mathcal{Y}$, respectively. If there is no point in the sample with $Y_i = y$ we denote $\delta_y(x) = \infty$. When $\Sigma = I_d$ these distances reduce to the respective Euclidean distance.

On the one hand, the error $L(\psi_n)$ can be interpreted as the expected loss of ψ_n on a test point (X, Y) generated by ν. The random variable representing the distances above of this test point to the sample has a marginal and conditional cumulative distribution given $Y = y$, respectively,

$$D(\delta) = \nu\left(\delta(X) \leqslant \delta\right) \qquad \text{and} \qquad D_y(\delta) = \frac{\nu\left(\delta_y(X) \leqslant \delta, Y = y\right)}{\nu(Y=y)}$$

for $\delta > 0$ and $y \in \mathcal{Y}$ with $\nu(Y = y) > 0$.

On the other hand, $\hat{\varepsilon}_n^{gbr} = L_n^{\mathcal{B}}(\psi_n)$ can be interpreted as the expected loss of ψ_n on a test point $(\hat{X}, \hat{Y})$ generated by the empirical law ν_n with

$$\nu_n = \frac{1}{n}\sum_{i=1}^{n} \mu_{X_i,Y_i,S_n}\delta_{Y_i}$$

in which μ_{X_i,Y_i,S_n} is a Gaussian measure with mean X_i and covariance matrix of form $\sigma^2_{Y_i,S_n}\Sigma$, in which σ_{Y_i,S_n} may depend on Y_i. The random variable representing the above distances of this test point to the sample has marginal and conditional cumulative distribution given $\hat{Y} = y$, respectively,

$$\hat{D}(\delta) = \nu_n\left(\delta(\hat{X}) \leqslant \delta\right) \qquad \text{and} \qquad \hat{D}_y(\delta) = \frac{\nu_n\left(\delta_y(\hat{X}) \leqslant \delta, \hat{Y} = y\right)}{\nu_n(\hat{Y}=y)}$$

for $\delta > 0$ and $y = Y_i$ for some $i = 1, \ldots, n$.

Method of moments estimators can be obtained for the free parameter $\sigma^2_{S_n}$ by assuming it does not depend on y and then equating the sample mean of D with the mean of $\hat{D}$. In classification problems, we can also equate the sample mean of D_y and the mean of $\hat{D}_y$ and solve for a σ^2_{y,S_n} for each class y to obtain estimators that depend on the class y.

The sample mean of D and the sample mean of D_y are, respectively,

$$\bar{\delta}_{S_n} = \frac{1}{n} \sum_{i=1}^{n} \min_{j \neq i} \{\delta(X_i, X_j)\}, \text{ and}$$

$$\bar{\delta}_{y,S_n} = \frac{1}{\sum_{i=1}^{n} \mathbb{1}_{Y_i = y}} \sum_{i=1}^{n} \min_{j \neq i, Y_j = Y_i} \{\delta(X_i, X_j)\} \; \mathbb{1}_{Y_i = y},$$

that are the mean distance of the points in S_n and the points in S_n with $Y_i = y$, respectively, to the respective sample $S_n \backslash \{X_i, Y_i\}$. The next lemma presents the mean of $\delta(\hat{X})$ and of $\delta_y(\hat{X})|\hat{Y} = y$, that is the mean of $\hat{D}$ and $\hat{D}_y$, respectively. Denote by p_{i,S_n} and p_{i,y,S_n} the densities of Gaussian distributions with mean X_i and covariance matrices $\sigma^2_{S_n}\Sigma$ and $\sigma^2_{y,S_n}\Sigma$, respectively.

Lemma 14 *For all $y \in \mathcal{Y}$ such that $Y_i = y$ for some $i = 1, \ldots, n$,*

$$\mathbb{E}\left[\delta_y(\hat{X})|\hat{Y} = y\right] = \frac{1}{\sum_{i=1}^{n} \mathbb{1}_{Y_i = y}} \sum_{i: Y_i = y} \sum_{j: Y_j = y} \int_{\mathcal{X}} \delta(x, X_j) \; p_{i,y,S_n}(x) \; \mathbb{1}_{\delta(x) = \delta(x, X_j)} \; dx$$

$$\mathbb{E}\left[\delta(\hat{X})\right] = \frac{1}{n} \sum_{i,j=1}^{n} \int_{\mathcal{X}} \delta(x, X_j) \; p_{i,S_n}(x) \; \mathbb{1}_{\delta(x) = \delta(x, X_j)} \; dx.$$

A method of moments estimator for σ_{S_n} and σ_{y,S_n} may be obtained by solving the equations

$$\mathbb{E}\left[\delta(\hat{X})\right] = \bar{\delta}_{S_n} \qquad \text{and} \qquad \mathbb{E}\left[\delta_y(\hat{X})|\hat{Y} = y\right] = \bar{\delta}_{y,S_n} \tag{20}$$

on σ_{S_n} and σ_{y,S_n}, respectively. For lower-dimensional data, the integrals in Lemma 14 may be solved exactly or via Monte Carlo integration. In Figure 4 we illustrate the estimation of σ_{S_n} by solving (20) via a grid search.

Since solving (20) may be computationally complex, especially for high-dimensional data, we propose an approximation for the expectations in Lemma 14 that implies estimators that are analogous to that proposed by Braga-Neto and Dougherty (2004). The next proposition states that, when

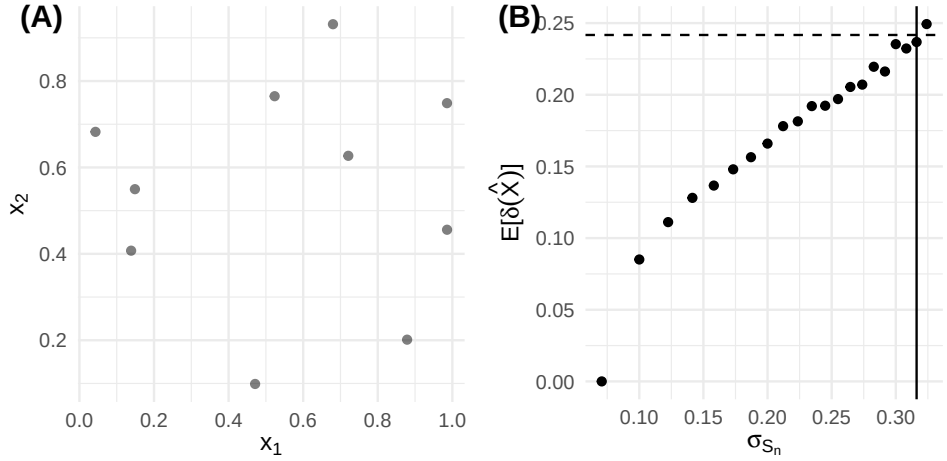

Figure 4: Illustration of the estimation of σ_{S_n} by the method of moments. **(A)** A sample of X in $d = 2$ dimensions with $\bar{\delta}_{S_n} = 0.24$. **(B)** A grid search to estimate σ_{S_n} by solving equation (20) considering $\Sigma = I_d$. The expectation $\mathbb{E}[\delta(\hat{X})]$ is computed via Monte Carlo integration for σ_{S_n} in a grid and the estimate $\hat{\sigma}_{S_n}$ is that with $\mathbb{E}[\delta(\hat{X})]$ closer to $\bar{\delta}_{S_n}$ that was $\hat{\sigma}_{S_n} = 0.31$.

σ_{S_n} converges to zero, $\sigma_{S_n}^{-1}\mathbb{E}[\delta(\hat{X})]$ converges to the expectation of a *chi* random variable with d degrees of freedom, recalling that d is the dimension of X. The same result follows for $\sigma_{y,S_n}^{-1}\mathbb{E}[\delta_y(\hat{X})|\hat{Y} = y]$.

Proposition 15 *Let χ be a random variable with a chi distribution with d degrees of freedom. Then,*

$$\frac{\mathbb{E}[\delta(\hat{X})]}{\sigma_{S_n}} \leqslant \mathbb{E}[\chi] \qquad \textit{and} \qquad \lim_{\sigma_{S_n}\to 0} \frac{\mathbb{E}[\delta(\hat{X})]}{\sigma_{S_n}} = \mathbb{E}[\chi]$$

for any fixed sample $S_n \in \mathcal{Z}^n$.

Proposition 15 yields the approximate method of moments estimators

$$\hat{\sigma}_{S_n} = \frac{\bar{\delta}_{S_n}}{\mathbb{E}[\chi]} \qquad \hat{\sigma}_{y,S_n} = \frac{\bar{\delta}_{y,S_n}}{\mathbb{E}[\chi]}. \tag{21}$$

Since $\mathbb{E}[\chi]$ is an upper bound for $\sigma_{S_n}^{-1}\mathbb{E}[\delta(\hat{X})]$, for any value of σ_{S_n}, the estimators in (21) are upper bounds for those obtained by solving (20), hence this approximation yields systematically more bolstering.

Remark 16 *In Braga-Neto and Dougherty (2004) it was proposed to divide by the median of a χ distribution in* (21)*. The median for d from* 1 *to* 5 *is, respectively, around* 0.67*,* 1.17*,* 1.53*,* 1.83 *and* 2.08*, which are values close to the mean, that is around* 0.79*,* 1.25*,* 1.59*,* 1.88 *and* 2.12 *for d from* 1 *to* 5*, respectively. Therefore, exchanging them in* (21) *should not have a great impact.*

Remark 17 *Equation* (20) *can be solved to estimate any kernel K_{S_n} that depends on only one unknown parameter, even if it is not of the form $\sigma^2_{S_n}\Sigma$.*

Remark 18 *Other distributions beyond Gaussian could be considered, and the deductions remain true by exchanging the densities p_{i,S_n} and p_{i,y,S_n}, and the χ distribution, for the respective densities and distribution of the distance to the mean.*

5.2 Maximum pseudo-likelihood estimator

In this section, we assume that each point (X_i, Y_i) has a distinct kernel Σ_{i,S_n}, that is a general positive-definite matrix, and we propose an estimator for these n matrices based on a pseudo-likelihood function. We denote these matrices simply by Σ_i to ease notation, and it should be implicit that they depend on the sample S_n.

We consider Gaussian bolstering with kernels $\Sigma_i = (n-1)^{-1}\tilde{\Sigma}_i$ which are obtained by maximizing

$$\mathcal{L}_{S_n}(\Sigma_1, \ldots, \Sigma_n, \pi_1, \ldots, \pi_n) = \prod_{i=1}^{n} \left(\sum_{j=1}^{n} \pi_j \; p_{j,\Sigma_j}(X_i) \; \mathbb{1}_{i \neq j} \right) \tag{22}$$

in which p_{j,Σ_j} is the density of a Gaussian distribution with mean X_j and covariance matrix Σ_j, and $0 \leqslant \pi_j \leqslant 1, j = 1, \ldots, n$, are probabilities that sum to one.

The function $\mathcal{L}_{S_n}$ is a *pseudo-likelihood* function of sample S_n that approximates the probability density function of X_i by a Gaussian mixture with means X_j and covariance matrices $\Sigma_j, j = 1, \ldots, n, j \neq i$. A maximum pseudo-likelihood estimator for $\Sigma_1, \ldots, \Sigma_n$ is obtained by maximizing (22).

The EM algorithm can be applied to obtain the kernels that maximize (22). The algorithm is an adaptation of the classical EM algorithm for the mixture of Gaussian distributions (Dempster et al., 1977), in which the means are known. We present the steps of the algorithm in Algorithm 1, where we denote by $\mathbf{\Sigma} = \{\Sigma_1, \ldots, \Sigma_n\} \coloneqq \{(n-1)^{-1}\tilde{\Sigma}_1, \ldots, (n-1)^{-1}\tilde{\Sigma}_n\}$ a collection of kernels.

We consider a hyperparameter λ in the E-step by updating the weights to

$$w_{i,j}^{(t)} = \frac{(\lambda + p_{i,\Sigma_i^{(t)}}(X_j))\ \mathbb{1}_{i\neq j}}{\lambda(n-1) + \sum_{k=1}^{n} p_{k,\Sigma_k^{(t)}}(X_j)\ \mathbb{1}_{k\neq j}}$$

to obtain a more stable solution. On the one hand, if $\lambda = 0$, then the solution is unstable, since $p_{i,\Sigma_i^{(t)}}(X_j)$ can be very small for all values of $j \neq i$ and the weights might not converge. On the other hand, when $\lambda \to \infty$ the matrices converge to

$$\Sigma_i = \frac{1}{(n-1)^2} \sum_j (X_j - X_i)(X_j - X_i)^T\, \mathbb{1}_{i\neq j}$$

that is $(n-1)^{-1}$ times the mean distance matrix from each point in the sample to X_i. In this paper we consider $\lambda = 1$, so the weights are a perturbation of $1/(n-1)$ by the Gaussian densities. We call the estimator $\hat{\mathbf{\Sigma}}$ obtained via the EM algorithm the *maximum pseudo-likelihood estimator* (MPE).

Algorithm 1 EM algorithm for kernel estimation in Gaussian bolstering.

1: **Initialize**: Kernels $\mathbf{\Sigma}^{(0)}, t = 0, \lambda > 0$ and $\epsilon_c > 0$.
2: **while** $\max_i \|\Sigma_i^{(t-1)} - \Sigma_i^{(t)}\|_\infty \geqslant \epsilon_c$ **do**
3: **E-step**: Compute the conditional probabilities

$$w_{i,j}^{(t)} = \frac{(\lambda + p_{i,\Sigma_i^{(t)}}(X_j))\ \mathbb{1}_{i\neq j}}{\lambda(n-1) + \sum_{k=1}^{n} p_{k,\Sigma_k^{(t)}}(X_j)\ \mathbb{1}_{k\neq j}}$$

4: **M-step**: Update the kernels as

$$\Sigma_i^{(t+1)} = \frac{1}{n-1} \sum_j w_{i,j}^{(t)} (X_j - X_i)(X_j - X_i)^T$$

5: t = t + 1

Figure 5 presents an example of the method of moments (MM) estimators, exact and approximated, with spherical matrix $\Sigma = I_d$, and the MPE for Gaussian and XY-Gaussian bolstering. We see that the exact chi approximation kernel is wider than the MM exact kernel, since the chi-based approximation is an upper bound for it. Furthermore, the MPE kernel is much smaller (in the matrix trace sense) than those obtained via the MM,

and it *points* to the direction where the other data points are in average *further* from the point. This means that the directions with more probability are those in which there are fewer sample points close to the i-th point. This makes sense, as these are the directions less represented in the sample, where the loss of ψ_n is not evaluated through the standard resubstitution.

