Journal of Machine Learning Research 22 (2021) 1-55 Submitted 7/19; Revised 11/20; Published 1/21

Aggregated Hold-Out

Guillaume Maillard GUILLAUME.MAILLARD@UNI.LU
Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France

Sylvain Arlot SYLVAIN.ARLOT@UNIVERSITE-PARIS-SACLAY.FR
Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France
Institut Universitaire de France (IUF)

Matthieu Lerasle MATTHIEU.LERASLE@UNIVERSITE-PARIS-SACLAY.FR
Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France

Editor: Shie Mannor

Abstract

Aggregated hold-out (agghoo) is a method which averages learning rules selected by hold-out (that is, cross-validation with a single split). We provide the first theoretical guarantees on agghoo, ensuring that it can be used safely: Agghoo performs at worst like the hold-out when the risk is convex. The same holds true in classification with the 0–1 risk, with an additional constant factor. For the hold-out, oracle inequalities are known for bounded losses, as in binary classification. We show that similar results can be proved, under appropriate assumptions, for other risk-minimization problems. In particular, we obtain an oracle inequality for regularized kernel regression with a Lipschitz loss, without requiring that the Y variable or the regressors be bounded. Numerical experiments show that aggregation brings a significant improvement over the hold-out and that agghoo is competitive with cross-validation.

Keywords: cross-validation, aggregation, bagging, hyperparameter selection, regularized kernel regression

1. Introduction

The problem of choosing from data among a family of learning rules is central to machine learning. There is typically a variety of rules which can be applied to a given problem — for instance, support vector machines, neural networks or random forests. Moreover, most machine learning rules depend on hyperparameters which have a strong impact on the final performance of the algorithm. For instance, k-nearest-neighbors rules (Biau and Devroye, 2015) depend on the number k of neighbors. A second example, among many others, is given by regularized empirical risk minimization rules, such as support vector machines (Steinwart and Christmann, 2008) or the lasso (Tibshirani, 1996; Bühlmann and van de Geer, 2011), which all depend on some regularization parameter. A related problem is model selection (Burnham and Anderson, 2002; Massart, 2007), where one has to choose among a family of candidate models.

In supervised learning, cross-validation (CV) is a general, efficient and classical answer to the problem of selecting a learning rule (Arlot and Celisse, 2010). It relies on the idea

of splitting data into a training sample —used for training a predictor with each rule in competition— and a validation sample —used for assessing the performance of each predictor. This leads to an estimator of the risk —the hold-out estimator when data are split once, the CV estimator when an average is taken over several data splits—, which can be minimized for selecting among a family of competing rules.

A completely different strategy, called aggregation, is to *combine* the predictors obtained with all candidates (Nemirovski, 2000; Yang, 2001; Tsybakov, 2004). Aggregation is the key step of ensemble methods (Dietterich, 2000), among which we can mention bagging (Breiman, 1996), adaboost (Freund and Schapire, 1997) and random forests (Breiman, 2001; Biau and Scornet, 2016). A major interest of aggregation is that it builds a learning rule that may not belong to the family of rules in competition. Therefore, it sometimes has a smaller risk than the best of all rules (Salmon and Dalalyan, 2011, Table 1). In contrast, cross-validation, which selects only one candidate, cannot outperform the best rule in the family.

1.1 Aggregated Hold-Out (Agghoo)

This paper studies a procedure mixing cross-validation and aggregation ideas, that we call *aggregated hold-out* (agghoo). Data are split several times; for each split, the hold-out selects one predictor; then, the predictors obtained with the different splits are aggregated. A formal definition is provided in Section 3. This procedure is as general as cross-validation and it has roughly the same computational cost (see Section 3.4). Agghoo is already popular among practicioners, and has appeared in the neuro-imaging literature (Hoyos-Idrobo et al., 2015; Varoquaux et al., 2017) under the name "CV + averaging". Yet, to the best of our knowledge, existing experimental studies do not give any indication on how to choose agghoo's parameters. No general mathematical definition has been provided, so it is unclear how to generalize agghoo beyond a given article's setting. Theoretical guarantees on agghoo have not been established yet, to the best of our knowledge. The closest results we found study other procedures, called ACV (Jung and Hu, 2015), EKCV (Jung, 2016), or Hall and Robinson (2009)'s bagged cross-validation (shortened into Hall's BCV, which should not be confused with other procedures combining bagging and cross-validation, see Section 3.3). These authors do not prove oracle inequalities. We explain in Section 3.3 why agghoo should be preferred to these procedures in the general prediction setting.

Because of the aggregation step, agghoo is an ensemble method, and like bagging, it combines resampling with aggregation. However, unlike agghoo, bagging applies to single estimators, and does not adress the problem of estimator selection. Hence, if there is a free hyperparameter, bagging must be combined with some estimator selection method, such as cross-validation. The application of bagging to the hold-out was first suggested by Breiman (1996) as a way to combine pruning and bagging of CART trees. We discuss in detail in Section 3.3 how agghoo relates to bagging and subagging combined with the hold-out. In particular, we explain why previous results on bagging or subagging do not apply to agghoo; new developments are required.

1.2 Contributions

In this article, agghoo's performance is studied both theoretically and experimentally. We consider agghoo from a prediction point of view. Performance is measured by a risk functional. On the theoretical side, the aim is to show that the risk of agghoo's final predictor is as low as the risk of the optimal rule among the given collection. This is known as an oracle inequality. By a convexity argument, agghoo always improves on the hold-out, provided that the risk is convex. Hence, agghoo can safely replace the hold-out in any application where this hypothesis holds true. Another consequence is that oracle inequalities for agghoo can be deduced from oracle inequalities for the hold-out.

This kind of result on the hold-out has already appeared in the literature: for example, Massart (2007, Corollary 8.8) proves a general theorem under an abstract noise assumption; more explicit results have been obtained in specific settings such as least-squares regression (Györfi et al., 2002, Theorem 7.1) or maximum-likelihood density estimation (Massart, 2007, Theorem 8.9). A review on cross-validation—which includes the hold-out—is done by Arlot and Celisse (2010).

Most existing theoretical guarantees on the hold-out have a limitation: they assume that the loss function is uniformly bounded. In regression, the variable Y and the regressors are also usually assumed to be bounded, which excludes some standard least-squares estimators. Even when the boundedness assumption holds true, constants arising from general bounds may be of the wrong order of magnitude, leading to vacuous results. By replacing uniform supremum bounds by local ones, we are able to relax these hypotheses in a general setting (Theorem 17). This enables us to prove an oracle inequality for the hold-out and agghoo in regularized kernel regression with a general Lipschitz loss (Theorem 11). This oracle inequality allows for instance to recover state-of-the-art convergence rates in median regression without knowing the regularity of the regression function (adaptivity), both in the general case and, for small enough regularity, also in the specific setting of Eberts and Steinwart (2013). To illustrate the implications of Theorem 11, we also apply it to ε-regression (Corollary 12). To the best of our knowledge, all these oracle inequalities are new, even for the hold-out. In addition to the RKHS setting studied here, Theorem 17 of this article has also been applied to sparse linear regression (Maillard, 2020a) and to least-squares density estimation (Maillard, 2020b).

A limitation of agghoo is that it does not cover settings where averaging does not make sense, such as classification. In classification with the 0–1 loss, the natural way to aggregate classifiers is to take a majority vote among them. This yields a procedure which we call majhoo. Using existing theory for the hold-out in classification, we prove that majhoo satisfies a general margin-adaptive oracle inequality (Theorem 13) under Tsybakov's margin assumption (Mammen and Tsybakov, 1999).

All our oracle inequalities are valid for any size of the aggregation ensemble. Qualitatively, since bagging and subagging are well-known for their stabilizing effects (Breiman, 1996; Bühlmann and Yu, 2002), we can expect agghoo to behave similarly. In particular, large ensembles should improve much the prediction performance of CV when the hold-out selected predictor is unstable.

For further insights into agghoo and majhoo, we conduct in Section 5 a numerical study on simulated data sets. Its results confirm our intuition: in all settings considered, agghoo

and majhoo actually perform much better than the hold-out, and sometimes better than CV, provided their parameters are well-chosen. When choosing the number of neighbors for k-nearest neighbors, the prediction performance of majhoo is much better than the one of CV, which illustrates the strong interest of using agghoo/majhoo when learning rules are "unstable". In support vector regression, agghoo can even perform as well as the oracle, an achievement that is not matched by CV, ACV, EKCV or bagging applied to K-fold cross-validation. Based upon our experiments, we also give in Section 5 some guidelines for choosing agghoo's parameters: the training set size and the number of data splits.

The remaining of the article is structured as follows. In Section 2, we introduce the general statistical setting. In Section 3, we give a formal definition of agghoo. In Section 4, we state the main theoretical results. In Section 5, we present our numerical experiments and discuss the results. Finally, in Section 6, we draw some qualitative conclusions about agghoo. The proofs are postponed to the Appendix.

2. Setting and Definitions

We consider a general statistical learning setting, following the book by Massart (2007).

2.1 Risk Minimization

The goal is to minimize over a set $\mathbb{S}$ a risk functional $\mathcal{L} : \mathbb{S} \to \mathbb{R} \cup \{+\infty\}$. The set $\mathbb{S}$ may be infinite dimensional for non-parametric problems. Assume that $\mathcal{L}$ attains its minimum over $\mathbb{S}$ at a point s, called a Bayes element. Then the *excess risk* of any $f \in \mathbb{S}$ is the nonnegative quantity

$$\ell(s, f) = \mathcal{L}(f) - \mathcal{L}(s) \ .$$

Suppose that the risk can be written as an expectation over an unknown probability distribution,

$$\mathcal{L}(f) = \mathbb{E}\big[\gamma(f, \xi)\big] \ ,$$

for a *contrast function* $\gamma : \mathbb{S} \times \Xi \to \mathbb{R}$ and a random variable ξ with values in some set Ξ and unknown distribution P, such that

$$\forall f \in \mathbb{S}, \qquad \widetilde{\xi} \in \Xi \mapsto \gamma(f, \widetilde{\xi}) \text{ is } P\text{-measurable} \ .$$

The statistical learning problem is to use data $D_n = (\xi_1, ..., \xi_n)$, where $\xi_1, ..., \xi_n$ are independent and identically distributed with common distribution P, to find an approximate minimizer for $\mathcal{L}$. The quality of this approximation is measured by the excess risk.

2.2 Examples

Supervised learning aims at predicting a quantity of interest $Y \in \mathcal{Y}$ using explanatory variables $X \in \mathcal{X}$. The statistician observes pairs $(X_1, Y_1), \dots (X_n, Y_n)$, so that $\Xi = \mathcal{X} \times \mathcal{Y}$, and seeks a predictor in $\mathbb{S} = \{f : \mathcal{X} \to \mathcal{Y} : f \text{ measurable}\}$. The contrast function is defined by $\gamma(f, (x, y)) = g(f(x), y)$ for some *loss function* $g : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Here, $g(y', y)$ measures the loss incurred by predicting y' instead of the observed value y. Two classical supervised learning problems are classification and regression, which we detail below.

Example 1 (Classification) *In classification, Y belongs to a finite set of labels, that is, $\mathcal{Y} = \{0, \dots, M-1\}$ for some $M \geqslant 2$. We wish to correctly label any new data point X, and the risk is the probability of error*:

$$\forall f \in \mathbb{S}, \qquad \mathcal{L}(f) = \mathbb{P}\big(f(X) \neq Y\big) \ ,$$

which corresponds to the loss function $g(y', y) = \mathbb{I}\{y' \neq y\}$. Classification with convex losses (such as the hinge loss or logistic loss) can also be described using the formalism of Section 2.1.

Example 2 (Regression) *In regression, we wish to predict a continuous variable Y that belongs to $\mathcal{Y} = \mathbb{R}^d$. The error made by predicting y' instead of y is measured by the loss function defined by $g(y', y) = \phi(\|y' - y\|)$ where $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is nondecreasing and convex. Some typical choices are $\phi(x) = x^2$ (least squares), $\phi(x) = |x|$ (median regression) or $\phi(x) = (|x| - \varepsilon)_+$ (Vapnik's ε-insensitive loss, leading to ε-regression). The risk is given by*

$$\mathcal{L}(f) = \mathbb{E}\Big[\phi\big(\|Y - f(X)\|\big)\Big] \ .$$

If ϕ is strictly convex, the minimizer of $\mathcal{L}$ over $\mathbb{S}$ is a unique function, up to modification on a set of probability 0 under the distribution of X.

In some applications, such as robust regression, it is of interest to define s and $\ell(s, f)$ even when $\phi(\|Y\|) \notin L^1$. This is possible for Lipschitz contrasts, by the following remark.

Remark 1 *When ϕ is convex and increasing (as in Example 2), assuming also that ϕ is Lipschitz-continuous, it is always possible to define*

$$s : x \mapsto \operatorname*{argmin}_{u \in \mathbb{R}} \mathbb{E}\big[\phi(\|Y - u\|) - \phi(\|Y\|) \,\big|\, X = x\big] \ .$$

When $s \in L^1(X)$, it is a Bayes element for the loss function $g(y', y) = \phi(\|y' - y\|) - \phi(\|y\|)$. Whenever $\phi(\|Y\|) \in L^1$, this loss yields the same Bayes element and excess risk as in Example 2.2.

This adjustment to the general definition allows to consider Example 2 when $\phi(\|Y - s(X)\|)$ is not integrable, for example when $Y = s(X) + \eta$, where η is independent from X and follows a multivariate Cauchy distribution with location parameter 0.

Some density estimation problems, such as maximum-likelihood or least-squares density estimation, also fit the formalism of Section 2.1, see the book by Massart (2007).

2.3 Learning Rules and Estimator Ensembles

Statistical procedures use data to compute an element of $\mathbb{S}$ which approximately minimizes $\mathcal{L}$. Since agghoo uses subsampling, we require learning rules to accept as input data sets of any size. Therefore, we define a learning rule to be a function which maps any data set to an element of $\mathbb{S}$.

Definition 2 *A* data set D_n *of length* n *is a finite sequence* $(\xi_i)_{1\leqslant i\leqslant n}$ *of independent and identically distributed* Ξ*-valued random variables with common distribution* P*.*

A learning rule $\mathcal{A}$ *is a measurable function*[1]

$$\mathcal{A} : \bigcup_{n=1}^{\infty} \Xi^n \to \mathbb{S} \ .$$

In the risk minimization setting, $\mathcal{A}$ should be chosen so as to minimize $\mathcal{L}(\mathcal{A}(D_n))$.

A generic situation is when a family $(\mathcal{A}_m)_{m\in\mathcal{M}}$ of learning rules is given, so that we have to select one of them (estimator selection), or to combine their outputs (estimator aggregation). For instance, when $\mathcal{X}$ is a metric space, we can consider the family $(\mathcal{A}_k^{\mathrm{NN}})_{k\geqslant 1}$ of nearest-neighbors classifiers —where k is the number of neighbors—, or, for a given kernel on $\mathcal{X}$, the family $(\mathcal{A}_\lambda^{\mathrm{SVM}})_{\lambda\in[0,+\infty)}$ of support vector machine classifiers —where λ is the regularization parameter. Not all rules in such families perform well on a given data set. Bad rules should be avoided when selecting the hyperparameter, or be given small weights if the outputs are combined in a weighted average. This requires a data-adaptive procedure, as the right choice of rule in general depends on the unknown distribution P.

Aggregation and parameter selection methods aim to resolve this problem, as described in the next section.

3. Cross-Validation and Aggregated Hold-Out (Agghoo)

This section recalls the definition of cross-validation for estimator selection, and introduces a new procedure called aggregated hold-out (agghoo). For more details and references on cross-validation, we refer the reader to the survey by Arlot and Celisse (2010).

3.1 Background: Cross-Validation

Cross-validation uses subsampling and the empirical risk. We first introduce some notation.

Definition 3 (Empirical risk) *For any data set* $D_n = (\xi_i)_{1\leqslant i\leqslant n}$ *and any* $f \in \mathbb{S}$*, the empirical risk of* f *over* D_n *is defined by*

$$P_n\gamma(f,\cdot) = \frac{1}{n}\sum_{i=1}^{n}\gamma(f,\xi_i) \ .$$

For any nonempty subset $T \subset \{1,\dots,n\}$*, let also*

$$D_n^T = (\xi_i)_{i\in T}$$

be the subsample of D_n *indexed by* T*, and define the associated empirical risk by*

$$\forall f \in \mathbb{S}, \qquad P_n^T\gamma(f,\cdot) = \frac{1}{|T|}\sum_{i\in T}\gamma(f,\xi_i) \ .$$

1. For any n,

$$\begin{cases} \Xi^n \times \Xi & \to \mathbb{R} \\ (\xi_{1:n},\xi) & \mapsto \gamma(\mathcal{A}(\xi_{1:n}),\xi) \end{cases}$$

is assumed to be measurable with respect to the product σ-algebra on Ξ^{n+1}.

The most classical estimator selection procedure is to *hold out* some data to calculate the empirical risk of each estimator, and then to select the estimator with the lowest empirical risk. This ensures that the data used to evaluate the risk are independent from the training data used to compute the learning rules.

Definition 4 (Hold-out) *For any data set D_n and any subset $T \subset \{1, \dots, n\}$, the associated hold-out risk estimator of a learning rule $\mathcal{A}$ is defined by*

$$\mathrm{HO}_T(\mathcal{A}, D_n) = P_n^{T^c} \gamma\left(\mathcal{A}(D_n^T), \cdot\right) \ .$$

Given a collection of learning rules $(\mathcal{A}_m)_{m \in \mathcal{M}}$, the hold-out procedure selects

$$\widehat{m}_T^{\mathrm{ho}}(D_n) \in \operatorname*{argmin}_{m \in \mathcal{M}} \mathrm{HO}_T(\mathcal{A}_m, D_n) \ ,$$

measurably with respect to D_n. The overall learning rule is then given by

$$\widehat{f}_T^{\mathrm{ho}}\left((\mathcal{A}_m)_{m \in \mathcal{M}}, D_n\right) = \mathcal{A}_{\widehat{m}_T^{\mathrm{ho}}(D_n)}(D_n^T) \ .$$

Hold-out depends on the arbitrary choice of a training set T, and is known to be quite unstable, despite its good theoretical properties (Massart, 2007, Section 8.5.1). Therefore, practicioners often prefer to use cross-validation instead, which considers several training sets.

Definition 5 (Cross-validation) *Let D_n denote a data set. Let $\mathcal{T}$ denote a collection of nonempty subsets of $\{1, \dots, n\}$. The associated cross-validation risk estimator of a learning rule $\mathcal{A}$ is defined by*

$$\mathrm{CV}_{\mathcal{T}}(\mathcal{A}, D_n) = \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} \mathrm{HO}_T(\mathcal{A}, D_n) \ .$$

The cross-validation procedure then selects

$$\widehat{m}_{\mathcal{T}}^{\mathrm{cv}}(D_n) \in \operatorname*{argmin}_{m \in \mathcal{M}} \mathrm{CV}_{\mathcal{T}}(\mathcal{A}_m, D_n) \ .$$

The final predictor obtained through this procedure is

$$\widehat{f}_{\mathcal{T}}^{\mathrm{cv}}\left((\mathcal{A}_m)_{m \in \mathcal{M}}, D_n\right) = \mathcal{A}_{\widehat{m}_{\mathcal{T}}^{\mathrm{cv}}(D_n)}(D_n) \ .$$

Depending on how $\mathcal{T}$ is chosen, this can lead to leave-one-out, leave-p-out, V-fold cross-validation or Monte-Carlo cross-validation, among others (Arlot and Celisse, 2010). In the following, we omit some of the arguments $\mathcal{A}, D_n$ which appear in Definitions 4 and 5, when they are clear from context. For example, we often write $\mathrm{HO}_T(\mathcal{A}), \widehat{m}_T^{\mathrm{ho}}, \widehat{f}_T^{\mathrm{ho}}$ instead of $\mathrm{HO}_T(\mathcal{A}, D_n), \widehat{m}_T^{\mathrm{ho}}(D_n), \widehat{f}_T^{\mathrm{ho}}\left((\mathcal{A}_m)_{m \in \mathcal{M}}, D_n\right)$, respectively.

3.2 Aggregated Hold-Out (Agghoo) Estimators

In this paper, we study another way to improve on the stability of hold-out selection, by *aggregating* the predictors $\widehat{f}_T^{\mathrm{ho}}$ obtained by the hold-out procedure applied repeatedly with different training sets $T \in \mathcal{T}$. When $\mathbb{S}$ is convex (for example, regression), *aggregated hold-out* (agghoo) consists in averaging them.