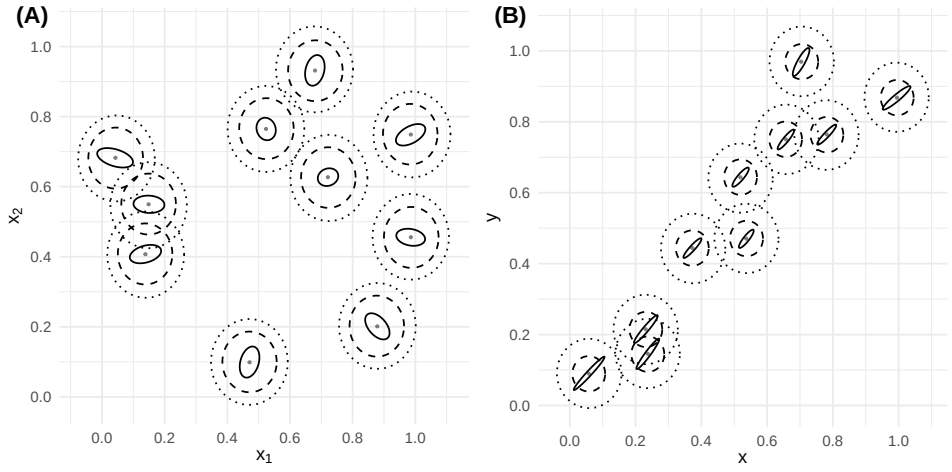

Figure 5: Example of method of moments and maximum pseudo-likelihood estimators for the kernel. The plots present the level curves of the Gaussian distributions with kernel estimated by the method of moments, exact (dashed) and approximated (dotted), and maximum pseudo-likelihood estimation (solid) in **(A)** Gaussian bolstering with $d = 2$ and **(B)** XY-Gaussian bolstering with $d = 1$. All the level curves are such that the probability inside the sphere/ellipse equals 0.05.

6. Experimental results

In this section, we present experimental results on error estimation for polynomial regression with synthetic data. The simulations were performed in **R** (R Core Team, 2023) with the *genree* package that we developed specifically for it and that is available at **https://github.com/dmarcondes/genree**. Bayesian polynomial regressions were fitted with the package *rstanarm* (Goodrich et al., 2023).

The synthetic data is generated from X uniformly distributed in $[0, 1]^d$ for $d = 1, 2, 3$ and $Y = \psi^\star(X) + \xi_\sigma$ in which

$$\psi^\star(x) = \left[1 + \sum_{i=1}^{d} x_i\right]^{p_g}$$

for $p_g = 1, 2, 3$, and ξ_σ is a random variable, independent of X, with the Gaussian distribution of mean zero and variance σ^2 for $\sigma = 0.25, 0.5$. We consider sample sizes of $n = 20, 50, 100$ and perform Monte Carlo integration with a sample size of $1,000$ to calculate ε_n and the error estimators.

The simulations are performed for prediction rules generated by least squares estimation of a degree p_f polynomial for $p_f = 1, 2$. We consider the following generalized error estimators: X-Gaussian bolstering with kernel estimated via maximum pseudo-likelihood (MPE) and exact method of moments (MM); XY-Gaussian bolstering with kernel estimated via maximum pseudo-likelihood (MPE); posterior-probability under degree p_f Bayesian polynomial regression; and Gaussian bolstering posterior-probability under MPE and degree p_f Bayesian polynomial regression. The degree p_f of Bayesian regression is the same degree of the fitted polynomial, but may differ from the degree p_g of the polynomial that generated the data. We also include cross-validation (CV) with 10 folds for comparison. The bias and root-mean-square error (RMSE) of each scenario are estimated as averages over 100 independent samples.

In Figures 6 to 8 we present the bias $\pm$ RMSE of the estimators for each scenario for $d = 1, 2$ and 3, respectively. The result for the X-Gaussian bolstering with kernel estimated via the method of moments is omitted from the figures for $d \geqslant 2$ since its results are, in general, significantly worse than the other estimators. Tables with the bias and RMSE of each scenario and estimator can be found in Appendix A.

As expected, the CV estimator has the smallest bias in the majority of the scenarios (59 out of 108), followed by the posterior-probability estimator (37 scenarios). While CV has smaller bias in low dimensions ($d = 1, 2$), the posterior-probability estimator has in general a smaller bias in $d = 3$, especially for target functions given by degree 2 and 3 polynomials. Considering only generalized resubstitution estimators, we see that the posterior-probability estimator has the least absolute bias in the majority of the scenarios (83 out of 108), with the exception of the scenarios with $d = 1$ and $\sigma = 0.5$ in which bolstering estimators had the least absolute bias in many cases. The bolstering estimators in the X direction, and the Gaussian bolstering posterior-probability estimator, perform poorly for $p_g = 2, 3$ especially for $d \geqslant 2$, evidencing that bolstering only on the X direction may not be suitable in higher dimensional scenarios and when the data is generated by more complex functions.

Conversely, the performance of the XY-bolstering estimator is not sensitive to the dimension d and the data-generating polynomial degree p_g. It is among the estimators with the least RMSE in the majority of cases, and has

the least one in 33 out of 108 scenarios. The posterior-probability estimator is also more robust to the dimension and the data-generating polynomial degree, and has the least RMSE in 20 scenarios. The MPE and MM X-bolstering estimators achieved the least RMSE in 18 cases each, most of them in lower-dimensional scenarios. Finally, the CV estimator did not perform as well from the RMSE point-of-view, achieving the least RMSE in only 13 cases, generally in lower-dimensional scenarios. These results indicate that generalized resubstitution error estimators have in general lower RMSE compared to CV in low-sample scenarios, especially in higher dimensions with more complex target functions.

The posterior-probability estimator depends on the degree p_f of the Bayesian polynomial regression, which is set equal to the degree of the fitted prediction rule but may differ from the degree p_g of the data-generating polynomial. It is therefore natural to ask how sensitive the estimator is to this potential misspecification. For $p_g = 1$ and 2, the posterior-probability estimator performs similarly regardless of whether p_f matches p_g or not, with differences in bias and RMSE between $p_f = 1$ and $p_f = 2$ remaining negligible across all dimensions and sample sizes. This suggests that the estimator is robust to mild degree misspecification in simpler scenarios. However, for $p_g = 3$, where both $p_f = 1$ and $p_f = 2$ are misspecified, the estimator with $p_f = 2$ consistently achieves substantially lower RMSE than with $p_f = 1$, with the gap widening as dimension increases. For instance, at $d = 3, \sigma = 0.25$, and $n = 20$ the RMSE drops from 5.34 to 0.24 when moving from $p_f = 1$ to $p_f = 2$. This indicates that, while the posterior-probability estimator does not require the Bayesian regression degree to match the data-generating degree exactly, using a higher-degree and therefore more flexible prior is beneficial when the target function is complex. Since in practice p_f is chosen to match the fitted model rather than the unknown p_g, this result also implies that fitting a more flexible prediction rule carries the additional benefit of a more accurate error estimate.

7. Discussion

In this paper, generalized resubstitution error estimators were extended to the general supervised learning framework, in particular to regression problems, and studied from a statistical learning distribution-free perspective. We formally defined generalized resubstitution error estimators as the expected loss under a generalized empirical distribution and, equivalently, as the resubstitution error of a generalized loss function. Sufficient conditions for the consistency of these estimators were established without making any

assumptions about the prediction rule or distribution; only an assumption about the moments of the loss function was made (cf. (11)). Consistency of these estimators was studied in general and specific cases, and the results were particularized to smoothing on the X, Y and both directions.

In particular, we extended to the regression case the result of Ghane and Braga-Neto (2022) by establishing the convergence when the generalized empirical measures converge to the point measure (cf. Corollary 6). This condition holds, for instance, in Gaussian bolstering for the squared loss when the entries of the kernel converge to zero (cf. Proposition 13). Moreover, we have also presented conditions under which the estimators are consistent even when the entries of the kernel do not converge to zero (cf. Proposition 12), a result that is also novel for classification. Two methods for learning the kernel in Gaussian bolstering were proposed, formalizing the heuristic approach of Braga-Neto and Dougherty (2004), by applying the EM algorithm for Gaussian mixtures. The estimators were applied to polynomial regression, and their superiority with respect to plain resubstitution was empirically observed.

Rather than studying the statistical properties of the proposed error estimators in specific statistical models, this paper aimed to understand under which general conditions they are consistent. Nevertheless, even though outside the scope of this paper, better understanding these estimators under a *distribution-dependent* framework for specific prediction rules is quite relevant, and we leave it for future research. The main questions within this line of inquiry would concern the estimators' convergence rate, efficiency, and asymptotic normality, from which inferential methods could be derived.

On the one hand, these questions can hardly be answered in a useful manner in a distribution-free, prediction-rule-agnostic, framework. In particular, the convergence rates we could deduce in this instance would be very pessimistic and therefore have little practical use, while any efficiency theory would depend heavily on the data distribution, which we assume is completely unknown. On the other hand, by making assumptions about the data-generating distribution and prediction rule, for example, under a generalized linear model (GLM), these questions could be properly answered. In this case, the convergence rate and asymptotic normality could be established by the usual techniques in GLM, and the families of generalized resubstitution estimators proposed in this paper could be compared from an efficiency perspective. For instance, the most efficient family for a given model could be theoretically determined, while the algorithms proposed in this paper could be applied to estimate the kernel from the data.

From an applied perspective, it would be interesting to apply these estimators to more complex regression models, e.g., neural networks. In such cases, overfitting is expected to be more prevalent in small-sample scenarios, and a greater benefit from smoothing the empirical measure could be obtained. We also leave this important topic for future research.

Acknowledgments

D. Marcondes was funded by grants #2022/06211-2 and #2023/00256-7, São Paulo Research Foundation (FAPESP). U. Braga-Neto was supported by NSF Award CCF-2225507. Most of this work was developed while D. Marcondes was a visiting scholar at the Department of Electrical and Computer Engineering, Texas A&M University.

References

P. L. Bartlett, P. M. Long, and R. C. Williamson. Fat-shattering and the learnability of real-valued functions. In *Proceedings of the seventh annual conference on Computational learning theory*, pages 299–310, 1994.

S. Bates, T. Hastie, and R. Tibshirani. Cross-validation: what does it estimate and how well does it do it? *Journal of the American Statistical Association*, 119(546):1434–1445, 2024.

U. Braga-Neto. *Fundamentals of pattern recognition and machine learning.* Springer, 2020.

U. Braga-Neto and E. Dougherty. Bolstered error estimation. *Pattern Recognition*, 37(6):1267–1281, 2004.

U. Braga-Neto and E. Dougherty. *Error Estimation for Pattern Recognition.* Wiley, New York, 2015.

C. Cortes, S. Greenberg, and M. Mohri. Relative deviation learning bounds and generalization with unbounded loss functions. *Annals of Mathematics and Artificial Intelligence*, 85:45–70, 2019.

L. Dalton and E. Dougherty. Bayesian minimum mean-square error estimation for classification error - part I: Definition and the bayesian mmse error estimator for discrete classification. *IEEE Transactions on Signal Processing*, 59(1):115–129, 2011a.

L. Dalton and E. Dougherty. Bayesian minimum mean-square error estimation for classification error - part II: Linear classification of gaussian models. *IEEE Transactions on Signal Processing*, 59(1):130–144, 2011b.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

B. Efron. Bootstrap methods: Another look at the jacknife. *Annals of Statistics*, 7:1–26, 1979.

B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382): 316–331, 1983.

B. Efron and R. Tibshirani. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92 (438):548–560, 1997.

P. Ghane and U. Braga-Neto. Generalized resubstitution for classification error estimation. *The Journal of Machine Learning Research*, 23(1):12811–12840, 2022.

B. Goodrich, J. Gabry, I. Ali, and S. Brilleman. rstanarm: Bayesian applied regression modeling via Stan., 2023. URL `https://mc-stan.org/rstanarm/`. R package version 2.26.1.

D. Hand. Recent advances in error rate estimation. *Pattern Recognition Letters*, 4:335–346, 1986.

A. Hefny and A. F. Atiya. A new monte carlo-based error rate estimator. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 37–47. Springer, 2010.

Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

P. Lachenbruch and M. Mickey. Estimation of error rates in discriminant analysis. *Technometrics*, 10:1–11, 1968.

G. Lugosi and M. Pawlak. On the posterior-probability estimate of the error rate of nonparametric classification rules. *IEEE Transactions on Information Theory*, 40(2):475–481, 1994.

G. McLachlan. Error rate estimation in discriminant analysis: recent advances. In A. Gupta, editor, *Advances in Multivariate Analysis*. D. Reidel, Dordrecht, 1987.

D. Pollard. Asymptotics via empirical processes. *Statistical science*, pages 341–354, 1989.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL `https://www.R-project.org/`.

B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do imagenet classifiers generalize to imagenet? *arXiv preprint arXiv:1902.10811*, 2019.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015.

R. Schiavo and D. Hand. Ten more years of error rate research. *International Statistical Review*, 68(3):295–310, 2000.

C. Smith. Some examples of discrimination. *Annals of Eugenics*, 18:272–282, 1947.

E. D. Sontag et al. Vc dimension of neural networks. *NATO ASI Series F Computer and Systems Sciences*, 168:69–96, 1998.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36: 111–147, 1974.

G. Toussaint. Bibliography on estimation of misclassification. *IEEE Transactions on Information Theory*, IT-20(4):472–479, 1974.

V. Vapnik. Statistical learning theory., 1998.

M. R. Yousefi, J. Hua, and E. R. Dougherty. Multiple-rule bias in the comparison of classification rules. *Bioinformatics*, 27(12):1675–1683, 2011.

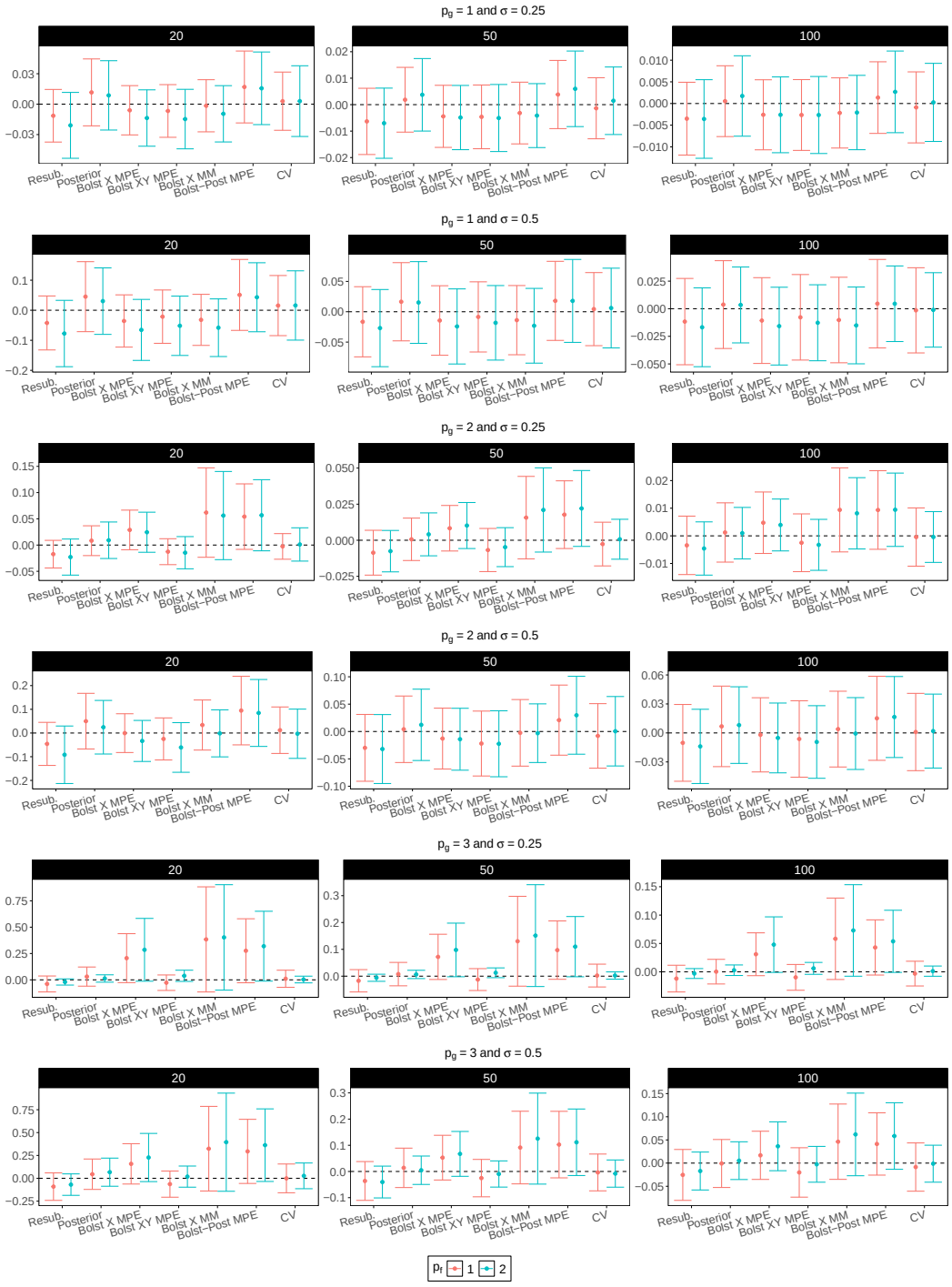

Figure 6: Bias $\pm$ RMSE for each estimator and sample size over the 100 samples for $d = 1$. Each plot represents a value of p_g and σ, and the colors refer to p_f.

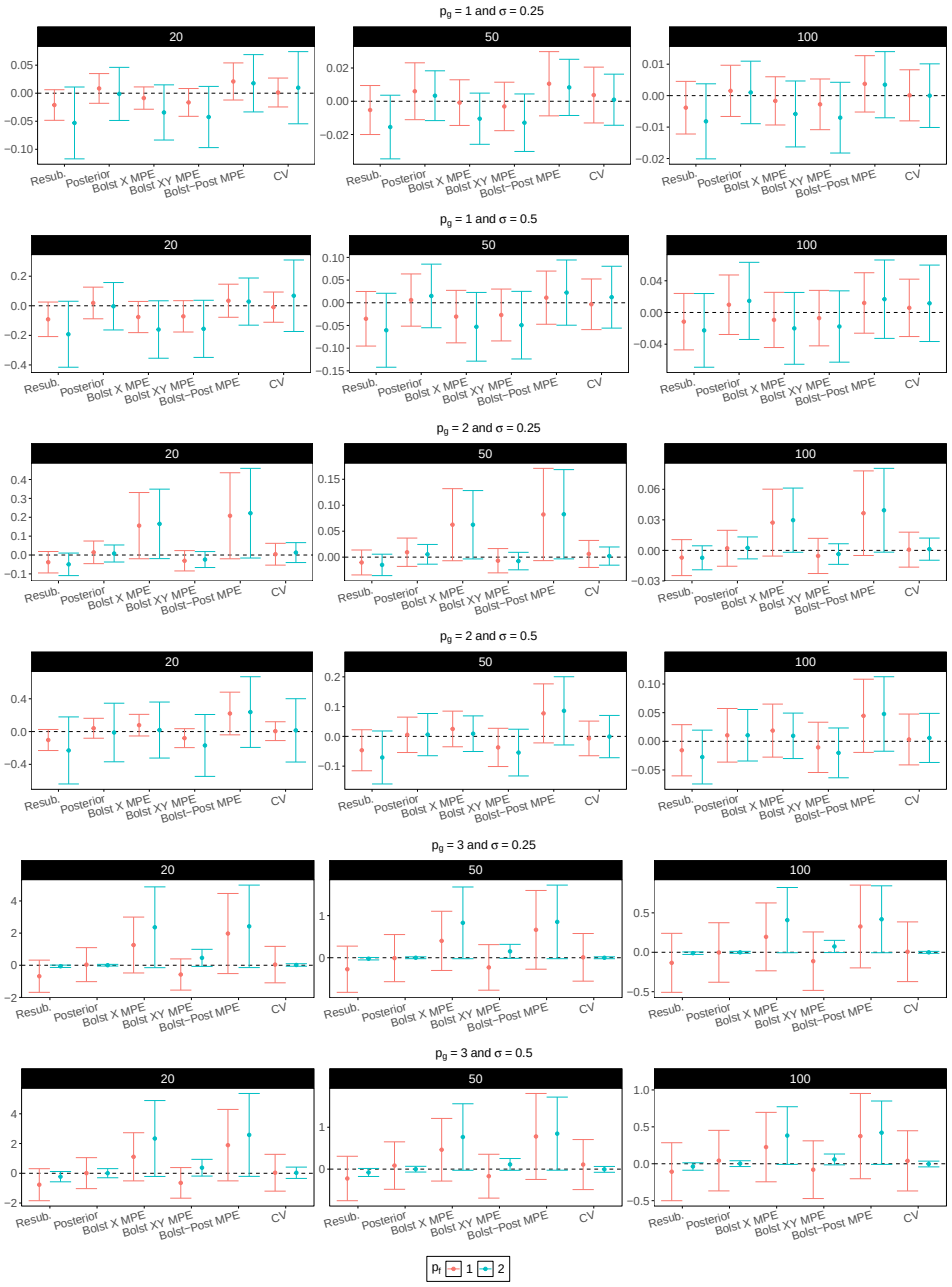

Figure 7: Bias $\pm$ RMSE for each estimator and sample size over the 100 samples for $d = 2$. Each plot represents a value of p_g and σ, and the colors refer to p_f.

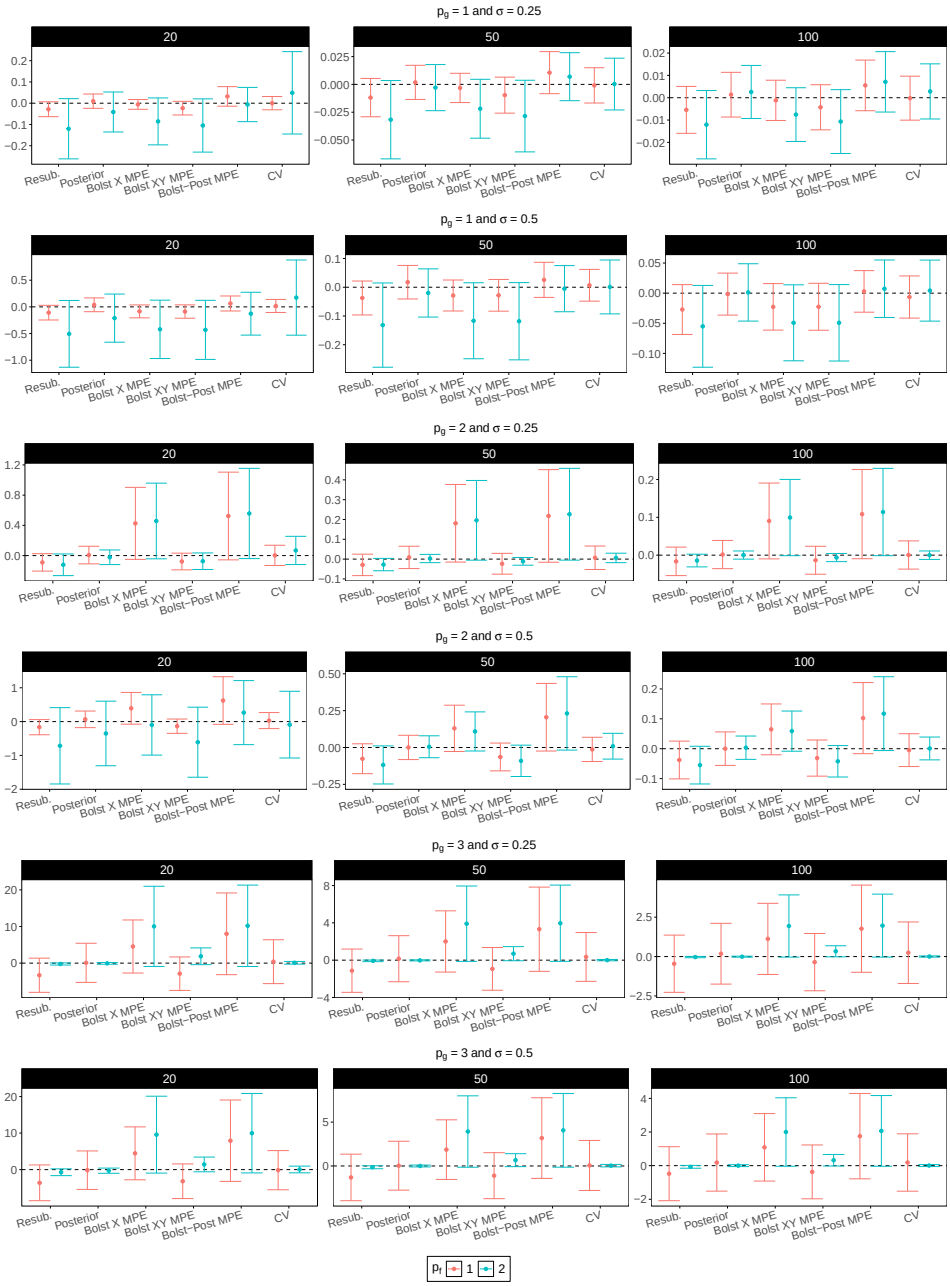

Figure 8: Bias $\pm$ RMSE for each estimator and sample size over the 100 samples for $d = 3$. Each plot represents a value of p_g and σ, and the colors refer to p_f.

Appendix A. Detailed results of the experiments

Table 1: Bias of the estimators in each scenario over the 100 samples for $d = 1$. The least absolute bias in each scenario is in boldface.

d	σ	n	p_g	p_f	Resub	Post	X - MPE	XY - MPE	X - MM	MPE - Post	CV
1	0.25	20	1	1	-0.011	0.012	-0.006	-0.0067	**-0.0016**	0.017	0.0029
1	0.25	20	1	2	-0.021	0.0087	-0.014	-0.015	-0.0095	0.016	**0.003**
1	0.25	20	2	1	-0.017	0.0083	0.029	-0.013	0.062	0.054	**-0.0026**
1	0.25	20	2	2	-0.023	0.0091	0.024	-0.015	0.056	0.057	**0.0011**
1	0.25	20	3	1	-0.039	0.031	0.21	-0.027	0.38	0.28	**0.01**
1	0.25	20	3	2	-0.02	0.013	0.29	0.038	0.4	0.32	**0.0031**
1	0.25	50	1	1	-0.0063	0.0019	-0.0044	-0.0046	-0.0032	0.0038	**-0.0014**
1	0.25	50	1	2	-0.007	0.0037	-0.0049	-0.0051	-0.0041	0.006	**0.0015**
1	0.25	50	2	1	-0.0086	**0.00064**	0.0083	-0.0068	0.016	0.018	-0.0027
1	0.25	50	2	2	-0.0076	0.0041	0.01	-0.0048	0.021	0.022	**0.00069**
1	0.25	50	3	1	-0.017	0.0077	0.072	-0.013	0.13	0.097	**0.0022**
1	0.25	50	3	2	-0.0059	0.0065	0.098	0.013	0.15	0.11	**0.0024**
1	0.25	100	1	1	-0.0035	**0.00054**	-0.0026	-0.0027	-0.0022	0.0014	-0.00089
1	0.25	100	1	2	-0.0036	0.0017	-0.0026	-0.0027	-0.0021	0.0027	**0.00026**
1	0.25	100	2	1	-0.0034	0.0012	0.0047	-0.0025	0.0094	0.0094	**-0.00044**
1	0.25	100	2	2	-0.0046	0.00097	0.0039	-0.0033	0.0082	0.0094	**-0.00042**
1	0.25	100	3	1	-0.012	**0.00026**	0.031	-0.0098	0.058	0.043	-0.0032
1	0.25	100	3	2	-0.003	0.0028	0.048	0.006	0.073	0.054	**0.0012**
1	0.5	20	1	1	-0.042	0.045	-0.036	-0.021	-0.032	0.051	**0.016**
1	0.5	20	1	2	-0.077	0.03	-0.065	-0.052	-0.058	0.043	**0.016**
1	0.5	20	2	1	-0.046	0.05	**-0.00053**	-0.025	0.034	0.095	0.012
1	0.5	20	2	2	-0.092	0.024	-0.033	-0.061	**-0.0015**	0.085	-0.0031
1	0.5	20	3	1	-0.091	0.045	0.16	-0.064	0.32	0.29	**-0.0011**
1	0.5	20	3	2	-0.069	0.067	0.23	**0.018**	0.39	0.36	0.027
1	0.5	50	1	1	-0.017	0.016	-0.014	-0.0085	-0.014	0.018	**0.0043**
1	0.5	50	1	2	-0.027	0.015	-0.024	-0.018	-0.023	0.018	**0.0061**
1	0.5	50	2	1	-0.03	0.0041	-0.013	-0.022	**-0.0025**	0.021	-0.008
1	0.5	50	2	2	-0.032	0.012	-0.014	-0.022	-0.0031	0.03	**0.00046**
1	0.5	50	3	1	-0.036	0.014	0.053	-0.025	0.091	0.1	**-0.0037**
1	0.5	50	3	2	-0.04	**0.0047**	0.067	-0.0097	0.13	0.11	-0.0081
1	0.5	100	1	1	-0.012	0.0038	-0.011	-0.0078	-0.01	0.0046	**-0.0015**
1	0.5	100	1	2	-0.017	0.0034	-0.016	-0.013	-0.015	0.0045	**-0.0011**
1	0.5	100	2	1	-0.01	0.0068	-0.002	-0.0063	0.004	0.015	**0.00095**
1	0.5	100	2	2	-0.014	0.0081	-0.0052	-0.0094	**-0.00062**	0.016	0.0019
1	0.5	100	3	1	-0.025	**-0.00083**	0.017	-0.02	0.046	0.041	-0.0085
1	0.5	100	3	2	-0.017	0.0053	0.036	-0.0026	0.062	0.058	**-0.0013**

Table 2: Bias of the estimators in each scenario over the 100 samples for $d = 2$. The least absolute bias in each scenario is in boldface.