Definition 6 (Agghoo) *Assume that $\mathbb{S}$ is a convex set. Let $(\mathcal{A}_m)_{m\in\mathcal{M}}$ denote a collection of learning rules, D_n a data set, and $\mathcal{T}$ a collection of subsets of $\{1,\dots,n\}$. Using the notation of Definition 4, the associated agghoo estimator is defined by*

$$\widehat{f}^{\mathrm{ag}}_{\mathcal{T}}\big((\mathcal{A}_m)_{m\in\mathcal{M}}, D_n\big) = \frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}} \widehat{f}^{\mathrm{ho}}_{T}\big((\mathcal{A}_m)_{m\in\mathcal{M}}, D_n\big) \ .$$

In the classification framework, as seen in Example 1, $\mathbb{S} = \{f : \mathcal{X} \to \{0,\dots,M-1\}\}$ which is not convex. However, there is still a natural way to aggregate several classifiers, by taking a majority vote.

Definition 7 (Majhoo) *Let $\mathcal{Y} = \{0,\dots,M-1\}$ be the set of labels. Given a collection of learning rules $(\mathcal{A}_m)_{m\in\mathcal{M}}$, a data set D_n and a collection $\mathcal{T}$ of subsets of $\{1,\dots,n\}$, the majority hold-out (majhoo) classifier is any measurable $\widehat{f}^{\mathrm{mv}}_{\mathcal{T}}\big((\mathcal{A}_m)_{m\in\mathcal{M}}, D_n\big) : \mathcal{X} \to \mathcal{Y}$ such that, using the notation $\widehat{f}^{\mathrm{ho}}_{T}$ introduced in Definition 4, for all $x \in \mathcal{X}$,*

$$\widehat{f}^{\mathrm{mv}}_{\mathcal{T}}\big((\mathcal{A}_m)_{m\in\mathcal{M}}, D_n\big)(x) \in \operatorname*{argmax}_{j\in\mathcal{Y}} \Big|\Big\{T\in\mathcal{T} \,:\, \widehat{f}^{\mathrm{ho}}_{T}\big((\mathcal{A}_m)_{m\in\mathcal{M}}, D_n\big)(x) = j\Big\}\Big| \ .$$

In most situations, it is clear how hold-out rules should be aggregated and there is no ambiguity in discussing hold-out aggregation. However, there is an important exception where both agghoo and majhoo can be used.

Remark 8 (Two options for binary classification) *In binary classification (Example 1 with $M = 2$), it is classical to consider classifiers of the form $x \mapsto \mathbb{I}_{f(x)\geqslant 0}$ where the function $f \in \mathbb{S}_{conv} = \{f : \mathcal{X} \to \mathbb{R}\}$ aims at minimizing a surrogate convex risk associated with the loss $g_{conv} : (y', y) \mapsto \phi[(2y'-1)(2y-1)]$ with $\phi : \mathbb{R} \to \mathbb{R}$ convex (Boucheron et al., 2005). Then, given a family of $\mathbb{S}_{conv}$-valued learning rules $\big(\mathcal{A}_m\big)_{m\in\mathcal{M}}$, one can either apply agghoo to the surrogate problem and get*

$$\mathbb{I}_{\widehat{f}^{\mathrm{ag}}_{\mathcal{T}}\left((\mathcal{A}_m)_{m\in\mathcal{M}}, D_n\right)\geqslant 0} \ ,$$

or apply majhoo to the binary classification problem and get

$$\widehat{f}^{\mathrm{mv}}_{\mathcal{T}}\left(\big(\mathbb{I}_{\mathcal{A}_m(\cdot)\geqslant 0}\big)_{m\in\mathcal{M}}, D_n\right) \ .$$

In the rest of Section 3, we focus on agghoo, though much of the following discussion applies also to majhoo.

Compared to cross-validation rules (Definition 5), agghoo reverses the order between aggregation (majority vote or averaging) and minimization of the risk estimator: instead of averaging hold-out risk estimators before selecting the hyperparameter, the selection step is made first to produce hold-out predictors $\big(\widehat{f}^{\mathrm{ho}}_{T}\big)_{T\in\mathcal{T}}$ (given by Definition 4) and then an average is taken.

3.3 Related Procedures

To the best of our knowledge, agghoo has not been studied theoretically before, though it is used in applications (Hoyos-Idrobo et al., 2015; Varoquaux et al., 2017), under the name "CV + averaging" in Varoquaux et al. (2017). According to Varoquaux et al. (2017), agghoo is commonly used by the machine learning community thanks to the SCIKIT-LEARN library (Pedregosa et al., 2011).

A related procedure is "K-fold averaging cross-validation" (ACV), proposed by Jung and Hu (2015). In linear regression, ACV corresponds to averaging the $\mathcal{A}_{\widehat{m}_T^{\mathrm{ho}}}(D_n)$, which are "retrained" on the whole data set, while agghoo averages the $\mathcal{A}_{\widehat{m}_T^{\mathrm{ho}}}(D_n^T)$. An advantage of averaging the rules $\mathcal{A}_{\widehat{m}_T^{\mathrm{ho}}}(D_n^T)$ is that they have been selected for their good performance on the validation set T^c, unlike the $\mathcal{A}_{\widehat{m}_T^{\mathrm{ho}}}(D_n)$ whose performance has not been assessed on independent data. Furthermore, similarly to bagging, using several distinct training sets may result in improvements for unstable methods through a reduction in variance. Note finally that the theoretical results of Jung and Hu (2015) on ACV are limited to a specific setting, and much weaker than an oracle inequality.

A second family of related procedures is averaging the chosen *parameters* $\left(\widehat{m}_T^{\mathrm{ho}}\right)_{T\in\mathcal{T}}$, contrary to agghoo which averages the chosen *prediction rules*. This leads to different procedures for learning rules that are not linear functions of their parameters. This is the approach taken by Jung and Hu (2015) for selecting a regularization parameter, still under the name of ACV. The idea has also been put forward under the name "bagged cross-validation" (that we call in this article Hall's BCV, for avoiding confusion with other ways of combining bagging and cross-validation) by Hall and Robinson (2009) —with numerical and theoretical results in the case of bandwidth choice in kernel density estimation—, and under the name "efficient K-fold cross-validation" (EKCV; Jung, 2016) for the choice of a regularization parameter in high-dimensional regression —with numerical results only. Unlike agghoo, which only depends on the set $\{\mathcal{A}_m : m \in \mathcal{M}\}$ of learning rules, ACV, EKCV and Hall's BCV depend on the parametrization $m \mapsto \mathcal{A}_m$. Sometimes, the most natural parametrization does not allow the use of such procedures: for example, model dimensions are integers, and averaging them does not make sense. In contrast, in regression, it is always possible to average the real-valued functions $\mathcal{A}_m(D_{n_t}) \in \mathbb{S}$.

Even when all procedures are applicable, averaging rules is generally safer than averaging hyperparameters. Often in regression, the risk $\mathcal{L}$ is known to be convex over $\mathbb{S}$, so given $f_1, \ldots, f_V \in \mathbb{S}$,

$$\mathcal{L}\left(\frac{1}{V}\sum_{i=1} f_i\right) \leqslant \frac{1}{V}\sum_{i=1}^{V} \mathcal{L}(f_i) \ .$$

Hence, averaging regressors (agghoo) always improves performance compared to selecting a single f_i at random (hold-out). On the other hand, if $(f_\theta)_{\theta\in\Theta}$ is a family of elements of $\mathbb{S}$ parametrized by a convex set Θ, there is no guarantee in general that the function $\theta \mapsto \mathcal{L}(f_\theta)$ is convex over Θ. So, for some $\theta_1, \ldots, \theta_V \in \Theta$, it may happen that

$$\mathcal{L}\left(f_{\frac{1}{V}\sum_{i=1}^{V}\theta_i}\right) \geqslant \frac{1}{V}\sum_{i=1}^{V} \mathcal{L}(f_{\theta_i}) \ .$$

In such a case, it is better to choose one parameter at random (hold-out) than to average them (ACV, EKCV or Hall's BCV).

A third family of related procedures is bagging or subagging applied to some CV predictor $D_n \mapsto \widehat{f}^{\mathrm{cv}}_{\mathcal{T}}((\mathcal{A}_m)_{m\in\mathcal{M}}, D_n)$ given by Definition 5. The bagging case (that we call "bagged CV") has been studied numerically by Petersen et al. (2007), but clearly differs from agghoo since it relies on bootstrap resamples, in which the original data can appear several times. As a consequence, some data can be shared between training and test samples, which breaks the independence between them. This is a major issue for theory, and it can degrade the performance as shown by our numerical experiments (Tables 1–2 in Section 5.1).

The problem disappears if sampling with replacement (bagging) is replaced with sampling without replacement (subagging). The resulting procedure, that can be called "subagged CV" in general, and "subagged hold-out" when the kind of CV considered is the hold-out, is not explicitly studied in the literature, to the best of our knowledge. Subagged hold-out is closer to agghoo but there is still a slight difference. In subagged hold-out, the sample is divided into three parts: the training part of the bagging subsample, the validation part of the bagging subsample, and the data not in the bagging subsample. Thus, part of the data is discarded in each iteration of subagging. With agghoo, the sample is only divided into two parts: training set and validation set. Thus, all the data is used, either to train the learning rules or to estimate their risk. As a result, agghoo is potentially more efficient in its use of the data. Another consequence is that theoretical results on bagging or subagging cannot be used directly for studying agghoo.

Note also that Petersen et al. (2007) recommend a different approach, where rather than bagging the whole CV procedure, bagging is instead applied *within* each CV fold, which leads to the (random) selection of a single (bagged) estimator. In contrast, agghoo selects different estimators for each fold, potentially reducing the variance as usual with ensemble methods (Catoni, 2001; Lecué, 2007; Genuer, 2012).

3.4 Computational Complexity

In general, for a given value of $V = |\mathcal{T}|$, both agghoo ($\widehat{f}^{\mathrm{ag}}_{\mathcal{T}}$) and CV ($\widehat{f}^{\mathrm{cv}}_{\mathcal{T}}$) must compute V hold-out risk estimators over all values of $m \in \mathcal{M}$. Assume for simplicity that all training data sets D_n^T, $T \in \mathcal{T}$, have the same size $|T| = n_t$, and denote by $n_v = n - n_t$ the size of the validation data set. Let $C_{ho}(\mathcal{M}, n_t, n_v)$ be the average computational complexity of the hold-out. Then the overall complexity of risk estimation is of order $V \times C_{ho}(\mathcal{M}, n_t, n_v)$ for both agghoo and CV. Next, CV must average V risk vectors of length $|\mathcal{M}|$ and find a single minimum, while agghoo computes V minima over $m \in \mathcal{M}$; these operations have similar complexity, of order $V \times |\mathcal{M}|$. Thus, computing the ensemble aggregated by agghoo takes about as much time as selecting a learning rule using cross-validation.

A potential difference occurs when evaluating agghoo and CV on new data. If there is no fast way to perform aggregation at training time, it is always possible to evaluate each predictor in the ensemble on the new data, and to average the results; then, agghoo is slower than CV by a factor of order V at test time.

4. Theoretical Results

The purpose of agghoo is to construct an estimator whose risk is as small as possible, compared to the (unknown) best rule in the class $(\mathcal{A}_m)_{m\in\mathcal{M}}$. This is guaranteed theoretically by proving "oracle inequalities" of the form

$$\mathbb{E}\big[\ell(s,\widehat{f}^{\mathrm{ag}}_{\mathcal{T}})\big] \leqslant C\mathbb{E}\Big[\inf_{m\in\mathcal{M}} \ell\big(s,\mathcal{A}_m(D_n)\big)\Big] + \varepsilon_n \ , \tag{1}$$

with ε_n negligible compared to the oracle excess risk $\mathbb{E}[\inf_{m\in\mathcal{M}} \ell(s,\mathcal{A}_m(D_n))]$ and C close to 1. Equation (1) then implies that agghoo performs as well as the best choice of $m \in \mathcal{M}$, up to the constant C. In the following, we actually prove slightly weaker inequalities that are more natural in our setting.

By definition, agghoo is an average of predictors chosen by hold-out over the collection $(\mathcal{A}_m)_{m\in\mathcal{M}}$. Therefore, when the risk is convex, an oracle inequality (1) can be deduced from an oracle inequality for the hold-out, provided that there exists an integer $n_t \in \{1,\dots,n-1\}$ such that

$$\mathcal{T} \text{ is independent from } D_n \qquad \text{and} \qquad \forall T\in\mathcal{T},\quad |T| = n_t \ . \tag{2}$$

We make this assumption in the rest of the article, and then define $n_v = n - n_t$ the size of the validation data set.

Most cross-validation methods satisfy hypothesis (2), including leave-p-out, V-fold cross-validation (with $n_t = n(V-1)/V$) and Monte-Carlo cross-validation (Arlot and Celisse, 2010).

In the remainder of this section, we introduce the RKHS setting of interest, and prove an oracle inequality for agghoo without changing the standard estimators or requiring Y to be bounded.

4.1 Agghoo in Regularized Kernel Regression

Kernel methods such as support vector machines (SVM), kernel least squares or ε-regression use a kernel function to map the data X_i into an infinite-dimensional function space, more specifically a reproducing kernel Hilbert space (RKHS) (Scholkopf and Smola, 2001; Steinwart and Christmann, 2008). We consider in this section regularized empirical risk minimization using a *training* loss function c, with a penalty proportional to the square norm of the RKHS, to solve the supervised learning problem (defined in Section 2.2) with loss function g for defining the risk. Hence, the contrast γ can be written $\gamma(f,(x,y)) = g(f(x),y) := (g\circ f)(x,y)$. We assume that g and c are convex in their first argument.

Definition 9 (Regularized kernel estimator) *Let $c : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ be convex in its first argument, and let $K : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ be a positive-definite kernel function. Given $\lambda > 0$ and training data $(X_i,Y_i)_{1\leqslant i\leqslant n_t}$, define the regularized kernel estimator as*

$$\mathcal{A}_\lambda(D_{n_t}) = \operatorname*{argmin}_{f\in\mathcal{H}} \Big\{ P_{n_t}(c\circ f) + \lambda \left\|f\right\|^2_{\mathcal{H}} \Big\} \ ,$$

where $\mathcal{H}$ is the reproducing kernel Hilbert space induced by K. By the representer theorem, $\mathcal{A}_\lambda$ can be computed explicitly:

$$\mathcal{A}_\lambda(D_{n_t})(x) = \sum_{j=1}^{n_t} \widehat{\theta}_{\lambda,j} K(X_j, x) \qquad \textit{where } \widehat{\theta}_{\lambda,j} \textit{ is the } j\textit{-th component of the vector}$$

$$\widehat{\theta}_\lambda = \operatorname*{argmin}_{\theta\in\mathbb{R}^{n_t}} \left\{ \frac{1}{n_t} \sum_{i=1}^{n_t} c\left(\sum_{j=1}^{n_t} \theta_j K(X_j, X_i), Y_i \right) + \lambda \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} \theta_i \theta_j K(X_i, X_j) \right\} \ . \tag{3}$$

The loss function c is used to measure the accuracy of the fit on the training data: for example, taking $c : (u, y) \mapsto (1 - uy)_+$ (the hinge loss) in Definition 9 corresponds to SVM. The loss function g used for risk evaluation may or may not be equal to c. For example, in classification, the 0–1 loss often cannot be used for training for computational reasons, hence a surrogate convex loss, such as the hinge loss, is used instead (see Remark 8), but there is no reason to use the hinge loss for risk estimation and hyperparameter selection.

In Definition 9, the hyperparameter of interest is λ (we assume that K is fixed). We show below some guarantees on agghoo's performance when it is applied to a finite subfamily $(\mathcal{A}_\lambda)_{\lambda\in\Lambda}$ of the one defined by Definition 9. We first state some useful assumptions.

Hypothesis $Comp_C(g, c)$: The functions $\mathcal{L}_c : f \mapsto P(c \circ f)$ and $\mathcal{L}_g$ have a common minimum $s \in \operatorname{argmin}_{f\in\mathbb{S}} \mathcal{L}_c(f) \cap \operatorname{argmin}_{f\in\mathbb{S}} \mathcal{L}_g(f)$ and for any $f \in \mathbb{S}$, $\mathcal{L}_c(f) - \mathcal{L}_c(s) \leqslant C\,[\mathcal{L}_g(f) - \mathcal{L}_g(s)]$.

Note that $Comp_1(g, c)$ is always satisfied when $g = c$. When $g \neq c$, some hypothesis relating c and g is necessary anyway for Definition 9 to be of interest, if only to ensure consistency (asymptotic minimization of the risk) for some sequence of hyperparameters $(\lambda_n)_{n\in\mathbb{N}}$.

In addition, some information about the evaluation loss g helps to obtain an oracle inequality (1) with a smaller remainder term ε_n.

Hypothesis $SC_{\rho,\nu}$: Let $\ell_X(u) = \mathbb{E}[g(u, Y)|X] - \inf_{v\in\mathbb{R}} \mathbb{E}[g(v, Y)|X]$. The triple (g, X, Y) satisfies $SC_{\rho,\nu}$ if and only if, for any $u, v \in \mathbb{R}$,

$$\mathbb{E}\Big[\big(g(u, Y) - g(v, Y)\big)^2 \,\Big|\, X \Big] \leqslant \Big[\rho \vee \big(\nu|u - v|\big) \Big] \big[\ell_X(u) + \ell_X(v) \big] \ . \tag{4}$$

For example, in the case of median regression, that is, $g(u, y) = |u - y|$, hypothesis $SC_{\rho,\nu}$ holds whenever there is a uniform lower bound on the concentration of Y around $s(X)$, as shown by the following proposition.

Proposition 10 *Let $g(u, y) = |u - y|$ for all $u, y \in \mathbb{R}$. For any $x \in \mathcal{X}$, let F_x be the conditional cumulative distribution function of Y knowing $X = x$. Assume that, for any $x \in \mathcal{X}$, F_x is continuous with a unique median $s(x)$ and that there exists $a(x) > 0, b(x) > 0$ such that*

$$\forall u \in \mathbb{R}, \qquad \Big| F_x(u) - F_x\big(s(x)\big) \Big| \geqslant a(x) \Big[\big| u - s(x) \big| \wedge b(x) \Big] \ . \tag{5}$$

For instance, this holds true if $\frac{dF_x}{du} \geqslant a(x)\mathbb{I}_{|u-s(x)|\leqslant b(x)}$ *for every* $x \in \mathcal{X}$. *Let*

$$a_m = \inf_{x\in\mathcal{X}}\big\{a(x)\big\} \qquad \text{and} \qquad \mu_m = \inf_{x\in\mathcal{X}}\big\{a(x)b(x)\big\} \ .$$

If $a_m > 0$ *and* $\mu_m > 0$, *then* (g, X, Y) *satisfies* $SC_{\frac{2}{a_m},\frac{2}{\mu_m}}$.

Proposition 10 is proved in Appendix C.1. We can now state our first main result.

Theorem 11 *Let* $\Lambda \subset \mathbb{R}_+^*$ *be a finite grid. Using the notation of Definition 6 and assuming that* (2) *holds true, let* $\widehat{f}_{\mathcal{T}}^{\,\mathrm{ag}}$ *be the output of agghoo, applied to the collection* $(\mathcal{A}_\lambda)_{\lambda\in\Lambda}$ *given by Definition 9. Assume that* $\lambda_m = \min\Lambda > 0$ *and* $\kappa = \sup_{x\in\mathcal{X}} K(x,x) < +\infty$. *Assume that* $Comp_C(g, c)$ *holds for a constant* $C > 0$ *and that* (g, X, Y) *satisfies* $SC_{\rho,\nu}$ *with constants* $\rho \geqslant 0, \nu \geqslant 0$. *Assume that* c *and* g *are convex and Lipschitz in their first argument, with Lipschitz constant less than* L. *Assume also that* $n_v \geqslant 100$ *and* $3 \leqslant |\Lambda| \leqslant \mathrm{e}^{\sqrt{n_v}}$. *Then, for any* $\theta \in (0, 1]$,

$$\begin{aligned}(1-\theta)\mathbb{E}\Big[\ell\big(s,\widehat{f}_{\mathcal{T}}^{\,\mathrm{ag}}\big)\Big] \leqslant (1+\theta)\mathbb{E}\Big[\min_{\lambda\in\Lambda}\ell\big(s,\mathcal{A}_\lambda(D_{n_t})\big)\Big] \\ + \max\left\{18\rho\frac{\log\big(n_v|\Lambda|\big)}{\theta n_v}, b_1\frac{\log^2\big(n_v|\Lambda|\big)}{\theta^3\lambda_m n_v^2}, b_2\frac{\log^{\frac{3}{2}}\big(n_v|\Lambda|\big)}{\theta\lambda_m n_v\sqrt{n_t}}\right\} \ ,\end{aligned} \tag{6}$$

where b_1, b_2 *do not depend on* $n_v, n_t, \lambda_m, \rho$ *or* θ *but only on* κ, L, ν *and* C.