d	σ	n	p_g	p_f	Resub	Post	X - MPE	XY - MPE	X - MM	MPE - Post	CV
2	0.25	20	1	1	-0.021	0.0085	-0.0087	-0.016	0.022	0.021	**0.0012**
2	0.25	20	1	2	-0.053	**-0.0013**	-0.034	-0.042	0.012	0.018	0.0098
2	0.25	20	2	1	-0.038	0.014	0.16	-0.03	0.66	0.21	**0.0042**
2	0.25	20	2	2	-0.049	**0.008**	0.16	-0.024	0.76	0.22	0.013
2	0.25	20	3	1	-0.68	**0.038**	1.26	-0.58	6.09	1.97	0.041
2	0.25	20	3	2	-0.064	**0.0027**	2.36	0.46	8.79	2.42	0.017
2	0.25	50	1	1	-0.0053	0.006	**-0.00082**	-0.0031	0.018	0.011	0.0037
2	0.25	50	1	2	-0.016	0.0033	-0.01	-0.013	0.0091	0.0083	**0.00093**
2	0.25	50	2	1	-0.01	0.0094	0.062	-0.0069	0.34	0.082	**0.0062**
2	0.25	50	2	2	-0.015	0.0054	0.062	-0.0076	0.35	0.083	**0.0021**
2	0.25	50	3	1	-0.27	**-0.0092**	0.4	-0.23	3.11	0.66	0.0097
2	0.25	50	3	2	-0.023	**-0.00033**	0.83	0.15	4.02	0.85	-0.00042
2	0.25	100	1	1	-0.0038	0.0015	-0.0017	-0.0028	0.0092	0.0037	**9.4e-05**
2	0.25	100	1	2	-0.0082	0.001	-0.0058	-0.007	0.0053	0.0035	**-1.1e-05**
2	0.25	100	2	1	-0.0071	0.0021	0.027	-0.0054	0.2	0.037	**0.00071**
2	0.25	100	2	2	-0.0073	0.0024	0.03	-0.0036	0.2	0.039	**0.0012**
2	0.25	100	3	1	-0.13	**-0.0027**	0.2	-0.11	1.81	0.33	0.0066
2	0.25	100	3	2	-0.012	**-0.00081**	0.41	0.074	2.15	0.42	-0.00088
2	0.5	20	1	1	-0.091	0.019	-0.076	-0.071	-0.033	0.034	**-0.0091**
2	0.5	20	1	2	-0.19	**-0.0031**	-0.16	-0.16	-0.071	0.028	0.068
2	0.5	20	2	1	-0.1	0.04	0.078	-0.08	0.53	0.22	**0.0049**

2	0.5	20	2	2	-0.23	**-0.011**	0.019	-0.17	0.68	0.24	0.015
2	0.5	20	3	1	-0.77	**0.0099**	1.11	-0.64	6.34	1.89	0.037
2	0.5	20	3	2	-0.23	**0.0062**	2.34	0.37	8.3	2.58	0.035
2	0.5	50	1	1	-0.035	0.006	-0.03	-0.027	-0.013	0.011	**-0.0033**
2	0.5	50	1	2	-0.061	0.015	-0.053	-0.049	-0.027	0.022	**0.012**
2	0.5	50	2	1	-0.047	**0.0052**	0.025	-0.037	0.3	0.077	-0.0069
2	0.5	50	2	2	-0.071	0.0059	0.009	-0.054	0.29	0.086	**-0.00069**
2	0.5	50	3	1	-0.22	**0.085**	0.46	-0.17	2.98	0.78	0.11
2	0.5	50	3	2	-0.08	**-0.0013**	0.77	0.11	3.71	0.85	-0.0061
2	0.5	100	1	1	-0.012	0.0098	-0.0094	-0.0071	**0.0014**	0.012	0.0058
2	0.5	100	1	2	-0.023	0.015	-0.02	-0.018	**-0.0082**	0.017	0.012
2	0.5	100	2	1	-0.016	0.011	0.019	-0.011	0.18	0.045	**0.0032**
2	0.5	100	2	2	-0.027	0.011	0.0096	-0.02	0.18	0.048	**0.0059**
2	0.5	100	3	1	-0.11	0.042	0.23	-0.08	1.8	0.37	**0.039**
2	0.5	100	3	2	-0.037	**0.0016**	0.38	0.059	2.08	0.42	-0.0032

Table 3: Bias of the estimators in each scenario over the 100 samples for $d = 3$. The least absolute bias in each scenario is in boldface.

d	σ	n	p_g	p_f	Resub	Post	X - MPE	XY - MPE	X - MM	MPE - Post	CV
3	0.25	20	1	1	-0.028	0.0096	-0.0057	-0.023	0.077	0.032	**0.00028**
3	0.25	20	1	2	-0.12	-0.041	-0.086	-0.1	0.065	**-0.0062**	0.049
3	0.25	20	2	1	-0.09	0.0069	0.43	-0.078	2.57	0.52	**0.0029**
3	0.25	20	2	2	-0.12	**-0.022**	0.46	-0.074	2.53	0.56	0.068
3	0.25	20	3	1	-3.31	**0.081**	4.54	-2.88	35.06	8	0.4
3	0.25	20	3	2	-0.23	**-0.065**	10.04	1.87	45.51	10.19	0.091
3	0.25	50	1	1	-0.012	0.0018	-0.0032	-0.0097	0.049	0.011	**-0.00089**
3	0.25	50	1	2	-0.032	-0.0029	-0.022	-0.028	0.04	0.0069	**0.00028**
3	0.25	50	2	1	-0.029	0.0086	0.18	-0.024	1.42	0.22	**0.0067**
3	0.25	50	2	2	-0.028	**0.0029**	0.2	-0.011	1.47	0.23	0.0059
3	0.25	50	3	1	-1.13	**0.16**	2.01	-0.93	20.49	3.32	0.35
3	0.25	50	3	2	-0.073	-0.022	3.91	0.7	24.96	3.96	**0.0034**
3	0.25	100	1	1	-0.0054	0.0014	-0.0012	-0.0043	0.034	0.0055	**-2e-04**
3	0.25	100	1	2	-0.012	**0.0026**	-0.0076	-0.011	0.03	0.0071	0.0028
3	0.25	100	2	1	-0.016	0.0016	0.09	-0.013	0.96	0.11	**0.00042**
3	0.25	100	2	2	-0.014	0.00025	0.099	-0.0064	0.99	0.11	**-9.8e-05**
3	0.25	100	3	1	-0.45	**0.18**	1.12	-0.35	13.7	1.77	0.24
3	0.25	100	3	2	-0.032	-0.007	1.94	0.33	17.13	1.96	**0.0024**
3	0.5	20	1	1	-0.11	0.039	-0.085	-0.087	**0.0086**	0.064	0.017
3	0.5	20	1	2	-0.51	-0.21	-0.42	-0.43	**-0.059**	-0.13	0.17
3	0.5	20	2	1	-0.16	0.067	0.39	-0.14	2.43	0.62	**0.03**
3	0.5	20	2	2	-0.72	-0.35	-0.099	-0.61	2.2	0.27	**-0.09**
3	0.5	20	3	1	-3.62	-0.16	4.45	-3.18	37.28	7.91	**-0.16**
3	0.5	20	3	2	-0.72	-0.3	9.57	1.41	45.95	9.97	**0.037**
3	0.5	50	1	1	-0.037	0.017	-0.029	-0.028	0.027	0.026	**0.0068**
3	0.5	50	1	2	-0.13	-0.02	-0.12	-0.12	-0.023	-0.005	**0.001**
3	0.5	50	2	1	-0.076	**0.00054**	0.13	-0.064	1.33	0.21	-0.013
3	0.5	50	2	2	-0.12	**0.0051**	0.11	-0.09	1.42	0.23	0.0088
3	0.5	50	3	1	-1.31	**0.033**	1.86	-1.11	21.75	3.18	0.06
3	0.5	50	3	2	-0.15	**-0.002**	3.93	0.66	25.38	4.06	0.036
3	0.5	100	1	1	-0.027	**-0.0016**	-0.023	-0.023	0.014	0.0029	-0.0064
3	0.5	100	1	2	-0.055	**0.0012**	-0.049	-0.049	-0.0039	0.0073	0.0042
3	0.5	100	2	1	-0.038	**0.00019**	0.065	-0.031	0.91	0.1	-0.0047
3	0.5	100	2	2	-0.055	0.0034	0.059	-0.042	0.97	0.12	**0.00089**
3	0.5	100	3	1	-0.48	**0.18**	1.09	-0.37	13.87	1.75	0.19
3	0.5	100	3	2	-0.07	**0.00087**	2	0.32	16.8	2.07	0.0065

Table 4: RMSE of the estimators in each scenario over the 100 samples for $d = 1$. The least RMSE in each scenario is in boldface.

d	σ	n	p_g	p_f	Resub	Post	X - MPE	XY - MPE	X - MM	MPE - Post	CV
1	0.25	20	1	1	0.026	0.033	**0.024**	0.026	0.026	0.035	0.029
1	0.25	20	1	2	0.032	0.034	0.028	0.029	**0.028**	0.036	0.035
1	0.25	20	2	1	0.026	0.028	0.038	0.025	0.085	0.062	**0.024**
1	0.25	20	2	2	0.034	0.035	0.038	**0.031**	0.084	0.068	0.032
1	0.25	20	3	1	0.076	0.091	0.23	**0.073**	0.5	0.3	0.081
1	0.25	20	3	2	**0.029**	0.035	0.3	0.053	0.5	0.33	0.032

1	0.25	50	1	1	0.013	0.012	0.012	0.012	0.012	0.013	**0.012**
1	0.25	50	1	2	0.013	0.014	0.012	0.013	**0.012**	0.014	0.013
1	0.25	50	2	1	0.016	**0.015**	0.016	0.015	0.029	0.023	0.015
1	0.25	50	2	2	0.014	0.015	0.016	**0.013**	0.029	0.026	0.014
1	0.25	50	3	1	0.042	0.044	0.084	**0.041**	0.17	0.11	0.043
1	0.25	50	3	2	**0.013**	0.015	0.1	0.018	0.19	0.11	0.014
1	0.25	100	1	1	0.0084	0.0082	**0.0081**	0.0082	0.0081	0.0083	0.0082
1	0.25	100	1	2	0.0091	0.0093	0.0087	0.0089	**0.0086**	0.0094	0.009
1	0.25	100	2	1	0.011	0.011	0.011	**0.01**	0.015	0.014	0.011
1	0.25	100	2	2	0.0096	0.0093	0.0094	0.0092	0.013	0.013	**0.0092**
1	0.25	100	3	1	0.024	**0.022**	0.038	0.023	0.072	0.049	0.022
1	0.25	100	3	2	**0.0087**	0.0093	0.049	0.01	0.081	0.055	0.009
1	0.5	20	1	1	0.09	0.12	0.087	0.089	**0.085**	0.12	0.1
1	0.5	20	1	2	0.11	0.11	0.1	0.099	**0.096**	0.11	0.12
1	0.5	20	2	1	0.091	0.12	**0.082**	0.089	0.11	0.14	0.098
1	0.5	20	2	2	0.12	0.11	**0.087**	0.1	0.099	0.14	0.1
1	0.5	20	3	1	0.15	0.17	0.22	**0.14**	0.46	0.35	0.16
1	0.5	20	3	2	0.12	0.15	0.26	**0.12**	0.54	0.4	0.14
1	0.5	50	1	1	0.058	0.064	0.057	0.058	**0.057**	0.065	0.06
1	0.5	50	1	2	0.064	0.067	0.062	**0.061**	0.062	0.068	0.066
1	0.5	50	2	1	0.061	0.061	**0.056**	0.059	0.061	0.064	0.059
1	0.5	50	2	2	0.063	0.065	0.057	0.06	**0.054**	0.071	0.063
1	0.5	50	3	1	0.074	0.075	0.085	0.071	0.14	0.13	**0.07**
1	0.5	50	3	2	0.061	0.054	0.086	**0.05**	0.17	0.13	0.052
1	0.5	100	1	1	0.039	0.04	0.039	0.039	0.039	0.04	**0.039**
1	0.5	100	1	2	0.036	0.034	0.035	0.034	0.035	0.034	**0.034**
1	0.5	100	2	1	0.04	0.042	**0.038**	0.04	0.039	0.043	0.04
1	0.5	100	2	2	0.038	0.04	**0.036**	0.038	0.037	0.042	0.038
1	0.5	100	3	1	0.055	**0.052**	0.052	0.053	0.081	0.067	0.052
1	0.5	100	3	2	0.041	0.041	0.053	**0.039**	0.089	0.072	0.04

Table 5: RMSE of the estimators in each scenario over the 100 samples for $d = 2$. The least RMSE in each scenario is in boldface.

d	σ	n	p_g	p_f	Resub	Post	X - MPE	XY - MPE	X - MM	MPE - Post	CV
2	0.25	20	1	1	0.027	0.027	**0.02**	0.025	0.033	0.033	0.026
2	0.25	20	1	2	0.064	0.047	0.049	0.055	**0.047**	0.051	0.064
2	0.25	20	2	1	0.057	0.06	0.18	**0.054**	0.73	0.23	0.058
2	0.25	20	2	2	0.06	0.045	0.18	**0.042**	0.84	0.24	0.053
2	0.25	20	3	1	1	1.06	1.74	**0.97**	6.88	2.49	1.13
2	0.25	20	3	2	0.074	**0.053**	2.51	0.53	9.76	2.57	0.079
2	0.25	50	1	1	0.015	0.017	**0.014**	0.015	0.024	0.019	0.017
2	0.25	50	1	2	0.019	**0.015**	0.015	0.017	0.018	0.017	0.015
2	0.25	50	2	1	0.024	0.027	0.069	**0.023**	0.36	0.089	0.026
2	0.25	50	2	2	0.021	0.019	0.066	**0.017**	0.38	0.086	0.018
2	0.25	50	3	1	0.55	0.56	0.7	**0.54**	3.37	0.94	0.57
2	0.25	50	3	2	0.029	**0.022**	0.85	0.17	4.27	0.87	0.023
2	0.25	100	1	1	0.0084	0.0081	**0.0077**	0.0081	0.013	0.009	0.0081
2	0.25	100	1	2	0.012	**0.0099**	0.01	0.011	0.011	0.011	0.01
2	0.25	100	2	1	0.018	0.018	0.033	0.017	0.2	0.042	**0.017**
2	0.25	100	2	2	0.012	0.011	0.032	**0.01**	0.21	0.041	0.011
2	0.25	100	3	1	0.37	0.38	0.43	**0.37**	1.95	0.53	0.38
2	0.25	100	3	2	0.016	**0.012**	0.41	0.077	2.25	0.42	0.012
2	0.5	20	1	1	0.12	0.11	0.11	0.11	**0.089**	0.11	0.1
2	0.5	20	1	2	0.22	0.16	0.19	0.19	**0.14**	0.16	0.24
2	0.5	20	2	1	0.13	0.12	0.13	0.12	0.59	0.26	**0.11**
2	0.5	20	2	2	0.41	0.36	**0.34**	0.38	0.85	0.43	0.39
2	0.5	20	3	1	1.08	1.04	1.62	**1.03**	7.18	2.4	1.24
2	0.5	20	3	2	0.35	**0.31**	2.55	0.56	9.05	2.79	0.38
2	0.5	50	1	1	0.06	0.058	0.058	0.057	**0.054**	0.059	0.056
2	0.5	50	1	2	0.081	0.07	0.076	0.074	**0.059**	0.072	0.068
2	0.5	50	2	1	0.069	0.059	0.06	0.064	0.33	0.099	**0.058**
2	0.5	50	2	2	0.089	0.071	**0.06**	0.078	0.32	0.11	0.071
2	0.5	50	3	1	0.53	0.57	0.75	**0.52**	3.24	1.03	0.6
2	0.5	50	3	2	0.096	**0.068**	0.79	0.14	3.9	0.87	0.07
2	0.5	100	1	1	0.036	0.038	0.035	0.035	**0.033**	0.038	0.036
2	0.5	100	1	2	0.046	0.049	0.045	0.045	**0.042**	0.049	0.048
2	0.5	100	2	1	0.045	0.047	0.046	**0.044**	0.2	0.064	0.044
2	0.5	100	2	2	0.047	0.045	**0.04**	0.043	0.19	0.065	0.043
2	0.5	100	3	1	0.39	0.41	0.47	**0.39**	1.93	0.58	0.41

2	0.5	100	3	2	0.05	0.039	0.39	0.073	2.2	0.43	**0.039**

Table 6: RMSE of the estimators in each scenario over the 100 samples for $d = 3$. The least RMSE in each scenario is in boldface.

d	σ	n	p_g	p_f	Resub	Post	X - MPE	XY - MPE	X - MM	MPE - Post	CV
3	0.25	20	1	1	0.035	0.034	**0.023**	0.032	0.088	0.046	0.031
3	0.25	20	1	2	0.14	0.094	0.11	0.13	0.11	**0.081**	0.19
3	0.25	20	2	1	0.12	0.12	0.48	**0.11**	2.7	0.58	0.13
3	0.25	20	2	2	0.14	**0.097**	0.5	0.11	2.65	0.6	0.19
3	0.25	20	3	1	4.66	5.34	7.24	**4.57**	37.52	11.15	5.99
3	0.25	20	3	2	0.29	**0.24**	10.94	2.29	48.25	11.11	0.37
3	0.25	50	1	1	0.017	0.015	**0.013**	0.016	0.053	0.019	0.016
3	0.25	50	1	2	0.035	**0.021**	0.026	0.032	0.047	0.022	0.023
3	0.25	50	2	1	0.054	0.056	0.2	**0.053**	1.46	0.23	0.059
3	0.25	50	2	2	0.031	0.021	0.2	**0.019**	1.51	0.23	0.024
3	0.25	50	3	1	2.32	2.47	3.28	**2.29**	21.29	4.52	2.62
3	0.25	50	3	2	0.085	**0.058**	4.04	0.76	25.71	4.09	0.071
3	0.25	100	1	1	0.011	0.01	**0.009**	0.01	0.036	0.011	0.0099
3	0.25	100	1	2	0.015	**0.012**	0.012	0.014	0.033	0.014	0.012
3	0.25	100	2	1	0.038	0.037	0.1	**0.037**	0.98	0.12	0.037
3	0.25	100	2	2	0.017	0.011	0.1	**0.011**	1.01	0.12	0.011
3	0.25	100	3	1	**1.82**	1.93	2.26	1.82	14.15	2.77	1.95
3	0.25	100	3	2	0.041	**0.032**	1.97	0.35	17.55	2	0.039
3	0.5	20	1	1	0.14	0.13	0.12	0.13	**0.11**	0.14	0.12
3	0.5	20	1	2	0.63	0.45	0.55	0.55	**0.3**	0.4	0.7
3	0.5	20	2	1	0.22	0.25	0.47	**0.21**	2.55	0.7	0.24
3	0.5	20	2	2	1.13	0.95	**0.89**	1.04	2.59	0.95	0.99
3	0.5	20	3	1	4.89	5.27	7.24	**4.75**	40.81	11.14	5.36
3	0.5	20	3	2	0.93	**0.71**	10.52	2.01	48.53	10.87	0.92
3	0.5	50	1	1	0.059	0.058	**0.054**	0.055	0.058	0.061	0.055
3	0.5	50	1	2	0.15	0.084	0.13	0.13	**0.066**	0.08	0.094
3	0.5	50	2	1	0.1	0.082	0.16	0.094	1.37	0.23	**0.082**
3	0.5	50	2	2	0.13	**0.074**	0.13	0.11	1.46	0.25	0.088
3	0.5	50	3	1	2.65	2.79	3.4	**2.62**	22.7	4.6	2.85
3	0.5	50	3	2	0.17	**0.11**	4.07	0.74	26.17	4.2	0.14
3	0.5	100	1	1	0.041	0.035	0.039	0.039	0.036	**0.035**	0.035
3	0.5	100	1	2	0.068	0.048	0.063	0.063	**0.039**	0.048	0.051
3	0.5	100	2	1	0.063	0.056	0.085	0.06	0.92	0.12	**0.055**
3	0.5	100	2	2	0.063	0.039	0.067	0.053	0.98	0.12	**0.038**
3	0.5	100	3	1	1.61	1.7	2.01	**1.6**	14.36	2.54	1.71
3	0.5	100	3	2	0.084	**0.056**	2.04	0.34	17.11	2.1	0.059

Appendix B. VC dimension and uniform relative deviation convergence

The complexity of a hypothesis space may be measured by its VC dimension under loss function ℓ, that is a special case of the pseudo-dimension of a family of real-valued functions (Pollard, 1989; Vapnik, 1998). We start by recalling the definition of the shatter coefficient and VC dimension of a class $\mathcal{A} \subset \mathcal{B}(\mathbb{R}^{d+m})$ of Borel-measurable sets.

Definition 19 *The n-th shatter coefficient of a class $\mathcal{A} \subset \mathcal{B}(\mathbb{R}^{d+m})$ is defined as*

$$\mathcal{S}(\mathcal{A}, n) = \sup_{z_1,\ldots,z_n \in \mathcal{Z}} \left| \left\{ (\mathbb{1}\{z_1 \in A\}, \ldots, \mathbb{1}\{z_n \in A\}) : A \in \mathcal{A} \right\} \right|.$$

The VC dimension of $\mathcal{A}$ is the greatest integer n such that $\mathcal{S}(\mathcal{A}, n) = 2^n$ and is denoted by $d_{VC}(\mathcal{A})$.

For a fixed loss function ℓ, $\psi \in \mathcal{H}$ and $b \in (0, C_\ell)$ define

$$A_{\ell,\psi,b} = \{z \in \mathcal{Z} : \ell(z, \psi) > b\} \in \mathcal{B}(\mathbb{R}^{d+m})$$

as the points in $\mathcal{Z}$ where $\ell(z, \psi) > b$ and let

$$\mathcal{A}^\star_{\mathcal{H},\ell} = \{A_{\ell,\psi,b} : \psi \in \mathcal{H}, b \in (0, C_\ell)\} \subset \mathcal{B}(\mathbb{R}^{d+m})$$

be the collection of such sets for $\psi \in \mathcal{H}$ and $b \in (0, C_\ell)$. We have that

$$\mathcal{S}(\mathcal{A}^\star_{\mathcal{H},\ell}, n) = \sup_{z_1,\ldots,z_n \in \mathcal{Z}} \left| \left\{ (\mathbb{1}\{\ell(z_1, \psi) > b\}, \ldots, \mathbb{1}\{\ell(z_n, \psi) > b\}) : \psi \in \mathcal{H}, b \in (0, C_\ell) \right\} \right|$$

which is the usual shatter coefficient of the space of binary functions of form $I(z) = \mathbb{1}\{\ell(z, \psi) > b\}$ for $z \in \mathcal{Z}$.

In classification problems with two classes, under the simple loss function, $\mathcal{S}(\mathcal{A}^\star_{\mathcal{H},\ell}, n)$ reduces to the usual definition of the shatter coefficient of a hypothesis space of binary functions:

$$\mathcal{S}(\mathcal{H}, n) \coloneqq \sup_{x_1,\ldots,x_n \in \mathcal{X}} \left| \left\{ (\psi(x_1), \ldots, \psi(x_n)) : \psi \in \mathcal{H} \right\} \right|.$$

This fact justifies the definition of the shatter coefficient and VC dimension of a hypothesis space $\mathcal{H}$ under a loss function ℓ as that of $\mathcal{A}^\star_{\mathcal{H},\ell}$. See Appendix E for more details about the VC dimension under a loss function.

Definition 20 *The n-th shatter coefficient of $\mathcal{H}$ under loss function ℓ is defined as $\mathcal{S}(\mathcal{H}, \ell, n) \coloneqq \mathcal{S}(\mathcal{A}^\star_{\mathcal{H},\ell}, n)$. The VC dimension of a hypothesis space $\mathcal{H}$ under loss function ℓ is defined as $d_{VC}(\mathcal{H}, \ell) \coloneqq d_{VC}(\mathcal{A}^\star_{\mathcal{H},\ell})$.*

Bounds on the rate of uniform relative deviation convergence based on the VC dimension have been obtained in the literature. This paper will rely on a result in Cortes et al. (2019), that is an improvement of results in Vapnik (1998), and we refer to the references in Cortes et al. (2019) for a historical overview of these results.

For $\epsilon, \tau > 0$ and $n \geqslant 1$, let $\mathbb{B}_{n,\epsilon,\tau} : \mathbb{Z}_+ \to \mathbb{R}_+$ be a function such that

$$\sum_{n=1}^{\infty} \mathbb{B}_{n,\epsilon,\tau}(k) < \infty \tag{23}$$

for all $k \in \mathbb{Z}_+$ and all $\epsilon, \tau > 0$ fixed. The results in Cortes et al. (2019) imply that, under the assumption (11), there exists $\mathbb{B}_{n,\epsilon,\tau}(d_{VC}(\mathcal{H}, \ell))$ satisfying (23) which is a bound for the rate of uniform relative deviation in $\mathcal{H}$ under loss function ℓ. See Cortes et al. (2019) for an explicit form for $\mathbb{B}_{n,\epsilon,\tau}$.

Proposition 21 *Fix a loss function ℓ and let $\mathcal{H}$ be a hypothesis space with $d_{VC}(\mathcal{H},\ell) < \infty$. If (11) holds, then there exists a function $\mathbb{B}_{n,\epsilon,\tau}$ satisfying (23) such that, for all $\epsilon > 0$, and for a $\tau > 0$ small enough,*

$$\mathbb{P}\left(\sup_{\psi\in\mathcal{H}}\left|\frac{L(\psi)-L_n(\psi)}{\sqrt[\alpha]{L^\alpha(\psi)+\tau}}\right| > \epsilon\right) < \mathbb{B}_{n,\epsilon,\tau}(d_{VC}(\mathcal{H},\ell)). \tag{24}$$

It follows from (23) and Borel-Cantelli Lemma that

$$\sup_{\psi\in\mathcal{H}}\left|\frac{L(\psi)-L_n(\psi)}{\sqrt[\alpha]{L^\alpha(\psi)+\tau}}\right| \to\ 0 \tag{25}$$

with probability one as $n \to \infty$.

Remark 22 *The constant τ avoids the discontinuity of the ratio when $\inf_{\psi\in\mathcal{H}} L^\alpha(\psi) = 0$. See Cortes et al. (2019) for more details.*

Remark 23 *The results in this paper rely only on (25) and not on the rate in (24). Therefore, they rely only on the finiteness of $d_{VC}(\mathcal{H},\ell)$ rather than the specific form of $\mathbb{B}_{n,\epsilon,\tau}$, so other complexity measures that guarantee a result analogous to (25) could be considered.*

Remark 24 *The main goal of this paper is to establish the consistency of generalized resubstitution errors in a general setting, rather than obtain sharp bounds for the rate of such convergence. Therefore, we do not give an explicit form to the bound $\mathbb{B}_{n,\epsilon,\tau}$ or consider better bounds, based for example on the n-th shatter coefficient.*

The next lemma shows that, under the assumptions of this paper, the consistency of a generalized resubstitution error $\hat{\varepsilon}_n^{\mathcal{B}}$ is equivalent to the convergence to zero of its relative deviation from ε_n.

Lemma 25 *Fix a loss function ℓ. The generalized resubstitution error $\hat{\varepsilon}_n^{\mathcal{B}}$ is consistent if, and only if, for any $\tau > 0$ fixed,*

$$\frac{\left|\hat{\varepsilon}_n^{\mathcal{B}} - \varepsilon_n\right|}{\sqrt[\alpha]{L^\alpha(\psi_n)+\tau}} \to 0$$

with probability one as $n \to \infty$.

Proof The result follows from condition (11) and inequality

$$\frac{\left|\hat{\varepsilon}_n^{\mathcal{B}} - \varepsilon_n\right|}{\sqrt[\alpha]{\sup_{\psi\in\mathcal{H}} L^\alpha(\psi) + \tau}} \leqslant \frac{\left|\hat{\varepsilon}_n^{\mathcal{B}} - \varepsilon_n\right|}{\sqrt[\alpha]{L^\alpha(\psi_n) + \tau}} \leqslant \frac{\left|\hat{\varepsilon}_n^{\mathcal{B}} - \varepsilon_n\right|}{\sqrt[\alpha]{\tau}}.$$

■

Appendix C. Proof of results

Proof [Proof of Proposition 3] The result follows from (11), (13) and the Dominated Convergence Theorem (Braga-Neto, 2020, Theorem A7) since

$$\left|\hat{\varepsilon}_n^{\mathcal{B}} - \varepsilon_n\right| \leqslant 2\max\left\{\varepsilon_n, \hat{\varepsilon}_n^{\mathcal{B}}\right\} \leqslant 2\max\left\{\sup_{\psi\in\mathcal{H}} L(\psi), L_n^{\mathcal{B}}(\psi_n)\right\} < \infty$$

with probability one for all $n \geqslant 1$. ■

Proof [Proof of Theorem 4] The result follows from Proposition 21, Lemma 25 and Borel-Cantelli Lemma by noting that, for any $\epsilon > 0$ and $\tau > 0$ fixed,

$$\begin{aligned}
\mathbb{P}\left(\frac{|\hat{\varepsilon}_n^r - \varepsilon_n|}{\sqrt[\alpha]{L^\alpha(\psi_n) + \tau}} > \epsilon\right) &= \mathbb{P}\left(\frac{|L_n(\psi_n) - L(\psi_n)|}{\sqrt[\alpha]{L^\alpha(\psi_n) + \tau}} > \epsilon\right) \\
&\leqslant \mathbb{P}\left(\left\{\exists \psi \in \mathcal{H} : \frac{|L_n(\psi) - L(\psi)|}{\sqrt[\alpha]{L^\alpha(\psi) + \tau}} > \epsilon\right\}\right) \\
&= \mathbb{P}\left(\sup_{\psi\in\mathcal{H}} \frac{|L(\psi) - L_n(\psi)|}{\sqrt[\alpha]{L^\alpha(\psi) + \tau}} > \epsilon\right).
\end{aligned}$$