Theorem 11 is proved in Appendix B as a consequence of a result valid in the general framework of Section 2.1 (Theorem 17). It shows that $\widehat{f}_{\mathcal{T}}^{\,\mathrm{ag}}$ satisfies an oracle inequality of the form (1), with $\mathcal{A}_\lambda(D_{n_t})$ instead of $\mathcal{A}_\lambda(D_n)$ on the right-hand side of the inequality. The fact that D_{n_t} appears in the bound instead of D_n is a limitation of our result, but it is natural since predictors aggregated by agghoo are only trained on part of the data. In most cases, it can be expected that $\ell(s, \mathcal{A}_\lambda(D_{n_t}))$ is close to $\ell(s, \mathcal{A}_\lambda(D_n))$ whenever $\frac{n_t}{n}$ is close to 1.

The assumption that K is bounded is mild. For instance, popular kernels such as Gaussian kernels, $(x, x') \mapsto \exp[-\|x - x'\|^2/(2h^2)]$ for some $h > 0$, or Laplace kernels, $(x, x') \mapsto \exp(-\|x - x'\|/h)$ for some $h > 0$, are bounded by $\kappa = 1$.

Taking $|\mathcal{T}| = 1$ in Theorem 11 yields a new oracle inequality for the hold-out. Oracle inequalities for the hold-out have already been proved in a variety of settings (see Arlot and Celisse, 2010, for a review), and used to obtain adaptive rates in regularized kernel regression (Steinwart and Christmann, 2008). However, this work has mostly been accomplished under the assumption that the contrast $\gamma\left(\mathcal{A}_\lambda(D_n), (X, Y)\right)$ is bounded uniformly (in n, D_n and $\lambda \in \Lambda$) by a constant. If this constant increases with n, bounds obtained in this manner may worsen considerably. As many "natural" regression procedures—including regularized kernel regression (Definition 9)—fail to satisfy such bounds, some theoreticians introduce "truncated" versions of standard procedures (Steinwart and Christmann, 2008), but truncation has no basis in practice. Theorem 11 avoids these complications.

In order to be satisfactory, Theorem 11 should prove that agghoo performs asymptotically as well as the best choice of $\lambda \in \Lambda$, at least for reasonable choices of Λ. This is the case

whenever the maximum in Equation (6) is negligible with respect to the oracle excess risk $\mathbb{E}[\min_{\lambda\in\Lambda} \ell(s, \mathcal{A}_\lambda(D_{n_t}))]$ as $n \to +\infty$. This depends on the range $[\lambda_m, +\infty)$ in which the hold-out is allowed to search for the optimal λ. On the one hand, it is desirable that this interval be wide enough to contain the true optimal value. On the other hand, if $\lambda_m = 0$, then inequality (6) becomes vacuous. We now provide precise examples where Theorem 11 applies with a remainder term in Equation (6) that is negligible relative to the oracle excess risk.

Take the example of median regression, in which $c(u, y) = g(u, y) = |u - y|$. Then $Comp_1(g, c)$ holds trivially. Make also the same assumptions as in Proposition 10, which ensures that $SC_{\rho,\nu}$ holds for some finite values of ρ and ν. Theorem 11 therefore applies as long as the kernel K is bounded and $\lambda_m > 0$. Choose $n_v = n_t = \frac{n}{2}$ and Λ of cardinality at most polynomial in n (which is sufficient in theory and in practice). Then Steinwart and Christmann (2008, Theorem 9.6) prove the consistency of $\mathcal{A}_{\lambda_n}(D_n)$ as $n \to +\infty$, provided that $\lambda_n^2 n \to +\infty$. This suggests choosing $\lambda_m = 1/\sqrt{n_t}$, in which case the remainder term of Equation (6) is of order $(\log n)^{3/2}/n$, which is negligible relative to nonparametric convergence rates in median regression.

In order to have a more precise idea of the order of magnitude of the oracle excess risk, let us consider median regression with a Gaussian kernel. Under some assumptions, one of which coincides with Proposition 10, Eberts and Steinwart (2013, Corollary 4.12) show that taking $\lambda_n = \frac{c_1}{n}$ leads to rates of order $n^{-\frac{2\alpha}{2\alpha+d}}$, where $d \in \mathbb{N}$ is the dimension of $\mathcal{X}$ and $\alpha > 0$ is the smoothness of s. Therefore, taking $\lambda_m = 1/n_t$ in Theorem 11, the remainder term of Equation (6) is at most of order $(\log n)^{3/2}/\sqrt{n}$, hence negligible relative to the above risk rates as soon as $2\alpha < d$.

Theorem 11 can handle situations where g is different from the training loss c, provided that $Comp_C(g, c)$ holds true. Such situations arise for instance in the case of support vector regression (Scholkopf and Smola, 2001, Chapter 9), which uses for training Vapnik's ε-insensitive loss $c_\varepsilon^{eps}(u, y) = (|u-y|-\varepsilon)_+$. This loss depends on a parameter ε, the choice of which is usually motivated by a tradeoff between sparsity and prediction accuracy (Scholkopf and Smola, 2001). Therefore, some other loss is typically used to measure predictive performance, independently of ε. We state one possible application of Theorem 11 to this case, as a corollary.

Corollary 12 (ε-regression) *Let $c = c_\varepsilon^{eps} : (u, y) \mapsto (|y-u|-\varepsilon)_+$ be Vapnik's ε-insensitive loss and assume that the evaluation loss is $g = c_0^{eps} : (u, y) \mapsto |u - y|$. Assume that for every x the conditional distribution of Y given $X = x$ has a unimodal density with respect to the Lebesgue measure, symmetric around its mode. Introduce the robust noise parameter*

$$\sigma = \sup_{x\in\mathcal{X}} \left\{ \inf \left\{ y \in \mathbb{R} \,:\, \mathbb{P}(Y \leqslant y \,|\, X = x) \geqslant \frac{3}{4} \right\} - \sup \left\{ y \in \mathbb{R} \,:\, \mathbb{P}(Y \leqslant y \,|\, X = x) \leqslant \frac{1}{4} \right\} \right\} .$$

Then, applying agghoo to a finite subfamily $(\mathcal{A}_\lambda)_{\lambda\in\Lambda}$ of the rules given by Definition 9 with $c = c_\varepsilon^{eps}$ and a kernel K such that $\|K\|_\infty \leqslant 1$ yields the following oracle inequality. Assuming

$n_v \geqslant 100$, $3 \leqslant |\Lambda| \leqslant \mathrm{e}^{\sqrt{n_v}}$, and that (2) holds true, we have that for any $\theta \in (0,1]$,

$$(1-\theta)\mathbb{E}\Big[\ell\big(s,\widehat{f}^{\mathrm{ag}}_{\mathcal{T}}\big)\Big] \leqslant (1+\theta)\mathbb{E}\Big[\min_{\lambda\in\Lambda}\ell\big(s,\mathcal{A}_\lambda(D_{n_t})\big)\Big] + \max\left\{72\sigma\frac{\log\big(n_v|\Lambda|\big)}{\theta n_v}\,,\, b_1\frac{\log^2\big(n_v|\Lambda|\big)}{\theta^3\lambda_m n_v^2}\,,\, b_2\frac{\log^{\frac{3}{2}}\big(n_v|\Lambda|\big)}{\theta\lambda_m n_v\sqrt{n_t}}\right\}\ ,$$

where b_1 and b_2 represent numerical constants.

Corollary 12 is proved in Appendix C.2.

When $\varepsilon = 0$, ε-regression becomes median regression, which is discussed above. The oracle inequality of Corollary 12 is then the same as that given by Theorem 11 and Proposition 10. Assumptions of unimodality and symmetry allow to give more explicit values of a_m and μ_m in terms of σ. When $\varepsilon > 0$, the unimodality and symmetry assumptions are used to prove hypothesis $Comp_C(g,c)$.

4.2 Classification

Loss functions are not all convex. When convexity fails, the aggregation procedure should be revised.

In classification, majhoo is a possible solution (see Definition 7). By Proposition 35 in Appendix D, majority voting satisfies a kind of "convexity inequality" with respect to the 0–1 loss; as a result, oracle inequalities for the hold-out imply oracle inequalities for majhoo.

Hold-out for binary classification with 0–1 loss has been studied by Massart (2007). In that work, Massart makes an assumption which is closely related to margin hypotheses, such as Tsybakov's noise condition (Mammen and Tsybakov, 1999) which we consider here. This approach allows to derive the following theorem.

Theorem 13 *Consider the classification setting described in Example 1 with $M = 2$ classes (binary classification). Let $(\mathcal{A}_m)_{m\in\mathcal{M}}$ be a collection of learning rules and $\mathcal{T}$ a collection of training sets satisfying assumption (2).*

Assume that there exists $\beta \geqslant 0$ and $r \geqslant 1$ such that for $\xi = (X,Y)$ with distribution P,

$$\forall h > 0, \qquad \mathbb{P}\Big(\big|2\eta(X)-1\big| \leqslant h\Big) \leqslant r h^{\beta} \tag{MA}$$

where $\eta(X) := \mathbb{P}(Y = 1 \,|\, X)$. Then, we have

$$\mathbb{E}\Big[\ell\big(s,\widehat{f}^{\mathrm{mv}}_{\mathcal{T}}\big)\Big] \leqslant 3\mathbb{E}\left[\inf_{m\in\mathcal{M}}\ell\big(s,\mathcal{A}_m(D_{n_t})\big)\right] + \frac{29 r^{\frac{1}{\beta+2}}\log\big(\mathrm{e}|\mathcal{M}|\big)}{n_v^{\frac{\beta+1}{\beta+2}}}\ .$$

Theorem 13 is proved in Appendix D. It shows that $\widehat{f}^{\mathrm{mv}}_{\mathcal{T}}$, like $\widehat{f}^{\mathrm{ag}}_{\mathcal{T}}$, satisfies an oracle inequality of the form (1) with $\mathcal{A}_m(D_{n_t})$ instead of $\mathcal{A}_m(D_n)$. Tsybakov's margin assumption (MA) only depends on the distribution of (X,Y) and not on the collection of learning rules. It is a standard hypothesis in classification, under which "fast" learning rates—faster than $n^{-1/2}$—are attainable (Tsybakov, 2004). In contrast with the results of Section 4.1, that are

valid for various losses but only for a specific type of learning rule, Theorem 13 holds true for *any* family of classification rules.

The constant 3 in front of the oracle excess risk can be replaced by any constant larger than 2, at the price of increasing the constant in the remainder term, as can be seen from the proof (in Appendix D). However, our approach cannot yield a constant lower than 2, because we use Proposition 35 instead of a convexity argument, since the 0–1 loss is not convex.

5. Numerical Experiments

This section investigates how agghoo and majhoo's performance vary with their parameters $V = |\mathcal{T}| =$ and $\tau = \frac{n_t}{n}$, and how it compares to the performance of CV and related methods at a similar computational cost—that is, for the same values of V and τ. Two settings are considered, corresponding to Corollary 12 (ε-regression) and Theorem 13 (classification).

5.1 ε-Regression

Consider the collection $(\mathcal{A}_\lambda)_{\lambda\in\Lambda}$ of regularized kernel estimators (see Definition 9) with loss function $c_\varepsilon^{eps}(u,y) = (|u-y|-\varepsilon)_+$ and Gaussian kernel $K(x,x') = \exp[-(x-x')^2/(2h^2)]$ over $\mathcal{X} = \mathbb{R}$.

5.1.1 Experimental Procedure

Agghoo and CV training sets $T \in \mathcal{T}$ are chosen independently and uniformly among the subsets of $\{1,\dots,n\}$ with cardinality $\lfloor \tau n \rfloor$, for different values of τ and $V = |\mathcal{T}|$; hence, CV corresponds to what is usually called "Monte-Carlo CV" (Arlot and Celisse, 2010). Each algorithm is run on 1000 independent samples of size $n = 500$, and independent test samples of size 1000 are used for estimating the excess risks $\ell(s, \widehat{f}_{\mathcal{T}}^{\text{ag}})$, $\ell(s, \widehat{f}_{\mathcal{T}}^{\text{cv}})$ and the oracle excess risk $\inf_{\lambda\in\Lambda} \ell(s, \mathcal{A}_\lambda(D_n))$. The risks (and excess risks) are evaluated using the L^1 loss $g(u,y) = |u-y|$. Expectations of these quantities are estimated by taking an average over the 1000 samples; we also compute standard deviations for these estimates, which are not displayed, since they are sufficiently small to ensure that visible "gaps" on the graph are statistically significant.

Agghoo and CV are applied to $(\mathcal{A}_\lambda)_{\lambda\in\Lambda}$ over the grid $\Lambda = \{\frac{2^{j-1}}{500 n_t} : 0 \leqslant j \leqslant 17\}$, corresponding to the grid $\{\frac{500}{2^j} : 0 \leqslant j \leqslant 17\}$ over the COST parameter $C = \frac{1}{2\lambda n_t}$ of the R implementation SVM from package E1071.

In a second step, V-fold agghoo and Monte-Carlo agghoo are compared with other averaging cross-validation procedures:

- ACV (Jung and Hu, 2015),
- EKCV (Jung, 2016),
- bagged K-FCV, that is, bagging applied to the K-fold CV predictor $\widehat{f}_{\mathcal{T}}^{\text{cv}}$ given by Definition 5, with $\mathcal{T} = \{\{1,\dots,n\}\backslash J_i : 1 \leqslant i \leqslant K\}$ for some partition $(J_i)_{1\leqslant i\leqslant K}$ of $\{1,\dots,n\}$ into K blocks of equal size $|J_i| = n_v = n/K$. Following the terminology detailed in Section 3.3, bagged K-FCV is a specific instance of bagged CV. Note that bagged K-FCV is *not* Hall's BCV.

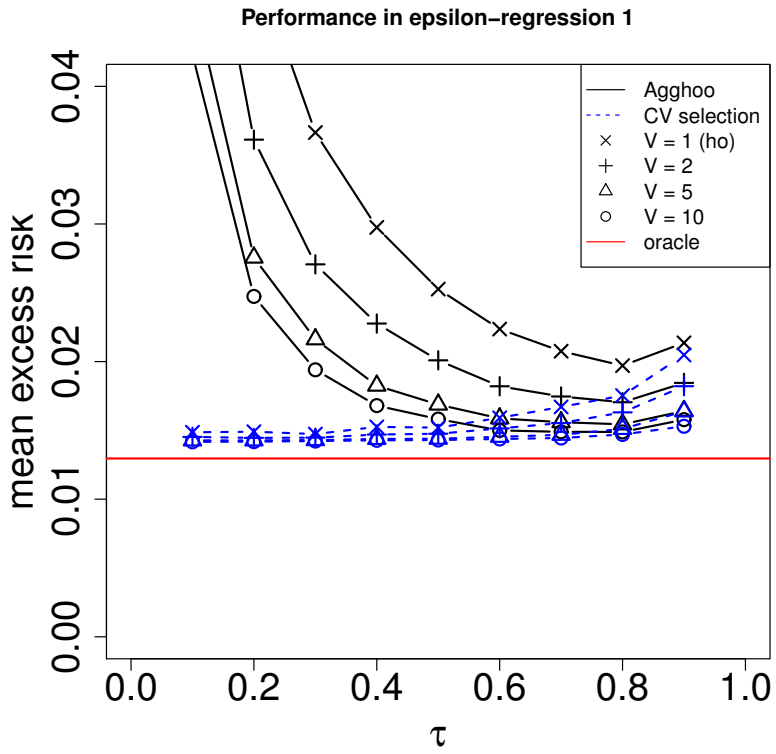

Figure 1: Performance of agghoo and CV for ε-regression in setup 1

These methods can be seen as having the same two hyperparameters: the fraction τ of data assigned to the training set in each fold of CV, and the number V of estimators—or parameters—being aggregated. Using the notation of Jung and Hu (2015) for ACV, we have $\tau = (K-1)/K$ and $V = K$ for some integer $K \geqslant 1$. Using the notation of Jung (2016) for EKCV, we have $\tau = \frac{K-1}{K}$ and $V = M$. For bagged K-FCV, V is the number of bagging resamples considered, and $\tau = (K-1)/K$ (or equivalently, $K = 1/(1-\tau)$). For all methods, we consider the choices $\tau \in \{0.8, 0.9\}$ and $V \in \{5, 10\}$. For ACV and EKCV, these are the values recommended by Jung and Hu (2015) and Jung (2016), respectively.

5.1.2 Setup 1

Data are generated as follows: $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent, with $X_i \sim \mathcal{N}(0, \pi^2)$, $Y_i = s(X_i) + Z_i$, with $Z_i \sim \mathcal{N}(0, 1/4)$ independent from X_i. The regression function is $s : x \mapsto \exp[\cos(x)]$, the kernel parameter is $h = \frac{1}{2}$ and the threshold for the ε-insensitive loss is $\varepsilon = \frac{1}{4}$.

Results for agghoo and CV in setup 1 are shown on Figure 1. The performance of agghoo strongly depends on both τ and V. For a fixed τ, increasing V significantly decreases the risk of the resulting estimator. This is not surprising and confirms that considering several data splits is always useful.

Most of the improvement occurs between $V = 1$ and $V = 5$, and taking V much larger seems useless—at least for $\tau \geqslant 0.5$—, a behavior previously observed for CV (Arlot and

τ	V	MC agghoo	V-fold agghoo	ACV	EKCV	Bagged K-FCV
0.8	5	2.48 ± 0.2	1.95 ± 0.2	1.66 ± 0.2	1.51 ± 0.2	5.46 ± 0.4
0.8	10	1.93 ± 0.2			1.47 ± 0.2	3.54 ± 0.3
0.9	5	3.46 ± 0.3			1.52 ± 0.2	6.03 ± 0.4
0.9	10	2.82 ± 0.2	2.35 ± 0.2	2.07 ± 0.2	1.38 ± 0.2	3.99 ± 0.3

Table 1: Setup 1, difference between the average risk of each method and the average risk of the oracle, multiplied by 10^3. The uncertainty is obtained using the formula $2 * \frac{\widehat{\sigma}}{\sqrt{n_s}}$, where $\widehat{\sigma}^2$ is the empirical variance and n_s is the number of simulations. Missing values occur when the given combination of τ and V is not allowed by the method; this is the case with ACV and V-fold agghoo, for which $\tau = \frac{V-1}{V}$.

Lerasle, 2016). For a fixed V, the risk strongly decreases when τ increases from 0.1 to 0.5, decreases slowly over the interval $[0.5, 0.8]$ and seems to rise for $\tau > 0.8$. It seems that $\tau \in [0.6, 0.9]$ yields the best performance, while taking τ close to 0 should clearly be avoided (at least for $V \leqslant 10$). Taking V large enough, say $V = 10$, makes the choice of τ less crucial: a large region of values of τ yields (almost) optimal performance. We do not know whether taking V larger can make the performance of agghoo with $\tau \leqslant 0.4$ close to the optimum.

As a function of τ, the risk of CV behaves quite differently from agghoo's. The performance does not degrade significantly when τ is small. The optimum is located around $\tau = 0.1$, but the risk curve is so flat that there is no perceptible difference between the values of $\tau \in [0.1, 0.4]$. In any case, the optimum is much smaller than for agghoo. A possible explanation is that the regressors produced by cross-validation are all trained on the whole sample, so that τ only impacts risk estimation. Furthermore, additional simulations show, as expected, that higher values of τ ($\tau = 0.8$ or $\tau = 0.9$) improve *risk estimation* while degrading the *hyperparameter selection* performance. Compared to agghoo, CV's performance depends much less on V: only $V = 2$ appears to be significantly worse than $V \geqslant 5$.

Let us now compare agghoo and CV. For small values of τ (that is, $\tau \leqslant 0.5$), agghoo generally performs much worse than CV for all values of V. In the case of the hold-out, this is unsurprising as the hold-out estimator is then trained on a much smaller sample than the CV estimator. Clearly, aggregation does not sufficiently compensate for this, at least for $V \leqslant 10$. On the other hand, for $\tau \in [0.6, 0.9]$, agghoo with $V = 10$ approximately matches CV's performance. The risks of the two methods are indistinguishable for $V = 10, \tau = 0.8$.

Comparison of agghoo with ACV, EKCV and bagged K-FCV in setup 1. According to the results summarized by Table 1, the best performing method in this experiment is EKCV, followed by ACV. The performance of EKCV does not vary very much over the tested values of V, τ, whereas other methods show stronger variation. Among the two agghoo methods, V-fold appears to perform better than Monte-Carlo for a given value of τ and equal or smaller value of V. Bagged K-FCV performs the worst out of all the methods, for all values of τ and V. Overall, in this simulation, the methods which select a single regressor from the collection $(\mathcal{A}_\lambda(D_n))_{\lambda \in \mathbb{R}_+}$ (CV, ACV and EKCV) generally perform better than the methods which aggregate them (agghoo and bagged K-FCV). This could be due to the fact that the regression function $\exp[\cos(x)]$ of setup 1 is very smooth (analytic) and bounded. Combined

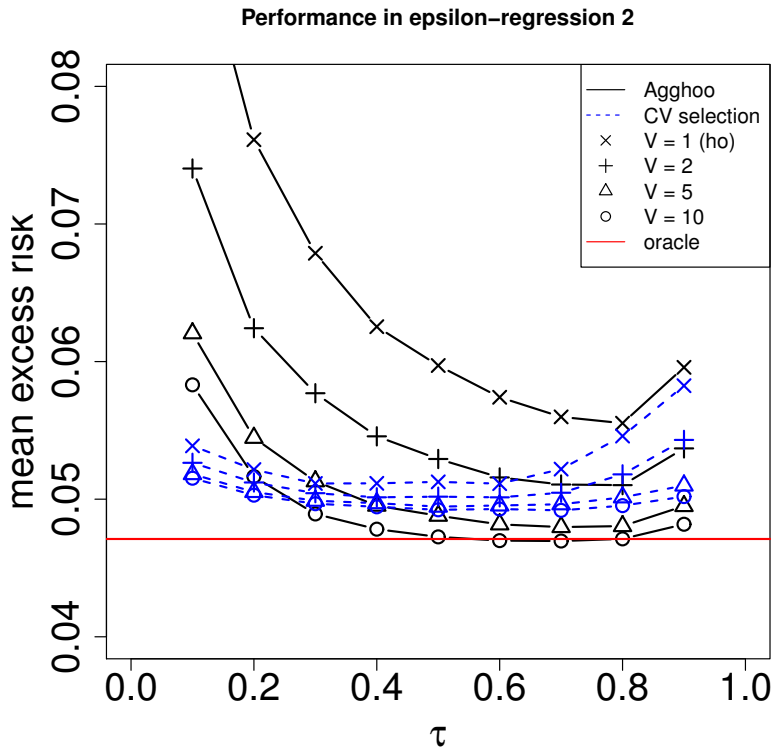

Figure 2: Performance of agghoo and CV for ε-regression in setup 2

with a one-dimensional variable X and Gaussian noise, this yields an "easy" non-parametric regression problem. As a result, the collection $(\mathcal{A}_\lambda(D_n))_{\lambda\in\mathbb{R}_+}$ may already have very good approximation properties and improving on it with aggregation may be difficult. Estimator aggregation could prove more useful in harder problems where s is less smooth and the dimension is higher. In order to assess this hypothesis, we carry out a second experiment.

5.1.3 Setup 2

Data are generated as follows: $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent, with $X_i \in \mathbb{R}^2$ distributed as Cauchy$(0,1)^{\otimes 2}$, $Y_i = s(X_i) + Z_i$, with $Z_i \sim \mathcal{N}(0, 1/4)$ independent from X_i. The regression function is defined almost everywhere by

$$s(x_1, x_2) = \frac{2\sin(x_1 x_2)}{x_1^2 + x_2^2} \ ,$$

the kernel parameter is $h = \frac{1}{2}$ and the threshold for the ε-insensitive loss is $\varepsilon = \frac{1}{4}$. This regression function is less regular than in the previous setup, since it has a discontinuity at $(0,0) \in \mathbb{R}^2$.

Results for agghoo and CV in setup 2 are shown on Figure 2. The qualitative conclusions about the behaviour of agghoo and CV, taken separately, are mostly the same as in setup 1, with the exception that CV now shows the expected increase in risk for the smallest values of τ.

τ	V	MC agghoo	V-fold agghoo	ACV	EKCV	Bagged K-FCV
0.8	5	0.94 ± 0.3	0.32 ± 0.2	2.1 ± 0.2	1.97 ± 0.2	6.97 ± 0.6
0.8	10	0.08 ± 0.2			3.66 ± 0.4	3.64 ± 0.5
0.9	5	2.39 ± 0.4			1.96 ± 0.2	8.81 ± 0.6
0.9	10	1.07 ± 0.3	0.72 ± 0.3	3.18 ± 0.4	3.65 ± 0.4	4.65 ± 0.5

Table 2: Setup 2, difference between the average risk of each method and the average risk of the oracle, multiplied by 10^3. The uncertainty is obtained using the formula $2 * \frac{\widehat{\sigma}}{\sqrt{n_s}}$, where $\widehat{\sigma}^2$ is the empirical variance and n_s is the number of simulations.

The main difference with setup 1 is that agghoo performs much better relative to CV and the oracle. For $V = 10$ and $\tau \in [0.4, 0.9]$, agghoo outperforms CV by a significant margin; for $V = 10$ and $\tau \in [0.6, 0.8]$, agghoo even matches the oracle's performance, up to statistical uncertainty.

Part of the explanation is that, on a given data set, agghoo can perform better than the oracle using aggregation whereas CV, as a parameter selection method, naturally cannot. Indeed, for a randomly drawn data set in setup 2, this situation can be observed to occur quite regularly.

Overall, if the computational cost of $V = 10$ data splits is not prohibitive, agghoo with optimized parameters ($V = 10$, $\tau \in [0.6, 0.8]$) clearly improves over CV with optimized parameters ($V = 10$, $\tau \in [0.5, 0.7]$). The same holds with $V = 5$.

Comparison of agghoo with ACV, EKCV and bagged K-FCV in setup 2. According to the results summarized by Table 2, aggregated hold-out is clearly the best performing method in setup 2, by a large margin. This shows the potential advantage of aggregating estimators (agghoo) rather than parameters (ACV, EKCV) when s is non-smooth. Overall, bagged K-FCV performs the worst, except for $(\tau, V) = (0.8, 10)$ where it is tied with EKCV. Its poor performance relative to agghoo can be explained by several factors: first, K-fold CV is more stable than the hold-out, which leads to a less diverse ensemble for aggregation. Secondly, bagging (sampling *with replacement*) breaks the independence between training and test samples, potentially leading to overfitting. Among the two agghoo methods, V-fold seems to outperform Monte-Carlo whenever both are defined, though the difference is not significant. However, the overall best performance is attained at $(\tau, V) = (0.8, 10)$, a combination which is not achievable using a V-fold subsampling scheme.

5.1.4 Computational Complexity

By Equation (3), regularized kernel regressors can be represented linearly by vectors of length n_t, therefore the aggregation step can be performed at training time by averaging these vectors. The complexity of this aggregation is at most $\mathcal{O}(V \times n_t)$. In general, this is negligible relative to the cost of computing the hold-out, as simply computing the kernel matrix requires $n_t(n_t + 1)/2$ kernel evaluations. Therefore, the aggregation step does not affect much the computational complexity of agghoo, so the conclusion of Section 3.4 that agghoo and CV have similar complexity applies in the present setting. The same holds for ACV and EKCV which rely on V-fold type splits. In contrast, bagged K-FCV has a higher

complexity, as one must carry out $1/(1-\tau)$-fold CV within each bagging sample. As a result, the hold-out must be computed $V/(1-\tau)$ times.

Evaluating agghoo, CV, ACV, EKCV and bagged K-FCV on new data $x \in \mathcal{X}$ also takes the same time in general, as all can be computed by evaluating $\sum_{j=1}^{n_t} \theta_j K(X_j, x)$ with a pre-computed value of θ. A potential difference occurs when the $\widehat{\theta}_\lambda$—given by Definition 9, Equation (3)—are sparse: aggregation increases the number of non-zero coefficients, so evaluating $\widehat{f}^{\mathrm{ag}}_{\mathcal{T}}$ on new data can be slower than evaluating $\widehat{f}^{\mathrm{cv}}_{\mathcal{T}}$ (or ACV and EKCV) if the implementation is designed to take advantage of sparsity.

5.2 k-Nearest-Neighbors Classification

We consider the collection $(\mathcal{A}_k^{\mathrm{NN}})_{k \geqslant 1,\, k \text{ odd}}$ of nearest-neighbors classifiers—assuming k is odd to avoid ties— on the following binary classification problem.

5.2.1 Experimental Setup

Data $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent, with X_i uniformly distributed over $\mathcal{X} = [0,1]^2$ and

$$\mathbb{P}(Y_i = 1 \,|\, X_i) = \sigma\left(\frac{h(X_i) - b}{\lambda}\right)$$

$$\text{where } \forall u, v \in \mathbb{R}, \qquad \sigma(u) = \frac{1}{1+\mathrm{e}^{-u}} \quad \text{and} \quad h\big((u,v)\big) = \exp\big[-(u^2+v)^3\big] + u^2 + v^2 \ ,$$

$b = 1.18$ and $\lambda = 0.05$. The Bayes classifier is $s : x \mapsto \mathbb{I}_{h(x) \geqslant b}$ and the Bayes risk, computed numerically using the scipy.integrate python library, is approximately equal to 0.242. Majhoo (the classification version of agghoo, see Definition 7) and CV are used with the collection $(\mathcal{A}_k^{\mathrm{NN}})_{k \geqslant 1,\, k \text{ odd}}$ and "Monte-Carlo" training sets as in Section 5.1. An experimental procedure similar to the one of Section 5.1 is used to evaluate the performance of agghoo and to compare it with Monte-Carlo cross-validation. Standard deviations of the excess risk were computed; they are smaller than 3.6% of the estimated value.

5.2.2 Results

As shown by Figure 3, the results are similar to the regression case (see Section 5.1), with a few differences. First, agghoo does not perform better than the oracle. In fact, all methods considered here remain far from the oracle, which has an excess risk around 0.0034 ± 0.0004; both agghoo and CV have excess risks at least 4 times larger. Second, risk curves as a function of τ for agghoo are almost U-shaped, with a significant rise of the risk for $\tau > 0.6$. Therefore, less data is needed for training, compared to Section 5.1. The optimal value of τ here is 0.6, at least for some values of V, up to statistical error. Third, the performance of CV as a function of τ has a similar U-shape, which makes the comparison between agghoo and CV easier. For a given τ, agghoo performs significantly better if $V \geqslant 10$, while CV performs significantly better if $V = 2$; the difference is mild for $V = 5$.

5.2.3 Alternative Methods

Neither ACV nor EKCV can be applied to k-nearest neighbors, as there are no models and the parameter k of k-NN cannot be averaged, as it is an integer. Bagged K-FCV can be

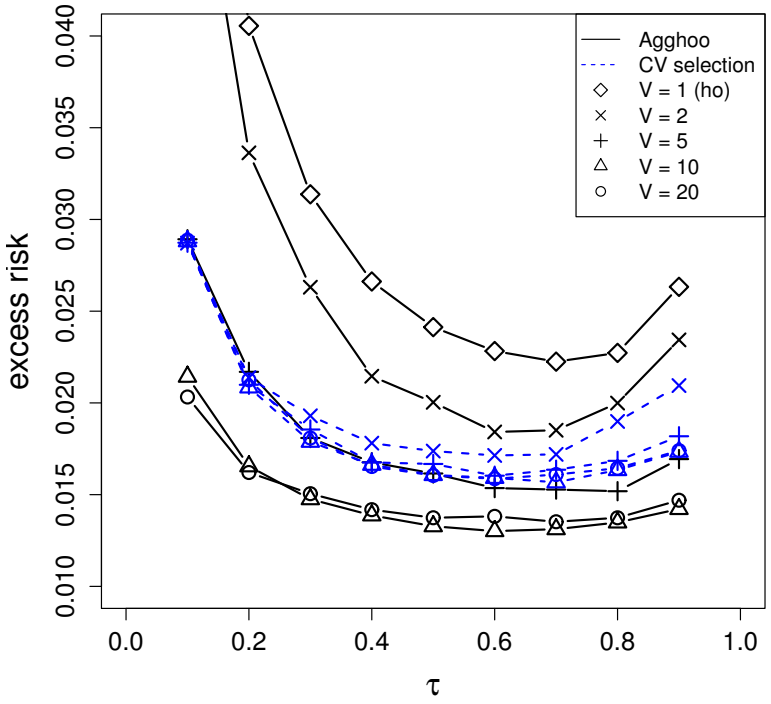

Figure 3: Classification performance of majhoo and CV for the k-NN family

used here, however it performs very poorly: using V bagging samples and K-fold CV with $(V, K) \in \{5, 10\}^2$ yields excess risks close to 0.072, again with negligible error (estimated standard deviation 8×10^{-4}), hence two to five times worse than agghoo.

Further investigations reveal that within a bagging sample, CV often chooses $k = 1$, so that bagged K-FCV practically reduces to bagged 1-NN. A plausible explanation for this is that sampling with replacement leads to the same data point appearing in both training and test sets; 1-NN perfectly classifies these repeated samples, which "artificially" improves its CV score.

5.2.4 Computational Complexity

As explained in Section 3.4, the complexity of computing the optimal parameters for CV $(\widehat{k}^{\mathrm{cv}}_{\mathcal{T}})$ is the same as for computing $(\widehat{k}^{\mathrm{ho}}_{T})_{T \in \mathcal{T}}$ for majhoo. Here, there is no simple way to represent the aggregated estimator, so aggregation may have to be performed at test time. In that case, the complexity of evaluating majhoo on new data is roughly V times greater than for CV, as explained in Section 3.4 for agghoo.

6. Discussion

Theoretical and numerical results of the paper show that agghoo can be used safely in RKHS regression, at least when its parameters are properly chosen; $V \geqslant 10$ and $\tau = 0.8$ seem to be safe choices. A variant, majhoo, can be used in supervised classification with the 0–1 loss, with a general guarantee on its performance (Theorem 13). Experiments show that

agghoo/majhoo actually performs much better than what the upper bounds of Section 4 suggest. In one simulation setup, it roughly matches CV's performance for well chosen V, τ. In two others setups, it outperforms CV by a significant margin, as long as $V \geqslant 5$ splits are used. Proving theoretically that agghoo can improve over CV is an open problem that deserves future works, solved in a specific setting during the revision of this article (Maillard, 2020b, Chapters 5–6).

Since agghoo and CV have the same training computational cost for any fixed (V, τ), agghoo—with properly chosen parameters V, τ—is competitive with CV in practice, unless aggregation is undesirable for some other reason, such as interpretability of the predictors, or computational complexity at test time.

Our results can be extended in several ways. First, our theoretical bounds directly apply to subagging hold-out, which also averages several hold-out selected estimators. As explained in Section 3.3, the difference with agghoo is that, in subagging, the training set size is $n - p - q$ and the validation set size is q, for some $q \in \{1, \dots, n - p - 1\}$, leading to slightly worse bounds than those we obtained for agghoo (at least if $\mathbb{E}\left[\ell(s, \mathcal{A}_m(D_n))\right]$ decreases with n). The difference should not be large in practice, if q is well chosen.

Oracle inequalities can also be obtained for agghoo in other settings, as a consequence of our general Theorems 16 and 17 in Appendix A. Since the first version of this paper appeared as a preprint, such results have been obtained in two settings. Maillard (2020a) applies Theorem 17 to sparse linear regression with the Huber loss function. Maillard (2020b, Chapter 6) applies Theorem 17 to the collection of empirical Fourier projections in least-squares density estimation, as a preliminary step to a deeper study of agghoo in that setting.

Acknowledgements

While finishing the writing of this article, the first author (Guillaume Maillard) has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 811017.

Appendix A. General Theorems

We need the following hypothesis, defined for two functions $w_i : \mathbb{R}_+ \to \mathbb{R}_+$, $i \in \{1, 2\}$ and a family $(t_m)_{m \in \mathcal{M}} \in \mathbb{S}^{\mathcal{M}}$.

Hypothesis $H(w_1, w_2, (t_m)_{m \in \mathcal{M}})$: w_1 and w_2 are nondecreasing, and for any $(m, m') \in \mathcal{M}^2$, some $c_{m'}^m \in \mathbb{R}$ exists such that, for all $k \geqslant 2$,

$$P\Big(\big|\gamma(t_m) - \gamma(t_{m'}) - c_{m'}^m\big|^k\Big) \leqslant k! \Big[w_1\big(\sqrt{\ell(s, t_m)}\big) + w_1\big(\sqrt{\ell(s, t_{m'})}\big)\Big]^2 \times \Big[w_2\big(\sqrt{\ell(s, t_m)}\big) + w_2\big(\sqrt{\ell(s, t_{m'})}\big)\Big]^{k-2} .$$

This hypothesis is similar to those used by Massart (2007) to study the hold-out and empirical risk minimizers. However, unlike Massart (2007), we intend to go beyond the setting of bounded risks.

We also need the following definition.

Definition 14 *Let $w : \mathbb{R}_+ \to \mathbb{R}_+$ and $r \in \mathbb{R}_+$. Let*

$$\delta(w, r) = \inf\left\{\delta \geqslant 0 : \forall x \geqslant \delta, w(x) \leqslant rx^2\right\},$$

with the convention $\inf \emptyset = +\infty$.

Remark 15

- *If $r > 0$ and $x \mapsto \frac{w(x)}{x}$ is nonincreasing, then $\delta(w, r)$ is the unique solution to the equation $\frac{w(x)}{x} = rx$.*
- *$r \mapsto \delta(w, r)$ is nonincreasing.*
- *If $w(x) = cx^\beta$ for $c > 0$ and $\beta \in [0, 2)$, then $\delta(w, r) = \left(\frac{c}{r}\right)^{\frac{1}{2-\beta}}$.*

A.1 Theorem Statements

We can now state two general theorems from which we deduce all the theoretical results of the paper. The first theorem is a general oracle inequality for the hold-out.

Theorem 16 *Let $(t_m)_{m\in\mathcal{M}}$ be a finite collection in $\mathbb{S}$, and*

$$\widehat{m} \in \operatorname*{argmin}_{m\in\mathcal{M}} P_{n_v}\gamma(t_m, \cdot) \ .$$

Assume that $H(w_1, w_2, (t_m)_{m\in\mathcal{M}})$ holds true. Let $x > 0$. Then, with probability larger than $1 - \mathrm{e}^{-x}$, for any $\theta \in (0, 1]$, we have

$$\begin{aligned}(1-\theta)\,\ell(s, t_{\widehat{m}}) \leqslant (1+\theta) \min_{m\in\mathcal{M}}\left\{\ell(s, t_m)\right\} + \sqrt{2}\theta\delta^2\left(w_1, \frac{\theta}{2}\sqrt{\frac{n_v}{x + \log|\mathcal{M}|}}\right) \\ + \frac{\theta^2}{2}\delta^2\left(w_2, \frac{\theta^2}{4}\frac{n_v}{x + \log|\mathcal{M}|}\right).\end{aligned} \tag{7}$$

If in addition, the two functions $x \mapsto \frac{w_j(x)}{x}$, $j = 1, 2$, are nonincreasing, then for any $x > 0$, with probability larger than $1 - \mathrm{e}^{-x}$, for all $\theta \in (0, 1]$, we have

$$\begin{aligned}(1-\theta)\ell(s, t_{\widehat{m}}) \leqslant (1+\theta) \min_{m\in\mathcal{M}}\left\{\ell(s, t_m)\right\} + \delta^2(w_1, \sqrt{n_v})\left[\theta + \frac{2(x + \log|\mathcal{M}|)}{\theta}\right] \\ + \delta^2(w_2, n_v)\left[\theta + \frac{(x + \log|\mathcal{M}|)^2}{\theta}\right] \ .\end{aligned} \tag{8}$$

Using Theorem 16, we prove the following general oracle inequality for agghoo.