■

Proof [Proof of Proposition 5] Observe that

$$\left|\hat{\varepsilon}_n^{\mathcal{B}} - \varepsilon_n\right| \leqslant \left|\hat{\varepsilon}_n^{\mathcal{B}} - \hat{\varepsilon}_n^r\right| + \left|\hat{\varepsilon}_n^r - \varepsilon_n\right|. \tag{26}$$

The result follows since the first term on the right-hand side of (26) converges to zero with probability one by hypothesis and the second one converges to zero with probability one by Theorem 4. ■

Proof [Proof of Corollary 6] We show that $|\hat{\varepsilon}_n^{\mathcal{B}} - \hat{\varepsilon}_n^r| \to 0$ with probability one, so the result follows from Proposition 5. Denote by $\hat{\nu}_n$ and ν_n the empirical measures

$$\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^{n} \beta_{Z_i,\psi_n,S_n} \qquad\qquad \nu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i}.$$

Since ℓ is bounded by a constant C_ℓ, we have that

$$\hat{\varepsilon}_n^r = \nu_n[\ell(Z,\psi_n)] = \lim_{m\to\infty} \sum_{k=1}^{m-1} \frac{C_\ell}{m} \nu_n\left(\ell(Z,\psi_n) > \frac{kC_\ell}{m}\right),$$

and that

$$\hat{\varepsilon}_n^{\mathcal{B}} = \hat{\nu}_n[\ell(Z,\psi_n)] = \lim_{m\to\infty} \sum_{k=1}^{m-1} \frac{C_\ell}{m} \hat{\nu}_n\left(\ell(Z,\psi_n) > \frac{kC_\ell}{m}\right),$$

in which $\nu_n[\cdot]$ and $\hat{\nu}_n[\cdot]$ mean expectation under ν_n and $\hat{\nu}_n$, respectively. From the representation of $\hat{\varepsilon}_n^r$ and $\hat{\varepsilon}_n^{\mathcal{B}}$ described above, we have that

$$\begin{aligned}
\left|\hat{\varepsilon}_n^{\mathcal{B}} - \hat{\varepsilon}_n^r\right| &= \left|\lim_{m\to\infty} \sum_{k=1}^{m-1} \frac{C_\ell}{m}\left(\hat{\nu}_n\left(\ell(Z,\psi_n) > \frac{kC_\ell}{m}\right) - \nu_n\left(\ell(Z,\psi_n) > \frac{kC_\ell}{m}\right)\right)\right| \\
&\leqslant \left|\lim_{m\to\infty} \sum_{k=1}^{m-1} \frac{C_\ell}{m} \sup_{\substack{\psi\in\mathcal{H} \\ 0\leqslant b\leqslant C_\ell}} \left(\hat{\nu}_n\left(\ell(Z,\psi) > b\right) - \nu_n\left(\ell(Z,\psi) > b\right)\right)\right| \\
&\leqslant C_\ell \sup_{\substack{\psi\in\mathcal{H} \\ 0\leqslant b\leqslant C_\ell}} \left|\hat{\nu}_n\left(\ell(Z,\psi) > b\right) - \nu_n\left(\ell(Z,\psi) > b\right)\right| \\
&= C_\ell \sup_{A\in\mathcal{A}^\star_{\mathcal{H},\ell}} \left|\hat{\nu}_n\left(A\right) - \nu_n\left(A\right)\right|
\end{aligned}$$

for all samples $S_n \in \mathcal{Z}^n$ and the result follows. The second assertion is a consequence of the inequality

$$\sup_{A\in\mathcal{A}^\star_{\mathcal{H},\ell}} \left|\hat{\nu}_n\left(A\right) - \nu_n\left(A\right)\right| \leqslant \sup_{\substack{A\in\mathcal{A}^\star_{\mathcal{H},\ell} \\ z\in\mathcal{Z},\psi\in\mathcal{H}}} \left|\beta_{z,\psi,S_n}(A) - \delta_z(A)\right|$$

which holds for all samples $S_n \in \mathcal{Z}^n$. ■

Proof [Proof of Corollary 7] We show that $|\hat{\varepsilon}_n^{\mathcal{B}} - \hat{\varepsilon}_n^r| \to 0$ with probability one, so the result follows from Proposition 5. The result follows since

$$\begin{aligned} |\hat{\varepsilon}_n^{\mathcal{B}} - \hat{\varepsilon}_n^r| &\leqslant \frac{1}{n} \sum_{i=1}^{n} |\ell_{\mathcal{B}}(Z_i, \psi_n) - \ell(Z_i, \psi_n)| \\ &\leqslant \sup_{z \in \mathcal{Z}, \psi \in \mathcal{H}} |\ell_{\mathcal{B}}(z, \psi) - \ell(z, \psi)| \end{aligned}$$

for all samples $S_n \in \mathcal{Z}^n$. ■

Proof [Proof of Proposition 8] Observe that

$$\left|\hat{\varepsilon}_n^{\mathcal{B}_n} - \varepsilon_n\right| \leqslant \left|\hat{\varepsilon}_n^{\mathcal{B}_n} - \varepsilon_n^{\mathcal{B}_n}\right| + \left|\varepsilon_n^{\mathcal{B}_n} - \varepsilon_n\right|.$$

The result follows since the first term of the sum above converges to zero with probability one due to Theorem 4 and the second converges to zero with probability one by hypothesis. ■

Proof [Proof of Proposition 9] Observe that, for all $\psi \in \mathcal{H}$,

$$\begin{aligned} L^{\mathcal{B}_n}(\psi) &= \int_{\mathcal{Z}} \ell_{\mathcal{B}_n}(z, \psi)\, d\nu(z) \\ &= \int_{\mathcal{Z}} \int_{\mathcal{Z}} \ell(z', \psi)\, d\beta_{z,\psi,n}(z')\, d\nu(z) \\ &= \int_{\mathcal{Z}} \int_{\mathcal{Z}} \ell(z', \psi) \rho_{z,\psi,n}(z')\, d\nu(z')\, d\nu(z) \\ &= \int_{\mathcal{Z}} \ell(z', \psi) \int_{\mathcal{Z}} \rho_{z,\psi,n}(z')\, d\nu(z)\, d\nu(z') \\ &= \int_{\mathcal{Z}} \ell(z', \psi)\, d\nu(z') = L(\psi) \end{aligned}$$

hence the condition of Proposition 8 holds trivially. ■

Proof [Proof of Corollary 10] The result follows since (15) implies that

$$\begin{aligned} L^{\mathcal{B}_n}(\psi) &= \int_{\mathcal{Z}} \int_{\mathcal{Z}} \ell(z', \psi) \rho_{z,\psi,n}(z')\, d\nu(z')\, d\nu(z) \\ &= \int_{\mathcal{Z}} \int_{\mathcal{Z}} \ell(z', \psi) \rho_{z',\psi,n}(z)\, d\nu(z')\, d\nu(z) \\ &= \int_{\mathcal{Z}} \ell(z', \psi) \int_{\mathcal{Z}} \rho_{z',\psi,n}(z)\, d\nu(z)\, d\nu(z') = L(\psi). \end{aligned}$$

■

Proof [Proof of Proposition 11] The result follows from Corollary 10 since (a) follows from the assumption that X is absolutely continuous wrt Lebesgue measure and (b) follows since $p_{x,\psi,n}(x') = p_{x',\psi,n}(x)$ for all $x, x' \in \mathbb{R}^d$ and $\psi \in \mathcal{H}$. ■

Proof [Proof of Proposition 12] For $z \in \mathcal{Z}, \psi \in \mathcal{H}$ and $S_n \in \mathcal{Z}^n$ it holds

$$\begin{aligned}
\ell_{\mathcal{B}}(z,\psi) &= \int_{\mathcal{Z}} \ell(z',\psi)\; d\beta_{z,\psi,S_n}(z') \\
&= \ell(z,\psi) + \mathbb{E}\left[(Z_{z,\psi,S_n} - z)\right] \cdot \nabla \ell(z,\psi) \\
&+ \frac{1}{2}\sum_{i,j=1}^{d+m} \int_{\mathcal{Z}} (z'_i - z_i)(z'_j - z_j) \frac{\partial^2 \ell}{\partial z_i \partial z_j}(\bar{z},\psi)\; d\beta_{z,\psi,S_n}(z') \\
&= \ell(z,\psi) + \frac{1}{2}\sum_{i,j=1}^{d+m} \int_{\mathcal{Z}} (z'_i - z_i)(z'_j - z_j) \frac{\partial^2 \ell}{\partial z_i \partial z_j}(\bar{z},\psi)\; d\beta_{z,\psi,S_n}(z')
\end{aligned}$$

in which $\bar{z}$ depends on z'. Observe that the first equality follows from the Taylor expansion of $\ell(\cdot,\psi)$ around z and the second follows since $\mathbb{E}(Z_{z,\psi,S_n}) = z$. Now,

$$\begin{aligned}
&|\ell_{\mathcal{B}}(z,\psi) - \ell(z,\psi)| \\
&\leqslant \frac{1}{2}\sum_{i,j=1}^{d+m} \left[\sup_{z\in\mathcal{Z}} \left|\frac{\partial^2 \ell}{\partial z_i \partial z_j}(z,\psi)\right|\right] \int_{\mathcal{Z}} |(z'_i - z_i)||(z'_j - z_j)|\; d\beta_{z,\psi,S_n}(z') \\
&\leqslant \frac{C_2}{2}\sum_{i,j=1}^{d+m} \sqrt{Var\left([Z_{z,\psi,S_n}]_i\right) Var\left([Z_{z,\psi,S_n}]_j\right)}
\end{aligned}$$

in which the second inequality follows from (17) and Cauchy-Schwarz inequality. The result follows by Corollary 7 since the variances converge to zero with probability one as $n \to \infty$. ■

Proof [Proof of Proposition 13] We will show that the conditions of Proposition 12 are in force. Let β_{z,ψ,S_n} be a degenerate Gaussian law with mean z and covariance matrix $\tilde{K}_{z,\psi,S_n}$ satisfying $[\tilde{K}_{z,\psi,S_n}]_{i,j} = [K_{z,\psi,S_n}]_{i,j}$ if $i,j < d+1$ and $[\tilde{K}_{z,\psi,S_n}]_{i,j} = 0$ if $i = d+1$ or $j = d+1$. Under these assumptions $\hat{\varepsilon}_n^{\mathcal{B}} = \hat{\varepsilon}_n^{gbr}$.

Recall that Z_{z,ψ,S_n} is a random variable with distribution β_{z,ψ,S_n} for $z \in \mathcal{Z}$ and $\psi \in \mathcal{H}$. Condition (b) of Proposition 12 is a direct consequence of (19). It remains to show that condition (a) of Proposition 12 is in force.

Since $\mathcal{Y}$ is compact, and the derivatives and second derivatives of ψ are uniformly bounded in $\mathcal{X}$ and $\mathcal{H}$, it follows that

$$\frac{\partial^2 \ell}{\partial x_i \partial x_j}((x,y),\psi) = 2(\psi(x)-y)\frac{\partial^2 \psi}{\partial x_i \partial x_j}(x) + 2\frac{\partial \psi}{\partial x_i}(x)\frac{\partial \psi}{\partial x_j}(x),$$

is uniformly bounded. Moreover,

$$\frac{\partial^2 \ell}{\partial x_i \partial y}((x,y),\psi) = -2\frac{\partial \psi}{\partial x_i}(x) \qquad \text{and} \qquad \frac{\partial^2 \ell}{\partial y^2}((x,y),\psi) = 2$$

are also uniformly bounded. Therefore, condition (a) of Proposition 12 also holds and the result follows. ■

Proof [Proof of Lemma 14] We will compute the mean of $\delta(\hat{X})$ and an analogous deduction, but considering only the points in the sample with $Y_i = y$, holds to compute the mean of $\delta_y(\hat{X})$ conditioned on $\hat{Y} = y$. Let I_n be a random variable and $N_i, i = 1,\ldots,n$, be random vectors taking values in $\{1,\ldots,n\}$ and $\mathcal{X} \times \mathcal{Y}$, respectively, all defined on the probability space $(\Omega, \mathcal{S}, \mathbb{P})$ and independent, such that $\mathbb{P}(I_n = i) = 1/n$ for $i = 1,\ldots,n$ and N_i has the probability law $\mu_{X_i,S_n}\delta_{Y_i}$. Define on $(\Omega, \mathcal{S}, \mathbb{P})$ the random vector $(\hat{X}, \hat{Y})$ by

$$(\hat{X}, \hat{Y})(\omega) = N_{I_n(\omega)}(\omega)$$

for $\omega \in \Omega$. It follows that $(\hat{X}, \hat{Y})$ has probability law ν_n and we will compute the mean $\mathbb{E}[\delta(\hat{X})]$.

Denote by p_{i,S_n} the density of a Gaussian distribution with mean X_i and covariance matrix $\sigma^2_{S_n}\Sigma$. It follows that

$$\begin{aligned}
\mathbb{E}\left[\delta(\hat{X})\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\delta(\hat{X})|I_n = i] \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{X}}\delta(x)p_{i,S_n}(x)\,dx \\
&= \frac{1}{n}\sum_{i,j=1}^{n}\int_{\mathcal{X}}\delta(x,X_j)p_{i,S_n}(x)\mathbb{1}_{\delta(x)=\delta(x,X_j)}\,dx.
\end{aligned}$$

■

Proof [Proof of Proposition 15] Define

$$M = \frac{1}{2} \min_{j \neq i} \delta(X_i, X_j)$$

and observe that

$$\sigma_{S_n} \int_{\mathbb{R}^d} \frac{\delta(x, X_i)}{\sigma_{S_n}} \, p_{i,S_n}(x) \, \mathbb{1}_{\delta(x,X_i)<M} \, dx \leqslant \mathbb{E}[\delta(\hat{X})|I_n = i] \leqslant \sigma_{S_n} \int_{\mathbb{R}^d} \frac{\delta(x, X_i)}{\sigma_{S_n}} \, p_{i,S_n}(x) \, dx$$

If $\hat{X}$ has a Gaussian distribution with mean X_i and kernel $\sigma_{S_n}\Sigma$, then $\sigma_{S_n}^{-1}\delta(\hat{X}, X_i)$ has a chi distribution with d degrees of freedom. Denoting by χ a random variable with this distribution, it follows that

$$\sigma_{S_n}\mathbb{E}[\chi \mathbb{1}_{\chi<M/\sigma_{S_n}}] \leqslant \mathbb{E}[\delta(\hat{X})|I_n = i] \leqslant \sigma_{S_n}\mathbb{E}[\chi]$$

The result follows by noting that

$$\lim_{\sigma_{S_n}\to 0} \frac{\mathbb{E}[\delta(\hat{X})]}{\sigma_{S_n}} = \frac{1}{n} \sum_{i=1}^{n} \lim_{\sigma_{S_n}\to 0} \frac{\mathbb{E}[\delta(\hat{X})|I_n = i]}{\sigma_{S_n}}.$$

■

Appendix D. L^∞-convergence and VC dimension

In this section, we estimate the VC dimension of a class from the VC dimension of a class that is *close* to it, in a sense we will make clear below. To ease notation, let Θ be a general set and consider the following sets of non-negative functions with domain $\mathcal{Z}$ indexed by the elements in Θ:

$$\mathcal{F} = \{f_\theta : \theta \in \Theta\} \qquad \mathcal{G} = \{g_\theta : \theta \in \Theta\} \qquad \mathcal{G}_n = \{g_\theta^{(n)} : \theta \in \Theta\}$$

for $n \geqslant 1$. The functions in these sets should be understood as loss functions of a hypothesis space indexed by Θ, for example, $f_\theta(z) = \ell(z, \psi_\theta)$. We define the VC dimension of these sets as, for example,

$$d_{VC}(\mathcal{F}) = d_{VC}(\{\{z \in \mathcal{Z} : f_\theta(z) > \beta\} : \theta \in \Theta, \beta \in (0, C_\mathcal{F})\}).$$

We define the distance between two sets $\mathcal{F}$ and $\mathcal{G}$ as

$$d(\mathcal{F},\mathcal{G}) = \sup_{\theta\in\Theta} \|f_\theta - g_\theta\|_\infty. \tag{27}$$

For each integer $1 \leqslant d \leqslant d_{VC}(\mathcal{F})$ let

$$S(\mathcal{F},d) = \{\boldsymbol{z} = \{z_1,\ldots,z_d\} \subset \mathcal{Z} : \mathcal{F} \text{ shatters } \{z_1,\ldots,z_d\}\}$$

be the collection of d points that $\mathcal{F}$ can shatter. We denote simply $S(\mathcal{F}) \coloneqq S(\mathcal{F}, d_{VC}(\mathcal{F}))$. For each set $\boldsymbol{z}$ in $S(\mathcal{F},d)$ and $b \in \{-1,+1\}^s, s \geqslant d$, define

$$\mathcal{F}(\boldsymbol{z},b) = \left\{\theta\in\Theta : \sup_\beta \min_{i=1,\ldots,d} b_i(f_\theta(z_i)-\beta) > 0\right\}$$

as the functions in $\mathcal{F}$ that realize dichotomy b with the points $\boldsymbol{z}$.