Theorem 17 *Assume that the hyperparameter space $\mathbb{S}$ is convex and that the risk $\mathcal{L}$ is convex. Let $(\mathcal{A}_m)_{m\in\mathcal{M}}$ be a finite collection of learning rules of size $|\mathcal{M}| \geqslant 3$. Let $\widehat{f}^{\mathrm{ag}}_{\mathcal{T}}$ be an agghoo estimator, according to Definition 6, with $\mathcal{T}$ satisfying assumption (2). Assume that $\widehat{w}_{1,1}, \widehat{w}_{1,2}$ are D_{n_t}-measurable random functions such that almost surely, hypothesis $H\big(\widehat{w}_{1,1}, \widehat{w}_{1,2}, (\mathcal{A}_m(D_{n_t}))_{m\in\mathcal{M}}\big)$ holds true. Assume also that for $i \in \{1, 2\}$, $x \mapsto \frac{\widehat{w}_{1,i}(x)}{x}$ is nonincreasing. Then for any $\theta \in (0, 1]$,*

$$(1-\theta)\mathbb{E}\Big[\ell\big(s, \widehat{f}^{\mathrm{ag}}_{\mathcal{T}}\big)\Big] \leqslant (1+\theta)\mathbb{E}\left[\min_{m\in\mathcal{M}} \ell\big(s, \mathcal{A}_m(D_{n_t})\big)\right] + R_1(\theta) \tag{9}$$

where $R_1(\theta) = R_{1,1}(\theta) + R_{1,2}(\theta)$ *with*

$$R_{1,1}(\theta) = \left(\theta + \frac{2\big(1+\log|\mathcal{M}|\big)}{\theta}\right)\mathbb{E}\Big[\delta^2\big(\widehat{w}_{1,1}, \sqrt{n_v}\big)\Big] \ ,$$
$$R_{1,2}(\theta) = \left(\theta + \frac{2\big(1+\log|\mathcal{M}|\big) + \log^2|\mathcal{M}|}{\theta}\right)\mathbb{E}\Big[\delta^2\big(\widehat{w}_{1,2}, n_v\big)\Big] \ .$$

For any D_{n_t}*-measurable functions* $\widehat{w}_{2,1}$ *and* $\widehat{w}_{2,2}$ *such that* $H(\widehat{w}_{2,1}, \widehat{w}_{2,2}, (\mathcal{A}_m(D_{n_t}))_{m\in\mathcal{M}})$ *holds true almost surely, and any* $x > 0$, $\theta \in (0,1]$, *we have*

$$(1-\theta)\mathbb{E}\Big[\ell\big(s, \widehat{f}^{\mathrm{ag}}_{\mathcal{T}}\big)\Big] \leqslant (1+\theta)\mathbb{E}\left[\min_{m\in\mathcal{M}} \ell\big(s, \mathcal{A}_m(D_{n_t})\big)\right] + R_2(\theta) \tag{10}$$

where $R_2(\theta) = R_{2,1}(\theta) + R_{2,2}(\theta) + R_{2,3}(\theta) + R_{2,4}(\theta)$ *with*

$$R_{2,1}(\theta) = \sqrt{2}\theta\mathbb{E}\left[\delta^2\left(\widehat{w}_{2,1}, \frac{\theta}{2}\sqrt{\frac{n_v}{x+\log|\mathcal{M}|}}\right)\right] \ ,$$
$$R_{2,2}(\theta) = \frac{\theta^2}{2}\mathbb{E}\left[\delta^2\left(\widehat{w}_{2,2}, \frac{\theta^2}{4}\frac{n_v}{x+\log|\mathcal{M}|}\right)\right] \ ,$$
$$R_{2,3}(\theta) = \mathrm{e}^{-x}\left(\theta + \frac{2(1+x+\log|\mathcal{M}|)}{\theta}\right)\mathbb{E}\Big[\delta^2\big(\widehat{w}_{1,1}, \sqrt{n_v}\big)\Big] \ ,$$
and $$R_{2,4}(\theta) = \mathrm{e}^{-x}\left(\theta + \frac{2(1+x+\log|\mathcal{M}|) + (x+\log|\mathcal{M}|)^2}{\theta}\right)\mathbb{E}\Big[\delta^2\big(\widehat{w}_{1,2}, n_v\big)\Big] \ .$$

A.2 Proof of Theorem 16

We start by proving three lemmas.

Lemma 18 *Let* w *be a nondecreasing function on* $\mathbb{R}_+$. *Let* $r > 0$. *Then*

$$\forall u \geqslant 0\,, \qquad w(u) \leqslant r\big(u^2 \vee \delta^2(w,r)\big) \ ,$$

where $\delta(w,r)$ *is given by Definition 14.*

Proof If $u > \delta(w,r)$, by Definition 14,

$$w(u) \leqslant ru^2.$$

If $u \leqslant \delta(w,r)$, since w is nondecreasing, for all $v > \delta(w,r)$,

$$w(u) \leqslant w(v) \leqslant rv^2.$$

By taking the infimum over v, we recover $w(u) \leqslant r\delta(w,r)^2$. ■

Lemma 19 *Let* w *be a nondecreasing function such that* $x \mapsto \frac{w(x)}{x}$ *is nonincreasing over* $(0,+\infty)$. *Let* $a \in \mathbb{R}_+$ *and* $b \in (0,+\infty)$. *For any* $\theta \in (0,1]$ *and* $u \geqslant 0$,

$$\frac{a}{b}w(\sqrt{u}) \leqslant \frac{\theta}{2}\big[u + \delta^2(w,b)\big] + \frac{a^2\delta^2(w,b)}{2\theta} \ .$$

Proof Since w is nondecreasing,

$$\begin{aligned} w(\sqrt{u}) &\leqslant w\big(\sqrt{u+\delta^2(w,b)}\big) \\ &= \sqrt{u+\delta^2(w,b)}\frac{w\big(\sqrt{u+\delta^2(w,b)}\big)}{\sqrt{u+\delta^2(w,b)}}. \end{aligned}$$

Since $\frac{w(x)}{x}$ is nonincreasing and $\delta(w,b)>0$,

$$\begin{aligned} w(\sqrt{u}) &\leqslant \sqrt{u+\delta^2(w,b)}\frac{w\big(\delta(w,b)\big)}{\delta(w,b)} \\ &\leqslant \sqrt{u+\delta^2(w,b)}\, b\, \delta(w,b) \qquad \text{by Definition 14.} \end{aligned}$$

Therefore, using the inequality $\sqrt{xy} \leqslant \frac{\theta}{2}x + \frac{y}{2\theta}$, valid for any $x>0, y>0$,

$$\frac{a}{b}w(\sqrt{u}) \leqslant \sqrt{a^2\big[u+\delta(w,b)^2\big]\delta(w,b)^2} \leqslant \frac{\theta}{2}\big[u+\delta(w,b)^2\big] + \frac{a^2\delta(w,b)^2}{2\theta} \ .$$

■

Lemma 20 *Let $n_v \in \mathbb{N}^*$. Let $\mathcal{M}$ be a finite set and let $(t_m)_{m\in\mathcal{M}} \in \mathbb{S}^{\mathcal{M}}$. Assume that there exists $p \in [0, 1/|\mathcal{M}|)$ and a function $R : (0,1] \to \mathbb{R}_+$ such that for any m, m' in $\mathcal{M}$, with probability greater than $1-p$,*

$$\forall \theta \in (0,1]\,, \qquad (P_{n_v}-P)\big[\gamma(t_m,\cdot)-\gamma(t_{m'},\cdot)\big] \leqslant \theta\ell(s,t_m)+\theta\ell(s,t_{m'})+R(\theta) \ .$$

Then, with probability greater than $1-|\mathcal{M}|p$, for any $\widehat{m} \in \operatorname{argmin}_{m\in\mathcal{M}} P_{n_v}\gamma(t_m,\cdot)$,

$$\forall \theta \in (0,1]\,, \qquad (1-\theta)\ell(s,t_{\widehat{m}}) \leqslant (1+\theta)\min_{m\in\mathcal{M}}\big\{\ell(s,t_m)\big\}+R(\theta) \ .$$

Proof Let $m_* \in \operatorname{argmin}_{m\in\mathcal{M}} P\gamma(t_m,\cdot)$. Then for any $m \in \mathcal{M}$, with probability greater than $1-p$,

$$\forall \theta \in (0,1]\,, \qquad (P_{n_v}-P)\big[\gamma(t_{m_*},\cdot)-\gamma(t_m,\cdot)\big] \leqslant \theta\ell(s,t_{m_*})+\theta\ell(s,t_m)+R(\theta) \ .$$

So by the union bound, with probability greater than $1-|\mathcal{M}|p$,

$$\forall \theta \in (0,1]\,, \forall m \in \mathcal{M}\,, \qquad (P_{n_v}-P)\big[\gamma(t_{m_*},\cdot)-\gamma(t_m,\cdot)\big] \leqslant \theta\ell(s,t_{m_*})+\theta\ell(s,t_m)+R(\theta) \ .$$

On that event, for all $\theta \in (0,1]$,

$$\begin{aligned} P\gamma(t_{\widehat{m}},\cdot) &= P_{n_v}\gamma(t_{\widehat{m}},\cdot)+(P-P_{n_v})\gamma(t_{\widehat{m}},\cdot) \\ &\leqslant P_{n_v}\gamma(t_{m_*},\cdot)+(P-P_{n_v})\gamma(t_{\widehat{m}},\cdot) \\ &= P\gamma(t_{m_*},\cdot)+(P-P_{n_v})\big[\gamma(t_{\widehat{m}},\cdot)-\gamma(t_{m_*},\cdot)\big] \\ &\leqslant P\gamma(t_{m_*},\cdot)+\theta\ell(s,t_{m_*})+\theta\ell(s,t_{\widehat{m}})+R(\theta) \ . \end{aligned}$$

Subtracting the Bayes risk $P\gamma(s,\cdot)$ on both sides, we get with probability greater than $1-|\mathcal{M}|p$, for all $\theta\in(0,1]$,

$$\ell(s,t_{\widehat{m}}) \leqslant \ell(s,t_{m_*}) + \theta\ell(s,t_{m_*}) + \theta\ell(s,t_{\widehat{m}}) + R(\theta) \ ,$$
$$\text{that is,} \qquad (1-\theta)\ell(s,t_{\widehat{m}}) \leqslant (1+\theta)\min_{m\in\mathcal{M}}\big\{\ell(s,t_m)\big\} + R(\theta) \ .$$

■

We now prove Theorem 16. Let $(m,m')\in\mathcal{M}^2$ be fixed. Let

$$\sigma := w_1\Big(\sqrt{\ell(s,t_m)}\Big) + w_1\Big(\sqrt{\ell(s,t_{m'})}\Big) \ ,$$
$$\text{and} \qquad c := w_2\Big(\sqrt{\ell(s,t_m)}\Big) + w_2\Big(\sqrt{\ell(s,t_{m'})}\Big) \ .$$

By hypothesis $H\big(w_1,w_2,(t_m)_{m\in\mathcal{M}}\big)$,

$$\exists c^m_{m'}\in\mathbb{R} \quad \text{such that} \quad \forall k\geqslant 2\,, \qquad P\big(\gamma(t_m,\cdot)-\gamma(t_{m'},\cdot)-c^m_{m'}\big)^k \leqslant k!\,\sigma^2c^{k-2} \ .$$

For all $y>0$, let $\Omega_y(m,m')$ be the event on which

$$(P_{n_v}-P)\big[\gamma(t_m,\cdot)-\gamma(t_{m'},\cdot)\big] \leqslant \sqrt{\frac{2y}{n_v}}\sigma + \frac{cy}{n_v} \ . \tag{11}$$

By Bernstein's inequality (Boucheron et al., 2013, Theorem 2.10),

$$\mathbb{P}\big(\Omega_y(m,m')\big) \geqslant 1-\mathrm{e}^{-y} \ .$$

Let $q=\frac{\theta}{2}\sqrt{\frac{n_v}{x+\log|\mathcal{M}|}}$. By Lemma 18 with $r=q$,

$$\sigma := w_1\Big(\sqrt{\ell(s,t_m)}\Big) + w_1\Big(\sqrt{\ell(s,t_{m'})}\Big) \leqslant q\big[\ell(s,t_m)\vee\delta^2(w_1,q) + \ell(s,t_{m'})\vee\delta^2(w_1,q)\big] \ .$$

Set $y=x+\log|\mathcal{M}|$ in Equation (11). Then,

$$\begin{aligned}
\sqrt{\frac{2y}{n_v}}\sigma &:= \sqrt{\frac{2(x+\log|\mathcal{M}|)}{n_v}}\sigma \\
&\leqslant \sqrt{\frac{2(x+\log|\mathcal{M}|)}{n_v}}\frac{\theta}{2}\sqrt{\frac{n_v}{x+\log|\mathcal{M}|}}\big[\ell(s,t_m)\vee\delta^2(w_1,q) + \ell(s,t_{m'})\vee\delta^2(w_1,q)\big] \\
&\leqslant \frac{\theta}{\sqrt{2}}\left[\ell(s,t_m)+\ell(s,t_{m'}) + 2\delta^2\left(w_1,\frac{\theta}{2}\sqrt{\frac{n_v}{x+\log|\mathcal{M}|}}\right)\right] \ .
\end{aligned} \tag{12}$$

As for the second term of Equation (11), by Lemma 18 with $r=q^2$, we have

$$c := w_2(\sqrt{\ell(s,t_m)}) + w_2(\sqrt{\ell(s,t_{m'})}) \leqslant q^2\big[\ell(s,t_m)\vee\delta^2(w_2,q^2) + \ell(s,t_{m'})\vee\delta^2(w_2,q^2)\big] \ .$$

Recall that q is shorthand for $\frac{\theta}{2}\sqrt{\frac{n_v}{x+\log|\mathcal{M}|}}$. Therefore,

$$\begin{aligned} c\frac{y}{n_v} &\leqslant \frac{x+\log|\mathcal{M}|}{n_v}\frac{\theta^2}{4}\frac{n_v}{x+\log|\mathcal{M}|}\big[\ell(s,t_m)\vee\delta^2(w_2,q^2)+\ell(s,t_{m'})\vee\delta^2(w_2,q^2)\big] \\ &= \frac{\theta^2}{4}\big[\ell(s,t_m)\vee\delta^2(w_2,q^2)+\ell(s,t_{m'})\vee\delta^2(w_2,q^2)\big] \\ &\leqslant \frac{\theta^2}{4}\left[\ell(s,t_m)+\ell(s,t_{m'})+2\delta^2\left(w_2,\frac{\theta^2}{4}\frac{n_v}{x+\log|\mathcal{M}|}\right)\right] \quad . \end{aligned} \tag{13}$$

Since $\sqrt{\frac{1}{2}}+\frac{1}{4}\leqslant 1$ and $\theta\in(0,1]$, plugging Equations (12) and (13) in Equation (11) yields, on the event $\Omega_{x+\log|\mathcal{M}|}(m,m')$, for all $\theta\in(0,1]$,

$$\begin{aligned} (P_{n_v}-P)\big[\gamma(t_m,\cdot)-\gamma(t_{m'},\cdot)\big] \leqslant \theta\big[\ell(s,t_m)+\ell(s,t_{m'})\big] &+ \sqrt{2}\theta\delta^2\left(w_1,\frac{\theta}{2}\sqrt{\frac{n_v}{x+\log|\mathcal{M}|}}\right) \\ &+ \frac{\theta^2}{2}\delta^2\left(w_2,\frac{\theta^2}{4}\frac{n_v}{x+\log|\mathcal{M}|}\right) \quad . \end{aligned} \tag{14}$$

Suppose now that $x\mapsto\frac{w_j(x)}{x}$ is nonincreasing for $j\in\{1,2\}$. Let $\theta\in(0,1]$. Let $y\geqslant 0$. By Lemma 19 with $a=\sqrt{2y}$ and $b=\sqrt{n_v}$,

$$\begin{aligned} \sqrt{\frac{2y}{n_v}}\sigma &= \sqrt{\frac{2y}{n_v}}\left[w_1\big(\sqrt{\ell(s,t_m)}\big)+w_1\big(\sqrt{\ell(s,t_{m'})}\big)\right] \\ &\leqslant \frac{\theta}{2}\ell(s,t_m)+\frac{\theta}{2}\ell(s,t_{m'})+\delta^2(w_1,\sqrt{n_v})\left(\theta+\frac{2y}{\theta}\right) \quad . \end{aligned} \tag{15}$$

By Lemma 19 with $a=y$ and $b=n_v$,

$$\begin{aligned} c\frac{y}{n_v} &= \frac{y}{n_v}\left[w_2\big(\sqrt{\ell(s,t_m)}\big)+w_2\big(\sqrt{\ell(s,t_{m'})}\big)\right] \\ &\leqslant \frac{\theta}{2}\ell(s,t_m)+\frac{\theta}{2}\ell(s,t_{m'})+\delta^2(w_2,n_v)\left[\theta+\frac{y^2}{\theta}\right]. \end{aligned} \tag{16}$$

Plugging Equations (15) and (16) in Equation (11) yields, on the event $\Omega_y(m,m')$, for all $\theta\in(0,1]$,

$$\begin{aligned} &(P_{n_v}-P)\big[\gamma(t_m,\cdot)-\gamma(t_{m'},\cdot)\big] \\ &\leqslant \theta\ell(s,t_m)+\theta\ell(s,t_{m'})+\delta^2(w_1,\sqrt{n_v})\left(\theta+\frac{2y}{\theta}\right)+\delta^2(w_2,n_v)\left(\theta+\frac{y^2}{\theta}\right) \quad . \end{aligned} \tag{17}$$

By Equation (14), Lemma 20 applies with $p=\exp(-x)/|\mathcal{M}|$ and

$$R(\theta)=\sqrt{2}\theta\delta^2\left(w_1,\frac{\theta}{2}\sqrt{\frac{n_v}{x+\log|\mathcal{M}|}}\right)+\frac{\theta^2}{2}\delta^2\left(w_2,\frac{\theta^2}{4}\frac{n_v}{x+\log|\mathcal{M}|}\right) \quad .$$

This yields Equation (7). By Equation (17), Lemma 20 applies with $p=\mathrm{e}^{-y}$ and

$$R(\theta)=\delta^2(w_1,\sqrt{n_v})\left[\theta+\frac{2y}{\theta}\right]+\delta^2(w_2,n_v)\left(\theta+\frac{y^2}{\theta}\right) \quad .$$

Setting $y=\log|\mathcal{M}|+x$ yields Equation (8). ■

A.3 Proof of Theorem 17

We start by proving two lemmas.

Lemma 21 *Let $f \in L^1(\mathbb{R}_+, \mathrm{e}^{-x}\mathrm{d}x)$ be a non-negative, nondecreasing function such that $\lim_{x\to+\infty} f(x) = +\infty$. Let X be a random variable such that*

$$\forall x \in \mathbb{R}_+ \,, \qquad \mathbb{P}\big(X > f(x)\big) \leqslant \mathrm{e}^{-x} \ .$$

Then

$$\mathbb{E}[X] \leqslant \int_0^{+\infty} f(x)\mathrm{e}^{-x}\mathrm{d}x \ .$$

Proof Let $g \in L^1(\mathbb{R}_+, \mathrm{e}^{-x}\mathrm{d}x)$ be a nondecreasing, differentiable function such that $g \geqslant f$. Then

$$\begin{aligned}
\mathbb{E}[X] &\leqslant \int_0^{+\infty} \mathbb{P}(X > t)\mathrm{d}t \\
&= \int_0^{g(0)} \mathbb{P}(X > t)\mathrm{d}t + \int_0^{+\infty} \mathbb{P}\big(X > g(x)\big)g'(x)\mathrm{d}x \\
&\leqslant g(0) + \int_0^{+\infty} \mathrm{e}^{-x}g'(x)\mathrm{d}x \quad \text{since } g \geqslant f \\
&= g(0) + \big[\mathrm{e}^{-x}g(x)\big]_0^{\infty} + \int_0^{+\infty} \mathrm{e}^{-x}g(x)\mathrm{d}x \\
&= \int_0^{+\infty} \mathrm{e}^{-x}g(x)\mathrm{d}x \ .
\end{aligned}$$

It remains to show that g can approximate f in $L^1(\mathbb{I}_{x\geqslant 0}\mathrm{e}^{-x}\mathrm{d}x)$. Let K be a nonnegative smooth function vanishing outside $[-1, 1]$, normalized such that $\int K(t)\mathrm{d}t = 1$. Let $\varepsilon > 0$. Define

$$f_\varepsilon(x) = \frac{1}{\varepsilon}\int f(t)K\left(\frac{x+\varepsilon-t}{\varepsilon}\right)\mathrm{d}t \tag{18}$$

$$= \frac{1}{\varepsilon}\int f(x+\varepsilon-t)K\left(\frac{t}{\varepsilon}\right)\mathrm{d}t \tag{19}$$

By Equation (18), f_ε is smooth. By Equation (19), f_ε is nondecreasing, moreover

$$\begin{aligned}
f_\varepsilon(x) - f(x) &= \frac{1}{\varepsilon}\int \big[f(x+\varepsilon-t) - f(x)\big]K\left(\frac{t}{\varepsilon}\right)\mathrm{d}t \qquad \text{since} \int K = 1 \\
&= \frac{1}{\varepsilon}\int_{-\varepsilon}^{\varepsilon} \big[f(x+\varepsilon-t) - f(x)\big]K\left(\frac{t}{\varepsilon}\right)\mathrm{d}t \qquad \text{since } K(u) = 0 \text{ when } |u| \geqslant 1 \\
&\geqslant 0 \qquad \text{since } f \text{ is nondecreasing and } K \geqslant 0 \ .
\end{aligned}$$

Thus $f_\varepsilon \geqslant f$. Finally, by Jensen's inequality and Fubini's theorem,

$$\begin{aligned}
\int \big|f_\varepsilon(x) - f(x)\big|\mathrm{e}^{-x}\mathrm{d}x &\leqslant \frac{1}{\varepsilon}\int_{-\varepsilon}^{\varepsilon} K\left(\frac{t}{\varepsilon}\right)\int \big|f(x+\varepsilon-t) - f(x)\big|\mathrm{e}^{-x}\mathrm{d}x\mathrm{d}t \\
&\leqslant \sup_{|\tau|\leqslant 2\varepsilon} \int \big|f(x+\tau) - f(x)\big|\mathrm{e}^{-x}\mathrm{d}x \ ,
\end{aligned}$$

which converges to 0 when $\varepsilon \to 0$ since $f \in L^1(\mathbb{R}_+, \mathrm{e}^{-x}\mathrm{d}x)$. ■

We use the following additional notation:

Definition 22 *Let g be the function defined by*

$$\forall (\theta, y, p, q) \in (0,1] \times \mathbb{R}_+^3\,, \qquad g(\theta, y, p, q) = \theta[p+q] + \frac{1}{\theta}(2yp + y^2 q) \ .$$

This function satisfies the following properties.

Lemma 23 *Let g be the function given in Definition 22. For any $\theta \in (0,1]$ and any real numbers $u > 0, p \geqslant 0, q \geqslant 0$,*

$$\mathrm{e}^u \int_u^{+\infty} g(\theta, y, p, q)\mathrm{e}^{-y}\mathrm{d}y = \left(\theta + \frac{2(1+u)}{\theta}\right) p + \left(\theta + \frac{2+2u+u^2}{\theta}\right) q \ .$$

Proof Using the formulae

$$\int_u^{+\infty} \mathrm{e}^{-x}\mathrm{d}x = \mathrm{e}^{-u}\,, \qquad \int_u^{+\infty} x\mathrm{e}^{-x}\mathrm{d}x = (1+u)\mathrm{e}^{-u}$$
$$\text{and} \qquad \int_u^{+\infty} x^2\mathrm{e}^{-x}\mathrm{d}x = (2+2u+u^2)\mathrm{e}^{-u} \ ,$$

we get

$$\begin{aligned}\mathrm{e}^u \int_u^{+\infty} g(\theta, y, p, q)\mathrm{e}^{-y}\mathrm{d}y &= \theta(p+q) + (1+u)\frac{2p}{\theta} + (2+2u+u^2)\frac{q}{\theta} \\ &= \left(\theta + \frac{2(1+u)}{\theta}\right) p + \left(\theta + \frac{2+2u+u^2}{\theta}\right) q \ .\end{aligned}$$

■

We can now proceed with the proof of Theorem 17. Let $\theta \in (0,1]$ be fixed. Let $(\widehat{f}_T^{\mathrm{ho}})_{T\in\mathcal{T}}$ be the individual hold out estimators, so that $\widehat{f}_{\mathcal{T}}^{\mathrm{ag}} = \frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}} \widehat{f}_T^{\mathrm{ho}}$. By convexity of the risk functional $\mathcal{L}$, we have

$$\mathcal{L}(\widehat{f}_{\mathcal{T}}^{\mathrm{ag}}) \leqslant \frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}} \mathcal{L}(\widehat{f}_T^{\mathrm{ho}}) \ .$$

It follows by subtracting $\mathcal{L}(s)$ that

$$\ell(s, \widehat{f}_{\mathcal{T}}^{\mathrm{ag}}) \leqslant \frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}} \ell(s, \widehat{f}_T^{\mathrm{ho}}) \ .$$

Since the data are independent and identically distributed, by assumption (2), all $\widehat{f}_T^{\mathrm{ho}}$ have the same distribution. Let $T_1 = \{1, \ldots, n_t\}$, so that $D_n^{T_1} = D_{n_t}$. Taking expectations yields

$$\mathbb{E}\big[\ell(s, \widehat{f}_{\mathcal{T}}^{\mathrm{ag}})\big] \leqslant \mathbb{E}\big[\ell(s, \widehat{f}_{T_1}^{\mathrm{ho}})\big] \ . \tag{20}$$

Since $H\left(\widehat{w}_{1,1}, \widehat{w}_{1,2}, (\mathcal{A}_m(D_{n_t}))_{m\in\mathcal{M}}\right)$ holds, we can apply Theorem 16 conditionally on D_{n_t}, with $t_m = \mathcal{A}_m(D_{n_t})$.

A.3.1 Proof of Equation (9)

For $i \in \{1,2\}$, let $\widehat{\delta}_{1,i} = \delta(\widehat{w}_{1,i}, \sqrt{n_v}^{\,i})$. Let g be given by Definition 22. By Theorem 16, Equation (8), for any $z \geqslant 0$, with probability greater than $1 - \mathrm{e}^{-z}$,

$$(1-\theta)\ell(s, \widehat{f}^{\mathrm{ho}}_{T_1}) \leqslant (1+\theta) \min_{m \in \mathcal{M}} \big\{\ell(s,t_m)\big\} + g\big(\theta, z + \log|\mathcal{M}|, \widehat{\delta}^2_{1,1}, \widehat{\delta}^2_{1,2}\big) \ . \tag{21}$$

As g is nondecreasing in its second variable, Lemma 21 applied to the random variable $(1-\theta)\ell(s, \widehat{f}^{\mathrm{ho}}_{T_1})$ yields

$$(1-\theta)\mathbb{E}\left[\ell(s, \widehat{f}^{\mathrm{ho}}_{T_1}) \middle| D_n^{T_1}\right] \leqslant (1+\theta) \min_{m \in \mathcal{M}} \big\{\ell(s,t_m)\big\} + \int_{\log|\mathcal{M}|}^{+\infty} g\big(\theta, y, \widehat{\delta}^2_{1,1}, \widehat{\delta}^2_{1,2}\big)\mathrm{e}^{-(y - \log|\mathcal{M}|)}\mathrm{d}y \ .$$

Lemma 23 yields

$$\begin{aligned}(1-\theta)\mathbb{E}\left[\ell(s, \widehat{f}^{\mathrm{ho}}_{T_1}) \middle| D_n^{T_1}\right] \leqslant (1+\theta) \min_{m \in \mathcal{M}} \big\{\ell(s,t_m)\big\} &+ \left(\theta + \frac{2\,(1+\log|\mathcal{M}|)}{\theta}\right)\widehat{\delta}^2_{1,1} \\ &+ \left(\theta + \frac{2\,(1+\log|\mathcal{M}|) + \log^2|\mathcal{M}|}{\theta}\right)\widehat{\delta}^2_{1,2} \ .\end{aligned}$$

Taking expectations with respect to $D_n^{T_1} = D_{n_t}$, we get

$$\begin{aligned}(1-\theta)\mathbb{E}\left[\ell(s, \widehat{f}^{\mathrm{ho}}_{T_1})\right] \leqslant (1+\theta)\mathbb{E}\left[\min_{m \in \mathcal{M}} \ell(s, \mathcal{A}_m(D_{n_t}))\right] &+ \left(\theta + \frac{2\,(1+\log|\mathcal{M}|)}{\theta}\right)\mathbb{E}\left[\widehat{\delta}^2_{1,1}\right] \\ &+ \left(\theta + \frac{2\,(1+\log|\mathcal{M}|) + \log^2|\mathcal{M}|}{\theta}\right)\mathbb{E}\left[\widehat{\delta}^2_{1,2}\right] \ .\end{aligned}$$

Equation (9) then follows from Equation (20).

A.3.2 Proof of Equation (10)

Fix $x \geqslant 0$. For $i \in \{1,2\}$, let $\widehat{\delta}_{2,i} = \delta\left(\widehat{w}_{2,i}, \left(\frac{\theta}{2}\sqrt{\frac{n_v}{x+\log|\mathcal{M}|}}\right)^i\right)$.

By Theorem 16, Equation (7), with probability larger than $1 - \mathrm{e}^{-x}$,

$$(1-\theta)\ell(s, \widehat{f}^{\mathrm{ho}}_{T_1}) \leqslant (1+\theta) \min_{m \in \mathcal{M}} \big\{\ell(s,t_m)\big\} + \sqrt{2}\theta\widehat{\delta}^2_{2,1} + \frac{\theta^2}{2}\widehat{\delta}^2_{2,2} \ . \tag{22}$$

Combining Equations (21) and (22), for any $z \geqslant 0$, with probability larger than $1 - \mathrm{e}^{-z}$,

$$\begin{aligned}(1-\theta)\ell(s, \widehat{f}^{\mathrm{ho}}_{T_1}) \leqslant (1+\theta) &\min_{m \in \mathcal{M}} \big\{\ell(s,t_m)\big\} \\ &+ \sqrt{2}\theta\widehat{\delta}^2_{2,1} + \frac{\theta^2}{2}\widehat{\delta}^2_{2,2} + \mathbb{I}_{z \geqslant x} g\big(\theta, z + \log|\mathcal{M}|, \widehat{\delta}^2_{1,1}, \widehat{\delta}^2_{1,2}\big) \ .\end{aligned}$$

By Lemma 21,

$$\begin{aligned}(1-\theta)\mathbb{E}\left[\ell(s, \widehat{f}^{\mathrm{ho}}_{T_1}) \middle| D_n^{T_1}\right] \leqslant (1+\theta) &\min_{m \in \mathcal{M}} \big\{\ell(s,t_m)\big\} + \sqrt{2}\theta\widehat{\delta}^2_{2,1} + \frac{\theta^2}{2}\widehat{\delta}^2_{2,2} \\ &+ \int_{x+\log|\mathcal{M}|}^{+\infty} g\big(\theta, y, \widehat{\delta}^2_{1,1}, \widehat{\delta}^2_{1,2}\big)\mathrm{e}^{-(y - \log|\mathcal{M}|)}\mathrm{d}y \ .\end{aligned}$$

By Lemma 23, it follows that

$$(1-\theta)\mathbb{E}\left[\ell(s,\widehat{f}^{\mathrm{ho}}_{T_1})\middle| D_n^{T_1}\right] \leqslant (1+\theta)\min_{m\in\mathcal{M}}\left\{\ell(s,t_m)\right\} + \sqrt{2}\theta\widehat{\delta}^2_{2,1} + \frac{\theta^2}{2}\widehat{\delta}^2_{2,2} + \mathrm{e}^{-x}\left(\theta + \frac{2(1+x+\log|\mathcal{M}|)}{\theta}\right)\widehat{\delta}^2_{1,1} + \mathrm{e}^{-x}\left(\theta + \frac{2(1+x+\log|\mathcal{M}|)+(x+\log|\mathcal{M}|)^2}{\theta}\right)\widehat{\delta}^2_{1,2} \ .$$

Taking expectations with respect to $D_n^{T_1}$ and using inequality (20) yields Equation (10) of Theorem 17. ■

Appendix B. RKHS Regression: Proof of Theorem 11

In the following, for any $g : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ and $t : \mathcal{X}\to\mathbb{R}$, the function $(x,y)\mapsto g(t(x),y)$ is denoted by $g\circ t$.

B.1 Preliminary Results

Remark first that the RKHS norm dominates the supremum norm.

Lemma 24 *If $\kappa = \sup_{x\in\mathcal{X}} K(x,x) < +\infty$ then for any $t\in\mathcal{H}$,*

$$\|t\|_\infty \leqslant \sqrt{\kappa}\,\|t\|_{\mathcal{H}} \ .$$

Proof By definition of an RKHS, $\forall t\in\mathcal{H}, \forall x\in\mathcal{X}, \langle t, K(x,\cdot)\rangle_{\mathcal{H}} = t(x)$. It follows that, for any $t\in\mathcal{H}$,

$$\begin{aligned}\|t\|^2_\infty = \sup_{x\in\mathcal{X}} t(x)^2 &= \sup_{x\in\mathcal{X}}\langle t, K(x,\cdot)\rangle^2_{\mathcal{H}} \\ &\leqslant \|t\|^2_{\mathcal{H}}\sup_{x\in\mathcal{X}}\langle K(x,\cdot), K(x,\cdot)\rangle \\ &\leqslant \|t\|^2_{\mathcal{H}}\sup_{x\in\mathcal{X}} K(x,x) \ .\end{aligned}$$

■

Using standard arguments, the following deviation inequality can be derived.

Proposition 25 *Let $\mathcal{H}$ denote a RKHS with bounded kernel $K : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$. Let $\kappa = \sup_{x\in\mathcal{X}} K(x,x)$ and $h : \mathbb{R}^2\to\mathbb{R}$ be Lipschitz in its first argument with Lipschitz constant L. For any $t\in\mathcal{H}$ and $r>0$, denote*

$$B_{\mathcal{H}}(t,r) = \left\{t'\in\mathcal{H} \ : \ \left\|t'-t\right\|_{\mathcal{H}} \leqslant r\right\} \ .$$

Let $t_0\in\mathcal{H}$. Then for any probability measure P on $\mathcal{X}\times\mathbb{R}$ and any $y>0$,

$$P^{\otimes n}\left[\sup_{(t_1,t_2)\in B_{\mathcal{H}}(t_0,r)^2}(P_n-P)\big(h\circ t_1 - h\circ t_2\big) \geqslant 2(2+\sqrt{2y})L\frac{r\sqrt{\kappa}}{\sqrt{n}}\right] \leqslant \mathrm{e}^{-y} \ .$$

Proof Let $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n}$ be a data set drawn from P. Let $(\sigma_i)_{1 \leqslant i \leqslant n}$ be independent Rademacher variables independent from D_n. Denote by

$$R_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i)\right]$$

the Rademacher complexity of a class $\mathcal{F}$ of real valued functions.

By Lemma 24, for any $(t_1, t_2) \in B_{\mathcal{H}}(t_0, r)^2$,

$$\|h \circ t_1 - h \circ t_2\|_\infty \leqslant L \|t_1 - t_2\|_\infty \leqslant L\left[\|t_1 - t_0\|_\infty + \|t_2 - t_0\|_\infty\right] \leqslant 2L\sqrt{\kappa} r \ .$$

By symmetry under exchange of t_1 and t_2, notice that

$$R_n\left(\{h \circ t_1 - h \circ t_2 \,:\, (t_1, t_2) \in B_{\mathcal{H}}(t_0, r)^2\}\right) = \sup_{(t_1, t_2) \in B_{\mathcal{H}}(t_0, r)^2} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i (h \circ t_1 - h \circ t_2)(X_i) \right| \ .$$

By the bounded difference inequality and Boucheron et al. (2005, Theorem 3.2), it follows that for any $y > 0$, with probability greater than $1 - \mathrm{e}^{-y}$,

$$\begin{aligned} &\sup_{(t_1, t_2) \in B_{\mathcal{H}}(t_0, r)^2} (P_n - P)(h \circ t_1 - h \circ t_2) \\ &\qquad \leqslant 2R_n\left(\{h \circ t_1 - h \circ t_2 \,:\, (t_1, t_2) \in B_{\mathcal{H}}(t_0, r)^2\}\right) + 2Lr\sqrt{\frac{2\kappa y}{n}} \ . \end{aligned}$$

Moreover,

$$\begin{aligned} &R_n\left(\{h \circ t_1 - h \circ t_2 \,:\, (t_1, t_2) \in B_{\mathcal{H}}(t_0, r)^2\}\right) \\ &\leqslant R_n\Big(\big\{h \circ t \,:\, t \in B_{\mathcal{H}}(t_0, r)\big\}\Big) + R_n\Big(\big\{-h \circ t \,:\, t \in B_{\mathcal{H}}(t_0, r)\big\}\Big) \\ &\leqslant 2LR_n\big(B_{\mathcal{H}}(t_0, r)\big) \qquad \text{by the contraction lemma} \\ &\qquad\qquad\qquad\qquad \text{as formulated by Meir and Zhang (2003, Theorem 7),} \\ &= 2LR_n\big(B_{\mathcal{H}}(0, r)\big) \qquad \text{by translation invariance.} \end{aligned}$$

Finally, by a classical computation (see for example Boucheron et al., 2005, Section 4.1.2),

$$\begin{aligned} &R_n\Big(\big\{h \circ t_1 - h \circ t_2 \,:\, (t_1, t_2) \in B_{\mathcal{H}}(t_0, r)^2\big\}\Big) \\ &\leqslant 2L\frac{r}{n}\mathbb{E}\sqrt{\sum_{i=1}^{n} K(X_i, X_i)} \\ &\leqslant 2Lr\sqrt{\frac{\kappa}{n}} \ . \end{aligned}$$

■

The proof of Theorem 11 also uses the following peeling lemma.

Lemma 26 *Let $(Z_u)_{u\in T}$ be a stochastic process and $d : T \to \mathbb{R}_+$ be a function. Let $a \geqslant 0$ and $b \in (0, 2]$ and assume that*

$$\forall r, y \geqslant 0\,, \qquad \mathbb{P}\left(\sup_{u\in T: d(u)\leqslant r} Z_u \geqslant r\frac{1+\sqrt{b(a+y)}}{\sqrt{n}}\right) \leqslant \mathrm{e}^{-y}\ . \tag{23}$$

Then, for any $\theta \in (0, +\infty)$,

$$\mathbb{P}\left(\exists u \in T\,,\ Z_u \geqslant \theta d^2(u) + \frac{2 + b\big[1.1 + 2(a+y)\big]}{\theta n}\right) \leqslant \mathrm{e}^{-y}\ .$$

Proof Let $x > 0$. Let $\eta \in (1, 2]$, $j_m \in \mathbb{N}^*$ and $y_0 \in \mathbb{R}$ be numerical constants that will be determined later. Then,

$$\begin{aligned}
&\mathbb{I}\left\{\sup_{u\in T}\frac{Z_u}{d^2(u)+x^2} \geqslant \frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right\}\\
&\leqslant \mathbb{I}\left\{\sup_{u\in T: d(u)\leqslant x}\frac{Z_u}{d^2(u)+x^2} \geqslant \frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right\}\\
&\quad + \sum_{j=0}^{+\infty}\mathbb{I}\left\{\sup_{u\in T:\eta^j x\leqslant d(u)\leqslant \eta^{j+1}x}\frac{Z_u}{d^2(u)+x^2} \geqslant \frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right\}\\
&\leqslant \mathbb{I}\left\{\sup_{u\in T: d(u)\leqslant x}\frac{Z_u}{x^2} \geqslant \frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right\}\\
&\quad + \sum_{j=0}^{+\infty}\mathbb{I}\left\{\sup_{u\in T:\eta^j x\leqslant d(u)\leqslant \eta^{j+1}x}\frac{Z_u}{(1+\eta^{2j})x^2} \geqslant \frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right\}\\
&\leqslant \mathbb{I}\left\{\sup_{u\in T: d(u)\leqslant x} Z_u \geqslant \frac{x\big(1+\sqrt{b(a+y)}\big)}{\sqrt{n}}\right\}\\
&\quad + \sum_{j=0}^{+\infty}\mathbb{I}\left\{\sup_{u\in T: d(u)\leqslant \eta^{j+1}x} Z_u \geqslant (1+\eta^{2j})\frac{x\big(1+\sqrt{b(a+y)}\big)}{\sqrt{n}}\right\}\ .
\end{aligned} \tag{24}$$

Notice that

$$\begin{aligned}
(1+\eta^{2j})\frac{x\big(1+\sqrt{b(a+y)}\big)}{\sqrt{n}} &= x\eta^{j+1}\times\frac{\eta^{2j}+1}{\eta^{j+1}}\times\frac{1+\sqrt{b(a+y)}}{\sqrt{n}}\\
&= x\eta^{j+1}\frac{1+\sqrt{b(a+z_j)}}{\sqrt{n}}\ ,
\end{aligned}$$

where

$$\begin{aligned}
z_j &= \frac{1}{b}\left(\frac{\eta^{2j}+1}{\eta^{j+1}} - 1 + \frac{\eta^{2j}+1}{\eta^{j+1}}\sqrt{b(a+y)}\right)^2 - a\\
&\geqslant \frac{1}{b}\left(\frac{\eta^{2j}+1}{\eta^{j+1}} - 1\right)^2 + \left(\frac{\eta^{2j}+1}{\eta^{j+1}}\right)^2 y \qquad \text{since } a \geqslant 0 \text{ and } \eta^{2j}+1 \geqslant \eta^{j+1}\ .
\end{aligned}$$

Taking expectations in Equation (24) and using hypothesis (23), we obtain

$$\mathbb{P}\left(\sup_{u\in T}\frac{Z_u}{d^2(u)+x^2}\geqslant\frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right)\leqslant \mathrm{e}^{-y}+\sum_{j=0}^{+\infty}\mathrm{e}^{-z_j}\ .$$

So for any $y\geqslant y_0$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{u\in T}\frac{Z_u}{d^2(u)+x^2}\geqslant\frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right)\\
&\quad\leqslant \mathrm{e}^{-y}+\mathrm{e}^{-y}\sum_{j=0}^{+\infty}\exp\left[-\frac{1}{b}\left(\frac{\eta^{2j}+1}{\eta^{j+1}}-1\right)^2-\left(\frac{(\eta^{2j}+1)^2}{(\eta^{j+1})^2}-1\right)y\right]\\
&\quad\leqslant \mathrm{e}^{-y}+\mathrm{e}^{-y}\sum_{j=0}^{+\infty}\exp\left[-\frac{1}{b}\left(\frac{\eta^{2j}+1}{\eta^{j+1}}-1\right)^2-\left(\frac{(\eta^{2j}+1)^2}{(\eta^{j+1})^2}-1\right)y_0\right]\ . \qquad (25)
\end{aligned}$$