For $d \geqslant 1$, we define the d-margin of a set $\mathcal{F}$ as

$$\delta_{\mathcal{F}}(d) = \begin{cases} \sup\limits_{\boldsymbol{z}\in S(\mathcal{F},d)} \min\limits_{b\in\{-1,1\}^d} \sup\limits_{\theta\in\mathcal{F}(\boldsymbol{z},b)} \sup\limits_\beta \min\limits_i b_i\left(f_\theta(z_i)-\beta\right) & \text{if } d \leqslant d_{VC}(\mathcal{F}) \\ \sup\limits_{\boldsymbol{z}\in S(\mathcal{F})} \sup\limits_{z_{d_{VC}(\mathcal{F})+1},\ldots,z_d} \min\limits_{b\in\{-1,1\}^d} \sup\limits_{\theta\in\mathcal{F}(\boldsymbol{z},b)} \sup\limits_\beta \min\limits_i b_i\left(f_\theta(z_i)-\beta\right), & \text{if } d > d_{VC}(\mathcal{F}). \end{cases}$$

Observe that $\delta_{\mathcal{F}}(d)$ is a non-increasing function of d. We denote $\delta_{\mathcal{F}} \coloneqq \delta_{\mathcal{F}}(d_{VC}(\mathcal{F}))$. The VC dimension of $\mathcal{F}$ can be defined based on $\delta_{\mathcal{F}}(d)$.

Lemma 26

$$d_{VC}(\mathcal{F}) = d \iff \delta_{\mathcal{F}}(d) > 0 \text{ and } \delta_{\mathcal{F}}(d+1) \leqslant 0$$

Proof We show that $\delta_{\mathcal{F}}(d) > 0$ if, and only if, $\mathcal{F}$ shatters some d points in $\mathcal{Z}$ so the result follows. On the one hand, if $\mathcal{F}$ shatters $\{z_1,\ldots,z_d\}$, then for all $b \in \{-1,1\}^d$ there exists $\theta_b \in \Theta$ and $\beta_b \in (0, C_{\mathcal{F}})$ such that

$$b_i\left(f_{\theta_b}(z_i) - \beta_b\right) > 0 \text{ for all } i = 1,\ldots,d,$$

and hence $\delta_{\mathcal{F}}(d) > 0$. On the other hand, if $\delta_{\mathcal{F}}(d) = \delta > 0$ then there exists a sequence $\{z_1,\ldots,z_d\}$ such that, for all $b \in \{-1,1\}^d$ there exists $\theta_b \in \Theta$ and $\beta_b \in (0, C_{\mathcal{F}})$ such that

$$b_i\left(f_{\theta_b}(z_i) - \beta_b\right) > \delta/2 > 0 \text{ for all } i = 1,\ldots,d,$$

and hence $\mathcal{F}$ shatters some d points in $\mathcal{Z}$. ■

The d-margins are also associated with the fat-shattering dimension (Bartlett et al., 1994) of a set of real-valued functions which can be defined based on them.

Definition 27 *The γ fat-shattering dimension of a set $\mathcal{F}$ is the greatest integer d such that $\delta_{\mathcal{F}}(d) \geqslant \gamma$.*

It follows from Lemma 26 that if $\mathcal{G}$ is close enough to $\mathcal{F}$, then its VC dimension cannot differ too much from that of $\mathcal{F}$.

Lemma 28 *For $d \geqslant 1$,*

$$d(\mathcal{F}, \mathcal{G}) < |\delta_{\mathcal{F}}(d_{VC}(\mathcal{F}) + d)| \implies d_{VC}(\mathcal{G}) < d_{VC}(\mathcal{F}) + d.$$

and

$$d(\mathcal{F}, \mathcal{G}) < \delta_{\mathcal{F}} \implies d_{VC}(\mathcal{F}) \leqslant d_{VC}(\mathcal{G}).$$

In particular,

$$d(\mathcal{F}, \mathcal{G}) < \min\{\delta_{\mathcal{F}}, \delta_{\mathcal{G}}\} \implies d_{VC}(\mathcal{F}) = d_{VC}(\mathcal{G}).$$

Proof We assume that $\delta_{\mathcal{F}}(d_{VC}(\mathcal{F}) + d) < 0$ otherwise the first assertion holds trivially since then $\mathcal{G} = \mathcal{F}$. Fix a $\delta > 0$. If $d(\mathcal{F}, \mathcal{G}) < \delta$ then, for any sequence $\{z_1, \ldots, z_d\}$ and $b \in \{-1, 1\}^d$, it holds

$$\left| \sup_{\theta,\beta} \min_i b_i \left(f_\theta(z_i) - \beta\right) - \sup_{\theta,\beta} \min_i b_i \left(g_\theta(z_i) - \beta\right) \right| < \delta$$

and therefore

$$\delta_{\mathcal{F}}(d) - \delta < \delta_{\mathcal{G}}(d) < \delta_{\mathcal{F}}(d) + \delta$$

for all $d \geqslant 1$. The result follows from the inequality above and Lemma 26 since, if $d(\mathcal{F}, \mathcal{G}) < |\delta_{\mathcal{F}}(d_{VC}(\mathcal{F}) + d)|$, then

$$\delta_{\mathcal{G}}(d_{VC}(\mathcal{F}) + d) < \delta_{\mathcal{F}}(d_{VC}(\mathcal{F}) + d) + |\delta_{\mathcal{F}}(d_{VC}(\mathcal{F}) + d)| = 0 \tag{28}$$

and if $d(\mathcal{F}, \mathcal{G}) < \delta_{\mathcal{F}}$, then $0 < \delta_{\mathcal{G}}(d_{VC}(\mathcal{F}))$. ■

It follows from Lemma 28 that the VC dimension is a continuous function with respect to the distance (27) when $\delta_{\mathcal{F}}(d_{VC}(\mathcal{F}) + 1) < 0$.

Lemma 29 *If $\delta_{\mathcal{F}}(d_{VC}(\mathcal{F}) + d) < 0$ for a $d \geqslant 1$ then*

$$\lim_{n\to\infty} d(\mathcal{G}_n, \mathcal{F}) = 0 \implies d_{VC}(\mathcal{F}) \leqslant \lim_{n\to\infty} d_{VC}(\mathcal{G}_n) < d_{VC}(\mathcal{F}) + d.$$

In particular, if $\delta_{\mathcal{F}}(d_{VC}(\mathcal{F}) + 1) < 0$ then

$$\lim_{n\to\infty} d(\mathcal{G}_n, \mathcal{F}) = 0 \implies \lim_{n\to\infty} d_{VC}(\mathcal{G}_n) = d_{VC}(\mathcal{F}).$$

Proof That $\lim d_{VC}(\mathcal{G}_n) \geqslant d_{VC}(\mathcal{F})$ follows from Lemma 28, and the fact that there exists a n_0 such that $d(\mathcal{G}_n, \mathcal{F}) < \delta_{\mathcal{F}}$ for all $n > n_0$ when this distance converges to zero. That $\lim d_{VC}(\mathcal{G}_n) < d_{VC}(\mathcal{F}) + d$ for $d \geqslant 1$ whenever $\delta_{\mathcal{F}}(d_{VC}(\mathcal{F}) + d) < 0$ also follows from Lemma 28 and the fact that there exists a n_0 such that $d(\mathcal{G}_n, \mathcal{F}) < |\delta_{\mathcal{F}}(d_{VC}(\mathcal{F}) + d)|$ for all $n > n_0$ when this distance converges to zero. ■

Denote by $\delta_{\mathcal{H},\ell}(d)$ the d-margin of hypothesis space $\mathcal{H}$ under loss function ℓ. We define its marginal VC dimension.

Definition 30 *The marginal VC dimension of a hypothesis space $\mathcal{H}$ under loss function ℓ is the least integer d such that $\delta_{\mathcal{H},\ell}(d) < 0$ and is denoted by $m_{VC}(\mathcal{H}, \ell)$. If $\delta_{\mathcal{H},\ell}(d) \geqslant 0$ for all $d \geqslant 1$ then $m_{VC}(\mathcal{H}, \ell) = \infty$.*

It follows from Lemma 26 that $d_{VC}(\mathcal{H}, \ell) < m_{VC}(\mathcal{H}, \ell)$. Let $\{\ell_n\}$ be a sequence of loss functions that converges in L^∞ to ℓ. The main result of this section states that if $m_{VC}(\mathcal{H}, \ell) < \infty$ then there exists a n_0 such that $d_{VC}(\mathcal{H}, \ell_n) < \infty$ for all $n > n_0$.

Proposition 31 *Fix a hypothesis space $\mathcal{H}$ and let $\{\ell_n\}$ and ℓ be loss functions such that*

$$\lim_{n\to\infty} \sup_{z\in\mathcal{Z},\psi\in\mathcal{H}} |\ell_n(z,\psi) - \ell(z,\psi)| = 0.$$

If $m_{VC}(\mathcal{H}, \ell) < \infty$, then

$$\limsup_{n\to\infty} d_{VC}(\mathcal{H}, \ell_n) < \infty.$$

Proof The result follows from inequality

$$\limsup_{n\to\infty} d_{VC}(\mathcal{H}, \ell_n) < m_{VC}(\mathcal{H}, \ell)$$

which is a direct consequence of Lemma 29. ■

The next lemma is a non-asymptotic version of Proposition 31.

Lemma 32 *Fix a hypothesis space $\mathcal{H}$ and let ℓ_1 and ℓ_2 be loss functions such that*

$$\sup_{z\in\mathcal{Z},\psi\in\mathcal{H}} |\ell_1(z,\psi) - \ell_2(z,\psi)| = \gamma$$

for a $\gamma > 0$. If there exists a $d \geqslant 1$ such that $\gamma < |\delta_{\mathcal{H},\ell_1}(d_{VC}(\mathcal{H},\ell_1)+d)|$, then

$$d_{VC}(\mathcal{H},\ell_2) < d_{VC}(\mathcal{H},\ell_1) + \min\{d \geqslant 1 : \gamma < |\delta_{\mathcal{H},\ell_1}(d_{VC}(\mathcal{H},\ell_1)+d)|\}.$$

In particular, if $d_{VC}(\mathcal{H},\ell_1) < \infty$ then $d_{VC}(\mathcal{H},\ell_2) < \infty$.

Proof The result is a direct consequence of Lemma 28. ■

If ℓ is a binary loss function, then $m_{VC}(\mathcal{H},\ell) = d_{VC}(\mathcal{H},\ell) + 2$.

Proposition 33 *If ℓ is a binary loss function and $\mathcal{H}$ is a hypothesis space with $d_{VC}(\mathcal{H},\ell) < \infty$, then $\delta_{\mathcal{H},\ell}(d_{VC}(\mathcal{H},\ell)+2) = -1/2$. In particular, $m_{VC}(\mathcal{H},\ell) = d_{VC}(\mathcal{H},\ell) + 2$.*

Proof To ease notation, we will show this result for $\mathcal{F}$ assuming that it is a set of binary functions. Denote $d = d_{VC}(\mathcal{F})$, and fix $\boldsymbol{z} \in S(\mathcal{F})$ and $\{z_{d+1}, z_{d+2}\} \subset \mathcal{Z}$. Observe that

$$\min_{b\in\{-1,1\}^{d+2}} \sup_{\theta\in\mathcal{F}(\boldsymbol{z},b)} \sup_{\beta} \min_{i=1,\ldots,d+2} b_i\left(f_\theta(z_i)-\beta\right) = \sup_{\beta} \min\{-\beta, -(1-\beta)\} = -1/2$$

as long as there exists a dichotomy $b \in \{-1,1\}^{d+2}$ with $b_{d+1} = 1$ and $b_{d+2} = -1$ such that for all $\theta \in \mathcal{F}(\boldsymbol{z},b)$, $f_\theta(z_{d+1}) = 0$ and $f_\theta(z_{d+2}) = 1$, or vice versa. Hence, it suffices to show this condition for all $\theta \in \mathcal{F}(\boldsymbol{z},b)$.

Define the following subsets of $\{-1,1\}^{d+2}$ for a $\boldsymbol{z} \in S(\mathcal{F})$ and $e \in \{-1,1\}$:

$$B(d+1,e) = \left\{b \in \{-1,1\}^{d+2} : b_{d+1} = e \text{ and } \mathcal{F}(\boldsymbol{z},b) \text{ cannot shatter } \boldsymbol{z} \cup \{z_{d+1}\} \text{ as } b\right\}$$
$$B(d+2,e) = \left\{b \in \{-1,1\}^{d+2} : b_{d+2} = e \text{ and } \mathcal{F}(\boldsymbol{z},b) \text{ cannot shatter } \boldsymbol{z} \cup \{z_{d+2}\} \text{ as } b\right\}$$

as the dichotomies b with $b_{d+1} = e$ and $b_{d+2} = e$ that $\mathcal{F}(\boldsymbol{z},b)$ cannot realize with $\boldsymbol{z} \cup \{z_{d+1}\}$ and $\boldsymbol{z} \cup \{z_{d+2}\}$, respectively. Since $\mathcal{F}$ cannot shatter $d+2$ points, it follows that $B(d+1,1) \cup B(d+1,-1) \neq \varnothing$ and the same is true for $d+2$.