Now, we have

$$\begin{aligned}
\exp\left[-\frac{1}{b}\left(\frac{\eta^{2j}+1}{\eta^{j+1}}-1\right)^2-\left(\frac{(\eta^{2j}+1)^2}{(\eta^{j+1})^2}-1\right)y_0\right]&\leqslant\exp\left[-\left(\frac{(\eta^{2j}+1)^2}{(\eta^{j+1})^2}-1\right)y_0\right]\\
&\leqslant\exp\left(y_0-\eta^{2(j-1)}y_0\right)\ . \qquad (26)
\end{aligned}$$

Let v denote the sequence $v_j=\exp\left(y_0-\eta^{2(j-1)}y_0\right)$. Then for $j\geqslant j_m$,

$$\begin{aligned}
\log v_{j+1}-\log v_j&=\eta^{2(j-1)}y_0-\eta^{2j}y_0\\
&=y_0(1-\eta^2)\eta^{2(j-1)}\\
&\leqslant y_0(1-\eta^2)\eta^{2(j_m-1)}\qquad\text{since }\eta>1\ .
\end{aligned}$$

Thus,

$$\forall j\geqslant j_m,\qquad v_{j+1}\leqslant v_j\exp\left(-y_0(\eta^2-1)\eta^{2(j_m-1)}\right)\ .$$

Therefore, we have

$$\forall j\geqslant 0,\qquad v_{j+j_m}\leqslant v_{j_m}\exp\left(-jy_0(\eta^2-1)\eta^{2(j_m-1))}\right)$$

and

$$\sum_{j=j_m}^{+\infty}v_j\leqslant v_{j_m}\left[1-\exp\left(-y_0(\eta^2-1)\eta^{2(j_m-1)}\right)\right]^{-1}\ .$$

It follows from Equations (25) and (26) that for any $y\geqslant y_0$, since $b\leqslant 2$,

$$\begin{aligned}
&\mathrm{e}^y\mathbb{P}\left(\sup_{u\in T}\frac{Z_u}{d^2(u)+x^2}\geqslant\frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right)\\
&\quad\leqslant 1+\sum_{j=0}^{j_m}\exp\left[-\frac{1}{2}\left(\frac{\eta^{2j}+1}{\eta^{j+1}}-1\right)^2-\left(\frac{(\eta^{2j}+1)^2}{(\eta^{j+1})^2}-1\right)y_0\right] \qquad (27)\\
&\qquad+\frac{\exp\left(y_0-\eta^{2(j_m-1)}y_0\right)}{1-\exp\left(-y_0(\eta^2-1)\eta^{2(j_m-1)}\right)}\ .
\end{aligned}$$

On the other hand, when $y \leqslant y_0$,

$$\mathbb{P}\left(\sup_{u\in T} \frac{Z_u}{d^2(u)+x^2} \geqslant \frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right) \leqslant 1 \leqslant \mathrm{e}^{y_0}\mathrm{e}^{-y} \ .$$

Taking $\eta = 1.18, j_m = 10, y_0 = 0.52$, the right-hand side of Equation (27) evaluates to $1.6765 < 1.7$ whereas $\mathrm{e}^{y_0} \leqslant 1.683 < 1.7$. It follows that for all $y > 0$,

$$\mathbb{P}\left(\sup_{u\in T} \frac{Z_u}{d^2(u)+x^2} \geqslant \frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right) \leqslant 1.7\mathrm{e}^{-y} \ . \tag{28}$$

Now take $x = \frac{1+\sqrt{b(a+y)}}{\theta\sqrt{n}}$ with $\theta > 0$. We can rewrite

$$\begin{aligned}
&\mathbb{P}\left(\sup_{u\in T} \frac{Z_u}{d^2(u)+x^2} \geqslant \frac{1+\sqrt{b(a+y)}}{x\sqrt{n}}\right) \\
&= \mathbb{P}\left(\exists u \in T\,,\ \frac{Z_u}{d^2(u)+x^2} \geqslant \theta\right) \\
&= \mathbb{P}\left(\exists u \in T\,,\ Z_u \geqslant \theta d^2(u) + \frac{1}{\theta n}\left[1+\sqrt{b(a+y)}\right]^2\right) \\
&\geqslant \mathbb{P}\left(\exists u \in T\,,\ Z_u \geqslant \theta d^2(u) + \frac{2+2b(a+y)}{\theta n}\right) \ .
\end{aligned}$$

It follows from Equation (28), with y replaced by $y + 0.55$, that

$$\mathbb{P}\left(\exists u \in T, Z_u \geqslant \theta d^2(u) + \frac{2+b\big[1.1+2(a+y)\big]}{\theta n}\right) \leqslant 1.7\mathrm{e}^{-0.55}\mathrm{e}^{-y} \leqslant \mathrm{e}^{-y} \ .$$

■

We need two other technical lemmas in the proof of Theorem 11.

Lemma 27 *For any nonnegative, continuous convex function h over a Hilbert space $\mathcal{H}$, and any $\lambda \in \mathbb{R}_+$, the elements of the regularization path,*

$$t_\lambda = \operatorname*{argmin}_{t\in\mathcal{H}} \left\{h(t) + \lambda \|t\|_{\mathcal{H}}^2\right\} \ ,$$

satisfy, for any $(\lambda, \mu) \in \mathbb{R}^2$ such that $0 < \lambda \leqslant \mu$,

$$\|t_\lambda - t_\mu\|_{\mathcal{H}}^2 \leqslant \|t_\lambda\|_{\mathcal{H}}^2 - \|t_\mu\|_{\mathcal{H}}^2 \ .$$

Proof By Barbu and Precupanu (2012, Theorem 2.11), t_λ exists for any $\lambda \in \mathbb{R}_+$. Moreover, it is unique by strong convexity of $\|\cdot\|_{\mathcal{H}}^2$. For a closed convex set $\mathcal{C} \subset \mathcal{H}$, let $\Pi_{\mathcal{C}}$ denote the orthogonal projection onto $\mathcal{C}$.

Let $\mu > 0$. The set $\{t : h(t) \leqslant h(t_\mu)\}$ is closed by continuity of h and convex by convexity of h. Moreover, for any $t \in \mathcal{H}$ such that $h(t) \leqslant h(t_\mu)$,

$$\begin{aligned}\mu \|t_\mu\|_{\mathcal{H}}^2 &\leqslant h(t_\mu) - h(t) + \mu \|t_\mu\|_{\mathcal{H}}^2 \\ &\leqslant \mu \|t\|_{\mathcal{H}}^2 \qquad \text{by definition of } t_\mu \ .\end{aligned}$$

Therefore, $t_\mu = \Pi_{\{t:h(t)\leqslant h(t_\mu)\}}(0)$. Let $\lambda \in (0, \mu)$. By definition of t_λ, t_μ,

$$\begin{aligned}\frac{h(t_\mu)}{\mu} + \|t_\mu\|_{\mathcal{H}}^2 &\leqslant \frac{h(t_\lambda)}{\mu} + \|t_\lambda\|_{\mathcal{H}}^2 \\ &= \frac{h(t_\lambda)}{\lambda} + \|t_\lambda\|_{\mathcal{H}}^2 + \left(\frac{1}{\mu} - \frac{1}{\lambda}\right) h(t_\lambda) \\ &\leqslant \frac{h(t_\mu)}{\lambda} + \|t_\mu\|_{\mathcal{H}}^2 + \left(\frac{1}{\mu} - \frac{1}{\lambda}\right) h(t_\lambda) \ ,\end{aligned}$$

which implies $(\mu^{-1} - \lambda^{-1})h(t_\mu) \leqslant (\mu^{-1} - \lambda^{-1})h(t_\lambda)$ and thus $h(t_\lambda) \leqslant h(t_\mu)$ since $\lambda < \mu$. For a projection $\Pi_{\mathcal{C}}$, it is well known that

$$\forall t \in \mathcal{H} \, , \, \forall t' \in \mathcal{C} \, , \qquad \langle t - \Pi_{\mathcal{C}}(t), \Pi_{\mathcal{C}}(t) - t' \rangle_{\mathcal{H}} \geqslant 0 \ .$$

Choosing $\mathcal{C} = \{t : h(t) \leqslant h(t_\mu)\}, t' = t_\lambda \in \mathcal{C}, t = 0$ yields $\langle -t_\mu, t_\mu - t_\lambda \rangle_{\mathcal{H}} \geqslant 0$. Therefore

$$\begin{aligned}\|t_\lambda\|_{\mathcal{H}}^2 &= \|t_\mu + (t_\lambda - t_\mu)\|_{\mathcal{H}}^2 \\ &= \|t_\mu\|_{\mathcal{H}}^2 + \|t_\lambda - t_\mu\|_{\mathcal{H}}^2 + 2\langle t_\mu, t_\lambda - t_\mu \rangle_{\mathcal{H}} \\ &\geqslant \|t_\mu\|_{\mathcal{H}}^2 + \|t_\lambda - t_\mu\|_{\mathcal{H}}^2 \ .\end{aligned}$$

■

Lemma 28 *Let $(b, c) \in \mathbb{R}_+^2$ and $l_{b,c}(x) = bx + c$. Let δ be given by Definition 14. For any $r \in \mathbb{R}_+$,*

$$\delta^2(l_{b,c}, r) \leqslant \frac{b^2}{r^2} + \frac{2c}{r} \ . \tag{29}$$

For $(a, b, c) \in \mathbb{R}_+^3$, let $g_{a,b,c}(x) = ax \vee (bx^3 + cx^2)^{\frac{1}{2}}$. For any $r \in \mathbb{R}_+$,

$$\delta^2\big(g_{a,b,c}, r\big) \leqslant \frac{a^2}{r^2} \vee \left(\frac{b^2}{r^4} + \frac{2c}{r^2}\right) \leqslant \frac{a^2}{r^2} + \frac{b^2}{r^4} + \frac{2c}{r^2} \ . \tag{30}$$

Proof Since $x \mapsto \frac{l_{b,c}(x)}{x}$ is nonincreasing, we have by Remark 15

$$\begin{aligned} b\delta(l_{b,c}, r) + c &= r\delta^2(l_{b,c}, r) \ , \\ \text{that is,} \qquad \delta^2(l_{b,c}, r) - \frac{b\delta(l_{b,c}, r)}{r} - \frac{c}{r} &= 0 \ ,\end{aligned}$$

hence $\delta(l_{b,c},r) = \frac{b}{2r} + \frac{1}{2}\sqrt{\frac{b^2}{r^2} + \frac{4c}{r}}$. Thus, we have

$$\delta^2(l_{b,c},r) \leqslant 2\left(\frac{b^2}{4r^2} + \frac{b^2}{4r^2} + \frac{c}{r}\right) \leqslant \frac{b^2}{r^2} + \frac{2c}{r} \ .$$

This proves Equation (29). For any $x > 0$, $g_{a,b,c}(x) \leqslant rx^2$ is equivalent to

$$ax \leqslant rx^2 \tag{31}$$

$$\text{and} \qquad bx^3 + cx^2 \leqslant r^2x^4 \ . \tag{32}$$

Equation (31) is equivalent to $x \geqslant \frac{a}{r}$. On the other hand, for every

$$x > \left(\frac{b^2}{r^4} + \frac{2c}{r^2}\right)^{\frac{1}{2}} \ ,$$

we have $x > \delta(l_{b,c},r^2)$ by Equation (29), hence $bx + c \leqslant r^2x^2$ by Definition 14, so that Equation (32) holds true. Therefore, whenever

$$x > \frac{a}{r} \vee \left(\frac{b^2}{r^4} + \frac{2c}{r^2}\right)^{\frac{1}{2}} \ ,$$

it holds that $g_{a,b,c}(x) \leqslant rx^2$. Equation (30) follows by Definition 14. ■

B.2 Uniform Control on the Empirical Process

From now on, until the end of the proof, the notation and hypotheses of Theorem 11 are used. Recall also the notation $g \circ t : (x,y) \mapsto g(t(x),y)$, for any $g : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ and $t : \mathcal{X} \to \mathbb{R}$. Fix a training set D_{n_t}. Start with the following definition.

Definition 29 *For $t_1, t_2 \in \mathcal{H}$, let*

$$d(t_1,t_2) = \min_{\lambda\in\Lambda}\left\{\|t_1 - s_\lambda\|_{\mathcal{H}}\right\} + \|t_1 - t_2\|_{\mathcal{H}} \ , \tag{33}$$

where $s_\lambda = \operatorname{argmin}_{t\in\mathcal{H}}\left\{P(c\circ t) + \lambda\|t\|_{\mathcal{H}}^2\right\}$. Furthermore, let

$$\widehat{y} = \frac{\lambda_m n_t}{32\kappa L^2} \times \sup_{(t_1,t_2)\in\mathcal{H}^2}\left\{(P_{n_t} - P)(c\circ t_1 - c\circ t_2) - \frac{\lambda_m}{2}d(t_1,t_2)^2\right\} \ ,$$

so that

$$\forall (t_1,t_2) \in \mathcal{H}^2, \qquad (P_{n_t} - P)(c\circ t_1 - c\circ t_2) \leqslant \frac{\lambda_m}{2}d(t_1,t_2)^2 + \frac{32\kappa L^2\widehat{y}}{\lambda_m n_t} \ . \tag{34}$$

We then have the following bounds on $\widehat{y}$.

Claim 30 *For all $x \geqslant 0$,*

$$\mathbb{P}\big(\widehat{y} \geqslant 2.6 + \log|\Lambda| + x\big) \leqslant \mathrm{e}^{-x} \ .$$

In particular, $\mathbb{E}[\widehat{y}] \leqslant 4 + \log|\Lambda|$.

Proof Let $t_1, t_2 \in \mathcal{H}$ be such that $d(t_1, t_2) \leqslant r$. Let $\lambda \in \Lambda$ be such that

$$\|t_1 - s_\lambda\|_{\mathcal{H}} + \|t_1 - t_2\|_{\mathcal{H}} \leqslant r \ .$$

By the triangle inequality, $t_1, t_2 \in B(s_\lambda, r)$, hence

$$\sup_{(t_1,t_2):d(t_1,t_2)\leqslant r} \big\{(P_{n_t} - P)(c\circ t_1 - c\circ t_2)\big\} \leqslant \max_{\lambda\in\Lambda} \sup_{(t_1,t_2)\in B(s_\lambda,r)^2} (P_{n_t} - P)(c\circ t_1 - c\circ t_2) \ . \quad (35)$$

From Proposition 25 and the union bound, it follows that, for any $x \geqslant 0$,

$$\mathbb{P}\left(\max_{\lambda\in\Lambda} \sup_{(t_1,t_2)\in B(s_\lambda,r)^2} \big\{(P_{n_t} - P)(c\circ t_1 - c\circ t_2)\big\} \geqslant 2\left(2 + \sqrt{2(x + \log|\Lambda|)}\right) L\frac{r\sqrt{\kappa}}{\sqrt{n_t}}\right) \leqslant \mathrm{e}^{-x} \ .$$

It follows by Equation (35) that, for all $x \geqslant 0$,

$$\mathbb{P}\left(\sup_{(t_1,t_2):d(t_1,t_2)\leqslant r} \frac{1}{4L\sqrt{\kappa}}\big\{(P_{n_t} - P)(c\circ t_1 - c\circ t_2)\big\} \geqslant \left(1 + \sqrt{\frac{x + \log|\Lambda|}{2}}\right)\frac{r}{\sqrt{n_t}}\right) \leqslant \mathrm{e}^{-x} \ .$$

By Lemma 26 with $\theta = \frac{\lambda_m}{8L\sqrt{\kappa}}$, $a = \log|\Lambda|$, $b = \frac{1}{2}$, with probability larger than $1 - \mathrm{e}^{-x}$,

$$\forall (t_1, t_2) \in \mathcal{H}^2 \,, \qquad (P_{n_t} - P)(c\circ t_1 - c\circ t_2) \leqslant \frac{\lambda_m}{2} d(t_1, t_2)^2 + 32L^2\frac{\kappa(2.6 + x + \log|\Lambda|)}{\lambda_m n_t} \ .$$

On the same event, $\widehat{y} \leqslant 2.6 + x + \log|\Lambda|$ by Definition 29. Therefore, by Lemma 21, $\mathbb{E}[\widehat{y}] \leqslant 3.6 + \log|\Lambda|$. ∎

Definition 29 and Claim 30 together imply a uniform control on the empirical process thanks to the drift term $\lambda_m d(t_1, t_2)^2$, whereas Proposition 25 only gives a bound on an RKHS ball of fixed radius.

B.3 Verifying the Assumptions of Theorem 17

Theorem 11 is a consequence of Theorem 17. For all $\lambda \in \Lambda$, let $\widehat{t}_\lambda = \mathcal{A}_\lambda(D_{n_t})$, where $\mathcal{A}_\lambda$ is given by Definition 9. To verify the assumptions of Theorem 17, adequate functions $(\widehat{w}_{i,j})_{(i,j)\in\{1,2\}^2}$ must be found such that for $i \in \{1, 2\}$, $H\left(\widehat{w}_{i,1}, \widehat{w}_{i,2}, (\widehat{t}_\lambda)_{\lambda\in\Lambda}\right)$ holds almost surely. This is the purpose of the present subsection.

The core of the proof of Theorem 11 lies in the following deterministic claim.

Claim 31 *For all $\lambda, \mu \in \Lambda$ such that $\lambda \leqslant \mu$, we have*

$$\big\|\widehat{t}_\lambda - \widehat{t}_\mu\big\|_\infty^2 \leqslant \frac{\kappa C}{\lambda_m}\ell(s, \widehat{t}_\mu) + 96L^2\frac{\kappa^2\widehat{y}}{\lambda_m^2 n_t} \ .$$

Proof Let $(\lambda,\mu)\in\Lambda^2$ with $\lambda\leqslant\mu$. Let s_μ be as in Definition 29. By convexity of c, the function $t\mapsto P(c\circ t)+\mu\left\|t\right\|_{\mathcal{H}}^2$ is μ-strongly convex. Since s_μ is its optimum, we get

$$\forall t\in\mathcal{H}\,,\qquad P(c\circ t)+\mu\left\|t\right\|_{\mathcal{H}}^2\geqslant P(c\circ s_\mu)+\mu\left\|s_\mu\right\|_{\mathcal{H}}^2+\mu\left\|t-s_\mu\right\|_{\mathcal{H}}^2\ .$$

Hence, taking $t=\widehat{t}_\mu$,

$$\begin{aligned}
&\lambda_m\left\|\widehat{t}_\mu-s_\mu\right\|_{\mathcal{H}}^2\\
&\leqslant\mu\left\|\widehat{t}_\mu-s_\mu\right\|_{\mathcal{H}}^2\\
&\leqslant P(c\circ\widehat{t}_\mu)+\mu\left\|\widehat{t}_\mu\right\|_{\mathcal{H}}^2-P(c\circ s_\mu)-\mu\left\|s_\mu\right\|_{\mathcal{H}}^2\\
&=P_{n_t}(c\circ\widehat{t}_\mu)+\mu\left\|\widehat{t}_\mu\right\|_{\mathcal{H}}^2-P_{n_t}(c\circ s_\mu)-\mu\left\|s_\mu\right\|_{\mathcal{H}}^2+(P-P_{n_t})(c\circ\widehat{t}_\mu-c\circ s_\mu)\ .
\end{aligned}$$

By Definition 9,

$$P_{n_t}(c\circ\widehat{t}_\mu)+\mu\left\|\widehat{t}_\mu\right\|_{\mathcal{H}}^2\leqslant P_{n_t}(c\circ s_\mu)+\mu\left\|s_\mu\right\|_{\mathcal{H}}^2\ .$$

Hence $\lambda_m\left\|\widehat{t}_\mu-s_\mu\right\|_{\mathcal{H}}^2\leqslant(P-P_{n_t})(c\circ\widehat{t}_\mu-c\circ s_\mu)=(P_{n_t}-P)(c\circ s_\mu-c\circ\widehat{t}_\mu)$. Now take $t_1=s_\mu$ and $t_2=\widehat{t}_\mu$ in Equation (34) of Definition 29 to get

$$\begin{aligned}
\lambda_m\left\|\widehat{t}_\mu-s_\mu\right\|_{\mathcal{H}}^2&\leqslant\frac{\lambda_m}{2}d(s_\mu,\widehat{t}_\mu)^2+32L^2\frac{\kappa\widehat{y}}{\lambda_m n_t}\\
&=\frac{\lambda_m}{2}\left\|s_\mu-\widehat{t}_\mu\right\|_{\mathcal{H}}^2+32L^2\frac{\kappa\widehat{y}}{\lambda_m n_t}\ .
\end{aligned}$$