We claim that if $B(d+1,1) \cap B(d+2,-1) \neq \varnothing$, then by taking $b \in B(d+1,1) \cap B(d+2,-1)$ it follows that for all $\theta \in \mathcal{F}(\boldsymbol{z},b)$, $f_\theta(z_{d+1}) = 0$ and $f_\theta(z_{d+2}) = 1$. Indeed, if $b \in B(d+1,1) \cap B(d+2,-1)$ then

$$\sup_{\theta\in\mathcal{F}(\boldsymbol{z},b)} \sup_{\beta} \left(f_\theta(z_{d+1}) - \beta\right) \leqslant 0 \quad \text{and} \quad \sup_{\theta\in\mathcal{F}(\boldsymbol{z},b)} \sup_{\beta} -\left(f_\theta(z_{d+2}) - \beta\right) \leqslant 0$$

from which follows that for all $\theta \in \mathcal{F}(\boldsymbol{z}, b)$, $f_\theta(z_{d+1}) = 0$ and $f_\theta(z_{d+2}) = 1$. The condition of interest also follows if $B(d+1,-1) \cap B(d+2,1) \neq \varnothing$.

We proceed by contradiction. Assume that

$$(B(d+1,1) \cap B(d+2,-1)) \cup (B(d+1,-1) \cap B(d+2,1)) = \varnothing. \quad (29)$$

This implies that $\mathcal{F}$ can realize all dichotomies in $\{-1,1\}^{d+2}$ such that $b_{d+1} \neq b_{d+2}$. In particular, it can realize all the dichotomies of the form (b', b_{d+2}) with $b' \in \{-1,1\}^{d+1}$ and $b_{d+2} \neq b_{d+1}$. But this implies that $\mathcal{F}$ shatters $\{z_1, \ldots, z_{d+1}\}$ which is a contradiction since $d_{VC}(\mathcal{F}) = d$. Hence, (29) does not hold.

Fix $\boldsymbol{z} \in S(\mathcal{F})$ and $z_{d+1} \in \mathcal{Z}$ and observe that

$$\min_{b \in \{-1,1\}^{d+1}} \sup_{\theta \in \mathcal{F}(\boldsymbol{z}, b)} \sup_{\beta} \min_{i=1,\ldots,d+1} b_i \left(f_\theta(z_i) - \beta\right) = \sup_{\beta} -\beta = 0$$

since $\mathcal{F}$ cannot shatter $d+1$ points. This result implies that $\delta_{\mathcal{F}}(d_{VC}(\mathcal{F}) + 1) = 0$. Therefore, $m_{VC}(\mathcal{H}, \ell) = d_{VC}(\mathcal{H}, \ell) + 2$ when ℓ is a binary loss function. ■

We combine Propositions 33 and Lemmas 29 and 32 to establish that if a sequence of loss functions $\{\ell_n\}$ converges to a binary loss ℓ, then the limiting VC dimension does not differ in more than one unit from $d_{VC}(\mathcal{H}, \ell)$.

Corollary 34 *Fix a hypothesis space $\mathcal{H}$ and a binary loss function with $d_{VC}(\mathcal{H}, \ell) < \infty$. Let $\{\ell_n\}$ be a sequence of loss functions such that*

$$\lim_{n \to \infty} \sup_{z \in \mathcal{Z}, \psi \in \mathcal{H}} |\ell_n(z, \psi) - \ell(z, \psi)| = 0.$$

Then

$$\limsup_{n \to \infty} d_{VC}(\mathcal{H}, \ell_n) \leqslant d_{VC}(\mathcal{H}, \ell) + 1.$$

In particular, if $\sup_{z \in \mathcal{Z}, \psi \in \mathcal{H}} |\ell_n(z, \psi) - \ell(z, \psi)| < 1/2$ for some n, then

$$d_{VC}(\mathcal{H}, \ell_n) \leqslant d_{VC}(\mathcal{H}, \ell) + 1.$$

Appendix E. VC dimension under real-valued loss functions

E.1 VC dimension under the quadratic loss function

If ℓ is the quadratic loss function, denoting, for $x \in \mathcal{X}, y \in \mathcal{Y}, 0 < \beta < C_\ell$ and $\psi \in \mathcal{H}$,

$$\begin{aligned} I_{\psi,\beta}(x,y) &= \mathbb{1}\left\{(\psi(x) - y)^2 > \beta\right\} \\ &= \mathbb{1}\left\{\psi(x) - y - \sqrt{\beta} > 0 \text{ or } \psi(x) - y + \sqrt{\beta} < 0\right\}, \end{aligned}$$

we can associate $\mathcal{H}$ with set

$$\mathcal{F}_{\mathcal{H},\ell} = \{I_{\psi,\beta} : \psi \in \mathcal{H}, \beta \in (0, C_\ell)\}$$

of classifiers in $\mathcal{Z}$ satisfying $d_{VC}(\mathcal{H}, \ell) = d_{VC}(\mathcal{F}_{\mathcal{H},\ell})$. Furthermore, we can bound $d_{VC}(\mathcal{H}, \ell)$ by a function of the VC dimension of the sets

$$\mathcal{F}^+_{\mathcal{H},\ell} = \left\{I^+_{\psi,\beta}(x, y) = \mathbb{1}\{\psi(x) - y - \sqrt{\beta} > 0\} : \psi \in \mathcal{H}, \beta \in (0, C_\ell)\right\}$$
$$\mathcal{F}^-_{\mathcal{H},\ell} = \left\{I^-_{\psi,\beta}(x, y) = \mathbb{1}\{\psi(x) - y + \sqrt{\beta} < 0\} : \psi \in \mathcal{H}, \beta \in (0, C_\ell)\right\}$$

of binary functions from $\mathcal{X} \times \mathcal{Y}$ to $\{0, 1\}$.

Lemma 35 *Fix a hypothesis space $\mathcal{H}$ and let ℓ be the quadratic loss function. Then,*

$$d_{VC}(\mathcal{H}, \ell) \leqslant 4(1 + \log 2) \max\{d_{VC}(\mathcal{F}^+_{\mathcal{H},\ell}), d_{VC}(\mathcal{F}^-_{\mathcal{H},\ell})\}.$$

In particular, if $d_{VC}(\mathcal{F}^+_{\mathcal{H},\ell}) < \infty$ and $d_{VC}(\mathcal{F}^-_{\mathcal{H},\ell}) < \infty$, then $d_{VC}(\mathcal{H}, \ell) < \infty$.

Proof [Proof of Lemma 35] Let $g : \{0, 1\}^2 \to \{0, 1\}$ be the Boolean function given by $g(b_1, b_2) = \max\{b_1, b_2\}$ and observe that

$$I_{\psi,\beta}(x, y) = g(I^+_{\psi,\beta}(x, y), I^-_{\psi,\beta}(x, y))$$

for $x \in \mathcal{X}, y \in \mathcal{Y}, 0 < \beta < C$ and $\psi \in \mathcal{H}$. Lemma 2 in Sontag et al. (1998) states that, given a Boolean function g, and two sets $\mathcal{H}_1$ and $\mathcal{H}_2$ of binary functions, if $\mathcal{F} = \{g(h_1, h_2) : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$ then

$$d_{VC}(\mathcal{F}) \leqslant 4(1 + \log 2) \max\{d_{VC}(\mathcal{H}_1), d_{VC}(\mathcal{H}_2)\}. \tag{30}$$

Since

$$\mathcal{F}_{\mathcal{H},\ell} \subseteq \{g(h_1, h_2) : h_1 \in \mathcal{F}^+_{\mathcal{H},\ell}, h_2 \in \mathcal{F}^-_{\mathcal{H},\ell}\},$$

it follows from (30) that

$$\begin{aligned} d_{VC}(\mathcal{H}, \ell) &= d_{VC}(\mathcal{F}_{\mathcal{H},\ell}) \\ &\leqslant d_{VC}(\{g(h_1, h_2) : h_1 \in \mathcal{F}^+_{\mathcal{H},\ell}, h_2 \in \mathcal{F}^-_{\mathcal{H},\ell}\}) \\ &\leqslant 4(1 + \log 2) \max\{d_{VC}(\mathcal{F}^+_{\mathcal{H},\ell}), d_{VC}(\mathcal{F}^-_{\mathcal{H},\ell})\} \end{aligned}$$

in which the first inequality follows from the fact that the VC dimension is non-decreasing on inclusion. ■

The bound of Lemma 35 is clearly not tight, but can be a tool to show that $d_{VC}(\mathcal{H}, \ell)$ is finite. For example, in the case of linear regression, in which

$$\mathcal{H} = \left\{ \psi(x) = a_0 + \sum_{i=1}^{d} a_i x_i : a_i \in \mathbb{R}, i = 0, \ldots, d \right\},$$

we have that, after some simple algebraic computations,

$$\mathcal{F}^{+}_{\mathcal{H},\ell} = \mathcal{F}^{-}_{\mathcal{H},\ell} = \left\{ \mathbb{1}\left\{ a_0 + \sum_{i=1}^{d} a_i x_i + a_{d+1} - y > 0 \right\} : a_i \in \mathbb{R}, i = 0, \ldots, d+1 \right\},$$

so $\mathcal{F}^{+}_{\mathcal{H},\ell}$ and $\mathcal{F}^{-}_{\mathcal{H},\ell}$ is the set of linear classifiers on $d+1$ variables. It is well-known that $d_{VC}(\mathcal{F}^{+}_{\mathcal{H},\ell}) = d_{VC}(\mathcal{F}^{-}_{\mathcal{H},\ell}) = d+2$ and hence

$$d_{VC}(\mathcal{H}, \ell) \leqslant 4(1 + \log 2)(d+2),$$

which is linear in the number of variables d.

E.2 VC dimension under $\ell_{\mathcal{B}_n}$

There is no general relation between $d_{VC}(\mathcal{H}, \ell)$ and $d_{VC}(\mathcal{H}, \ell_{\mathcal{B}})$ which holds for any loss function ℓ and smoothing probabilities $\mathcal{B}$. Actually, it is possible to construct pathological cases in which $d_{VC}(\mathcal{H}, \ell) < \infty$, but $d_{VC}(\mathcal{H}, \ell_{\mathcal{B}}) = \infty$. Nevertheless, it is possible to associate these VC dimensions in some specific cases.

We start by considering linear regression under the quadratic loss function.

Lemma 36 *Fix a class $\mathcal{B}_n$ of probability measures for $n \geqslant 1$ with mean z and kernel $K_{n,\psi}$ that does not depend on z, and let $\mathcal{H}$ be a hypothesis space of linear functions in $d \geqslant 1$ variables and ℓ be the quadratic loss function. Then,*

$$d_{VC}(\mathcal{H}, \ell_{\mathcal{B}_n}) = d_{VC}(\mathcal{H}, \ell).$$

Proof For each $\psi \in \mathcal{H}$, denote by

$$R(\psi) = \frac{1}{2} \sum_{i,j=1}^{d+1} \frac{\partial^2 \ell}{\partial z_i \partial z_j}(\bar{z}, \psi)[K_{n,\psi}]_{i,j}$$

the remainder of the degree two Taylor expansion of $\ell_{\mathcal{B}_n}(z, \psi)$ on z so it holds

$$\ell_{\mathcal{B}_n}(z, \psi) = \ell(z, \psi) + R(\psi).$$

Observe that $R(\psi)$ does not depend on z since $\ell(z,\psi)$ is a degree two polynomial on z. Since

$$\begin{aligned}\mathcal{A}^{\star}_{\mathcal{H},\ell_{\mathcal{B}_n}} &= \{\{z \in \mathcal{Z} : \ell(z,\psi) + R(\psi) > \beta\} : \psi \in \mathcal{H}, 0 < \beta < C_\ell + R(\psi)\} \\ &= \{\{z \in \mathcal{Z} : \ell(z,\psi) > \beta - R(\psi)\} : \psi \in \mathcal{H}, 0 < \beta < C_\ell\} \\ &= \{\{\mathcal{Z}\}\} \bigcup \{\{z \in \mathcal{Z} : \ell(z,\psi) > \beta\} : \psi \in \mathcal{H}, 0 < \beta < C_\ell\} \\ &= \{\{\mathcal{Z}\}\} \bigcup \mathcal{A}^{\star}_{\mathcal{H},\ell}.\end{aligned}$$

and $\{\{\mathcal{Z}\}\}$ generates the dichotomy of all ones, which can already be realized by sets in $\mathcal{A}^{\star}_{\mathcal{H},\ell}$ for all n, the VC dimension remains the same. ■

Remark 37 *From the proof of Lemma 36 follows that* $d_{VC}(\mathcal{H},\ell_{\mathcal{B}_n}) \leqslant d_{VC}(\mathcal{H},\ell)+1$ *whenever* $\ell_{\mathcal{B}_n}(z,\psi) = \ell(z,\psi) + f(\psi)$ *in which* f *is a function that depends solely on* ψ *and not on* z.

From Lemma 32 follows a general relation between $d_{VC}(\mathcal{H},\ell_{\mathcal{B}_n})$ and $d_{VC}(\mathcal{H},\ell)$ based on the d-margins of $\mathcal{H}$ under ℓ.

Corollary 38 *Fix a loss function* ℓ *and a collection* $\mathcal{B}_n$ *of probability measures for each* $n \geqslant 1$, *and let* $\mathcal{H}$ *be a hypothesis space with* $d_{VC}(\mathcal{H},\ell) < \infty$. *If there exists a* $d \geqslant 1$ *such that*

$$\limsup_{n\to\infty} \sup_{z\in\mathcal{Z},\psi\in\mathcal{H}} |\ell_{\mathcal{B}_n}(z,\psi) - \ell(z,\psi)| < |\delta_{\mathcal{H},\ell}(d_{VC}(\mathcal{H},\ell) + d)|$$

then

$$\limsup_{n\to\infty} d_{VC}(\mathcal{H},\ell_{\mathcal{B}_n}) < d_{VC}(\mathcal{H},\ell) + d < \infty.$$

Corollary 38 combined with Proposition 33 yields a sufficient condition for the finiteness of $\limsup_{n\to\infty} d_{VC}(\mathcal{H},\ell_{\mathcal{B}_n})$ in classification problems.

Corollary 39 *Fix a binary loss function* ℓ *and a collection* $\mathcal{B}_n$ *of probability measures for each* $n \geqslant 1$, *and let* $\mathcal{H}$ *be a hypothesis space with* $d_{VC}(\mathcal{H},\ell) < \infty$. *If*

$$\limsup_{n\to\infty} \sup_{z\in\mathcal{Z},\psi\in\mathcal{H}} |\ell_{\mathcal{B}_n}(z,\psi) - \ell(z,\psi)| < 1/2$$

then

$$\limsup_{n\to\infty} d_{VC}(\mathcal{H},\ell_{\mathcal{B}_n}) \leqslant d_{VC}(\mathcal{H},\ell) + 1 < \infty.$$