Therefore,

$$\left\|\widehat{t}_\mu-s_\mu\right\|_{\mathcal{H}}^2\leqslant 64L^2\frac{\widehat{y}\kappa}{\lambda_m^2 n_t}\ . \tag{36}$$

Now $\left\|\widehat{t}_\lambda-\widehat{t}_\mu\right\|_{\mathcal{H}}^2$ can be bounded as follows. Since $t\mapsto P_{n_t}(c\circ t)+\lambda\left\|t\right\|_{\mathcal{H}}^2$ is λ-strongly convex and $\widehat{t}_\lambda$ is its optimum,

$$\begin{aligned}
\lambda_m\left\|\widehat{t}_\lambda-\widehat{t}_\mu\right\|_{\mathcal{H}}^2&\leqslant\lambda\left\|\widehat{t}_\lambda-\widehat{t}_\mu\right\|_{\mathcal{H}}^2\\
&\leqslant P_{n_t}(c\circ\widehat{t}_\mu)-P_{n_t}(c\circ\widehat{t}_\lambda)+\lambda\left\|\widehat{t}_\mu\right\|_{\mathcal{H}}^2-\lambda\left\|\widehat{t}_\lambda\right\|_{\mathcal{H}}^2\ .
\end{aligned}$$

By Lemma 27 with $h(t)=P_{n_t}(c\circ t)$, $\left\|\widehat{t}_\lambda-\widehat{t}_\mu\right\|_{\mathcal{H}}^2\leqslant\left\|\widehat{t}_\lambda\right\|_{\mathcal{H}}^2-\left\|\widehat{t}_\mu\right\|_{\mathcal{H}}^2$. Hence

$$\begin{aligned}
(\lambda_m+\lambda)\left\|\widehat{t}_\lambda-\widehat{t}_\mu\right\|_{\mathcal{H}}^2&\leqslant P_{n_t}(c\circ\widehat{t}_\mu)-P_{n_t}(c\circ\widehat{t}_\lambda)\\
&=P(c\circ\widehat{t}_\mu)-P(c\circ\widehat{t}_\lambda)+(P_{n_t}-P)\left(c\circ\widehat{t}_\mu-c\circ\widehat{t}_\lambda\right)\\
&\leqslant P(c\circ\widehat{t}_\mu)-\min_{t\in\mathbb{S}}P\left(c\circ t\right)+(P_{n_t}-P)(c\circ\widehat{t}_\mu-c\circ\widehat{t}_\lambda)\\
&\leqslant C\ell(s,\widehat{t}_\mu)+(P_{n_t}-P)\left(c\circ\widehat{t}_\mu-c\circ\widehat{t}_\lambda\right)
\end{aligned}$$

by hypothesis $Comp_C(g,c)$. By Equation (34) with $t_1=\widehat{t}_\mu$ and $t_2=\widehat{t}_\lambda$, we have

$$\begin{aligned}
(\lambda_m+\lambda)\left\|\widehat{t}_\lambda-\widehat{t}_\mu\right\|_{\mathcal{H}}^2&\leqslant C\ell(s,\widehat{t}_\mu)+\frac{\lambda_m}{2}\left(\left\|\widehat{t}_\mu-s_\mu\right\|_{\mathcal{H}}+\left\|\widehat{t}_\lambda-\widehat{t}_\mu\right\|_{\mathcal{H}}\right)^2+32L^2\frac{\kappa\widehat{y}}{\lambda_m n_t}\\
&\leqslant C\ell(s,\widehat{t}_\mu)+\frac{\lambda_m}{2}\left(8\frac{L\sqrt{\widehat{y}\kappa}}{\lambda_m\sqrt{n_t}}+\left\|\widehat{t}_\lambda-\widehat{t}_\mu\right\|_{\mathcal{H}}\right)^2+32L^2\frac{\kappa\widehat{y}}{\lambda_m n_t}
\end{aligned}$$

by Equation (36). For any $a, b \in \mathbb{R}$, $(a+b)^2 \leqslant 2a^2 + 2b^2$, hence

$$(\lambda + \lambda_m) \left\| \widehat{t}_\lambda - \widehat{t}_\mu \right\|_{\mathcal{H}}^2 \leqslant C\ell(s, \widehat{t}_\mu) + \frac{\lambda_m}{2} \left(128L^2 \frac{\widehat{y}\kappa}{\lambda_m^2 n_t} + 2 \left\| \widehat{t}_\lambda - \widehat{t}_\mu \right\|_{\mathcal{H}}^2 \right) + 32L^2 \frac{\kappa \widehat{y}}{\lambda_m n_t} \ .$$

This yields

$$\lambda \left\| \widehat{t}_\lambda - \widehat{t}_\mu \right\|_{\mathcal{H}}^2 \leqslant C\ell(s, \widehat{t}_\mu) + 96L^2 \frac{\kappa \widehat{y}}{\lambda_m n_t} \ ,$$

and finally, since $\lambda \geqslant \lambda_m$,

$$\left\| \widehat{t}_\lambda - \widehat{t}_\mu \right\|_{\mathcal{H}}^2 \leqslant \frac{C\ell(s, \widehat{t}_\mu)}{\lambda_m} + 96L^2 \frac{\kappa \widehat{y}}{\lambda_m^2 n_t} \ .$$

Now, by Lemma 24,

$$\begin{aligned} \left\| \widehat{t}_\lambda - \widehat{t}_\mu \right\|_\infty^2 &\leqslant \kappa \left\| \widehat{t}_\lambda - \widehat{t}_\mu \right\|_{\mathcal{H}}^2 \\ &\leqslant \frac{\kappa C}{\lambda_m} \ell(s, \widehat{t}_\mu) + 96L^2 \frac{\kappa^2 \widehat{y}}{\lambda_m^2 n_t} \ . \end{aligned}$$

This proves Claim 31. ■

Using hypothesis $SC_{\rho,\nu}$—Equation (4)—, a refined bound can be obtained on

$$P\left[\left(g \circ \widehat{t}_\lambda - g \circ \widehat{t}_\mu \right)^2 \right] \ .$$

Claim 32 *For any* $(\lambda, \mu) \in \Lambda^2$,

$$P\left[\left(g \circ \widehat{t}_\lambda - g \circ \widehat{t}_\mu \right)^2 \right] \leqslant \widehat{w}_B \left(\sqrt{\ell(s, \widehat{t}_\lambda)} \right)^2 + \widehat{w}_B \left(\sqrt{\ell(s, \widehat{t}_\mu)} \right)^2$$

where

$$\widehat{w}_B(x)^2 = \max \left\{ \rho x^2, \ \nu \frac{4}{3} \sqrt{\frac{\kappa C}{\lambda_m}} x^3 + 10\nu L \frac{\kappa \sqrt{\widehat{y}}}{\lambda_m \sqrt{n_t}} x^2 \right\} \ .$$

Proof By hypothesis $SC_{\rho,\nu}$—Equation (4)—with $u = \widehat{t}_\lambda(X)$ and $v = \widehat{t}_\mu(X)$,

$$\begin{aligned} \mathbb{E}\Big[\left(g \circ \widehat{t}_\lambda - g \circ \widehat{t}_\mu \right)^2 (X, Y) \,\Big|\, X \Big] &\leqslant \left[\rho \vee \left(\nu |\widehat{t}_\lambda(X) - \widehat{t}_\mu(X)| \right) \right] \left[\ell_X(\widehat{t}_\lambda(X)) + \ell_X(\widehat{t}_\mu(X)) \right] \\ &\leqslant \left[\rho \vee \left(\nu \left\| \widehat{t}_\lambda - \widehat{t}_\mu \right\|_\infty \right) \right] \left[\ell_X(\widehat{t}_\lambda(X)) + \ell_X(\widehat{t}_\mu(X)) \right] \ , \end{aligned}$$

where $\ell_X(u) = \mathbb{E}[g(u, Y)|X] - \min_{v \in \mathbb{R}} \mathbb{E}[g(v, Y)|X]$. Integrating this inequality with respect to X, it follows that

$$P\left[\left(g \circ \widehat{t}_\lambda - g \circ \widehat{t}_\mu \right)^2 \right] \leqslant \left[\rho \vee \left(\nu \left\| \widehat{t}_\lambda - \widehat{t}_\mu \right\|_\infty \right) \right] \left[\ell(s, \widehat{t}_\lambda) + \ell(s, \widehat{t}_\mu) \right] \ .$$

Assume without loss of generality that $\lambda \leqslant \mu$. By Claim 31,

$$\begin{aligned}
&P\left[\left(g\circ\widehat{t}_\lambda - g\circ\widehat{t}_\mu\right)^2\right] \\
&\leqslant \left(\rho \vee \nu\left[\sqrt{\frac{\kappa C}{\lambda_m}}\sqrt{\ell(s,\widehat{t}_\mu)} + 10\frac{L\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}}\right]\right)\left[\ell(s,\widehat{t}_\lambda)+\ell(s,\widehat{t}_\mu)\right] \\
&\leqslant \max\Bigg\{\rho\Big[\ell(s,\widehat{t}_\lambda)+\ell(s,\widehat{t}_\mu)\Big],\ \nu\Bigg[\sqrt{\frac{\kappa C}{\lambda_m}}\left(\sqrt{\ell(s,\widehat{t}_\mu)}\ell(s,\widehat{t}_\lambda)+\sqrt{\ell(s,\widehat{t}_\mu)}^3\right) \\
&\quad + 10\frac{L\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}}\Big[\ell(s,\widehat{t}_\lambda)+\ell(s,\widehat{t}_\mu)\Big]\Bigg]\Bigg\}\ . \qquad (37)
\end{aligned}$$

Using the inequality $ab \leqslant \frac{a^p}{p} + \frac{b^q}{q}$ with Hölder conjugates $p=3$, $q=\frac{3}{2}$, we have

$$\begin{aligned}
\sqrt{\ell(s,\widehat{t}_\mu)}\,\ell(s,\widehat{t}_\lambda) + \sqrt{\ell(s,\widehat{t}_\mu)}^3 &\leqslant \frac{1}{3}\sqrt{\ell(s,\widehat{t}_\mu)}^3 + \frac{2}{3}\ell(s,\widehat{t}_\lambda)^{\frac{3}{2}} + \sqrt{\ell(s,\widehat{t}_\mu)}^3 \\
&\leqslant \frac{4}{3}\left[\sqrt{\ell(s,\widehat{t}_\lambda)}^3 + \sqrt{\ell(s,\widehat{t}_\mu)}^3\right]\ . \qquad (38)
\end{aligned}$$

Claim 32 then follows from Equations (37) and (38), using the elementary inequality

$$\forall a,b,c,d\in\mathbb{R}\,,\qquad (a+b)\vee(c+d) \leqslant a\vee c + b\vee d\ .$$

■

As g is L-Lipschitz in its first argument, it follows from Claim 31 that for all $\lambda,\mu\in\Lambda$ such that $\lambda\leqslant\mu$,

$$\begin{aligned}
\left\|g\circ\widehat{t}_\lambda - g\circ\widehat{t}_\mu\right\|_\infty &\leqslant L\left\|\widehat{t}_\lambda - \widehat{t}_\mu\right\|_\infty \\
&\leqslant L\sqrt{\frac{\kappa C}{\lambda_m}}\sqrt{\ell(s,\widehat{t}_\mu)} + 10L^2\frac{\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}} \\
&\leqslant \widehat{w}_A\left(\sqrt{\ell(s,\widehat{t}_\mu)}\right) + \widehat{w}_A\left(\sqrt{\ell(s,\widehat{t}_\lambda)}\right)\ , \qquad (39)
\end{aligned}$$

where

$$\widehat{w}_A(x) = L\sqrt{\frac{\kappa C}{\lambda_m}}x + 5L^2\frac{\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}}\ . \qquad (40)$$

If follows that for all $k\geqslant 2$,

$$\begin{aligned}
P\left[\left(g\circ\widehat{t}_\lambda - g\circ\widehat{t}_\mu\right)^k\right] &\leqslant \left\|g\circ\widehat{t}_\lambda - g\circ\widehat{t}_\mu\right\|_\infty^k \\
&\leqslant \left[\widehat{w}_A\left(\sqrt{\ell(s,\widehat{t}_\mu)}\right) + \widehat{w}_A\left(\sqrt{\ell(s,\widehat{t}_\lambda)}\right)\right]^k\ .
\end{aligned}$$

This proves that hypothesis $H\left(\widehat{w}_A,\widehat{w}_A,(\widehat{t}_\lambda)_{\lambda\in\Lambda}\right)$, as defined in Appendix A, holds true.

It follows from Claim 32 and Equation (39) that, for all $k \geqslant 2$,

$$\begin{aligned} P\big[|g \circ \widehat{t}_\lambda - g \circ \widehat{t}_\mu|^k\big] &\leqslant \big\|g \circ \widehat{t}_\lambda - g \circ \widehat{t}_\mu\big\|_\infty^{k-2} P\left[\big(g(\widehat{t}_\lambda(X), Y) - g(\widehat{t}_\mu(X), Y)\big)^2\right] \\ &\leqslant \left[\widehat{w}_A\Big(\sqrt{\ell(s, \widehat{t}_\lambda)}\Big) + \widehat{w}_A\Big(\sqrt{\ell(s, \widehat{t}_\mu)}\Big)\right]^{k-2} \\ &\quad \times \left[\widehat{w}_B\Big(\sqrt{\ell(s, \widehat{t}_\lambda)}\Big) + \widehat{w}_B\Big(\sqrt{\ell(s, \widehat{t}_\mu)}\Big)\right]^2 \end{aligned}$$

which proves that $H\left(\widehat{w}_B, \widehat{w}_A, (\widehat{t}_\lambda)_{\lambda\in\Lambda}\right)$ holds true.

B.4 Conclusion of the Proof

We have proved that $H\left(\widehat{w}_B, \widehat{w}_A, (\widehat{t}_\lambda)_{\lambda\in\Lambda}\right)$ and $H\left(\widehat{w}_A, \widehat{w}_A, (\widehat{t}_\lambda)_{\lambda\in\Lambda}\right)$ hold, where $\widehat{w}_B$ is defined in Claim 32 and $\widehat{w}_A$ in Equation (40). Moreover, $x \mapsto \frac{\widehat{w}_A(x)}{x}$ is nonincreasing. Therefore, Theorem 17 applies with $\widehat{w}_{1,1} = \widehat{w}_A, \widehat{w}_{1,2} = \widehat{w}_A, \widehat{w}_{2,1} = \widehat{w}_B, \widehat{w}_{2,2} = \widehat{w}_A$, $x = \log n_v$ and it remains to bound the remainder terms $(R_{2,i})_{1\leqslant i\leqslant 4}$ of Equation (10). For each i, we bound $R_{2,i}(\theta)$ by a numerical constant times $\max\{T_1(\theta), T_2(\theta), T_3(\theta)\}$, where

$$\begin{aligned} T_1(\theta) &= \frac{6\rho}{100}\frac{\log(n_v|\Lambda|)}{\theta n_v} \\ T_2(\theta) &= (\nu \vee L)^2\, \kappa C \frac{\log^2(n_v|\Lambda|)}{\theta^3 \lambda_m n_v^2} \\ T_3(\theta) &= L(\nu \vee L)\kappa \frac{\log^{\frac{3}{2}}(n_v|\Lambda|)}{\theta \lambda_m n_v \sqrt{n_t}} \ . \end{aligned}$$

Summing up these bounds yields Theorem 11.

B.4.1 Bound on $R_{2,1}(\theta) = \sqrt{2}\theta\mathbb{E}\left[\delta^2\left(\widehat{w}_B, \frac{\theta}{2}\sqrt{\frac{n_v}{\log(n_v|\Lambda|)}}\right)\right]$

Recall that $\widehat{w}_B(x)^2 := \max\left\{\rho x^2,\ \nu\frac{4}{3}\sqrt{\frac{\kappa C}{\lambda_m}}x^3 + 10\nu L\frac{\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}}x^2\right\}$.

By Equation (30) in Lemma 28 with $a = \sqrt{\rho}$, $b = \nu\frac{4}{3}\sqrt{\frac{\kappa C}{\lambda_m}}$ and $c = 10\nu L\frac{\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}}$, we have

$$\delta^2\left(\widehat{w}_B, \frac{\theta}{2}\sqrt{\frac{n_v}{\log(n_v|\Lambda|)}}\right) \leqslant 4\rho\frac{\log(n_v|\Lambda|)}{\theta^2 n_v} + 29\nu^2\kappa C\frac{\big[\log(n_v|\Lambda|)\big]^2}{\theta^4\lambda_m n_v^2} + 80\nu L\kappa\frac{\big[\log(n_v|\Lambda|)\big]\sqrt{\widehat{y}}}{\theta^2\lambda_m n_v\sqrt{n_t}} \ .$$

Therefore,

$$R_{2,1}(\theta) \leqslant 4\sqrt{2}\rho\frac{\log(n_v|\Lambda|)}{\theta n_v} + 29\sqrt{2}\nu^2\kappa C\frac{\big[\log(n_v|\Lambda|)\big]^2}{\theta^3\lambda_m n_v^2} + 80\sqrt{2}\nu L\kappa\frac{\big[\log(n_v|\Lambda|)\big]\sqrt{\mathbb{E}[\widehat{y}]}}{\theta\lambda_m n_v\sqrt{n_t}} \ .$$

By Claim 30, $\mathbb{E}[\widehat{y}] \leqslant 4 + \log|\Lambda|$. Since $n_v \geqslant 100 \geqslant \mathrm{e}^4$, $\mathbb{E}[\widehat{y}] \leqslant \log(n_v|\Lambda|)$. As a result,

$$\begin{aligned} R_{2,1}(\theta) &\leqslant 6\rho\frac{\log(n_v|\Lambda|)}{\theta n_v} + 42\nu^2\kappa C\frac{\big[\log(n_v|\Lambda|)\big]^2}{\theta^3\lambda_m n_v^2} + 114\nu L\kappa\frac{\big[\log(n_v|\Lambda|)\big]^{\frac{3}{2}}}{\theta\lambda_m n_v\sqrt{n_t}} \\ &\leqslant 100T_1(\theta) + 42T_2(\theta) + 114T_3(\theta) \\ &\leqslant 256 \times \max\big\{T_1(\theta), T_2(\theta), T_3(\theta)\big\} \ . \end{aligned}$$

B.4.2 Bound on $R_{2,2}(\theta) = \frac{\theta^2}{2}\mathbb{E}\left[\delta^2\left(\widehat{w}_A, \frac{\theta^2}{4}\frac{n_v}{\log(n_v|\Lambda|)}\right)\right]$

Recall that by Equation (40), $\widehat{w}_A(x) = L\sqrt{\frac{\kappa C}{\lambda_m}}x + 5L^2\frac{\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}}$.

By Equation (29) in Lemma 28 with $b = L\sqrt{\frac{\kappa C}{\lambda_m}}$ and $c = 5L^2\frac{\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}}$, we have

$$\delta^2\left(\widehat{w}_A, \frac{\theta^2}{4}\frac{n_v}{\log(n_v|\Lambda|)}\right) \leqslant 16L^2\kappa C\frac{\log^2(n_v|\Lambda|)}{\theta^4\lambda_m n_v^2} + 40L^2\kappa\frac{\big[\log(n_v|\Lambda|)\big]\sqrt{\widehat{y}}}{\theta^2\lambda_m n_v\sqrt{n_t}} \ .$$

As $\mathbb{E}[\widehat{y}] \leqslant \log(n_v|\Lambda|)$ by Claim 30, it follows that

$$\begin{aligned} R_{2,2}(\theta) &\leqslant 8L^2\kappa C\frac{\log^2(n_v|\Lambda|)}{\theta^2\lambda_m n_v^2} + 20L^2\kappa\frac{\log^{\frac{3}{2}}(n_v|\Lambda|)}{\lambda_m n_v\sqrt{n_t}} \\ &\leqslant 8\theta T_2(\theta) + 20\theta T_3(\theta) \\ &\leqslant 28 \times \max\{T_1(\theta), T_2(\theta), T_3(\theta)\} \qquad \text{since } \theta \in (0,1] \ . \end{aligned}$$

B.4.3 Bound on $R_{2,3}(\theta) = \frac{1}{n_v}\left(\theta + \frac{2\big[1+\log(n_v|\Lambda|)\big]}{\theta}\right)\mathbb{E}\left[\widehat{\delta}^2\big(\widehat{w}_A, \sqrt{n_v}\big)\right]$

By Equation (29) in Lemma 28 with $b = L\sqrt{\frac{\kappa C}{\lambda_m}}$ and $c = 5L^2\frac{\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}}$, we have

$$\delta^2(\widehat{w}_A, \sqrt{n_v}) \leqslant L^2\frac{\kappa C}{\lambda_m n_v} + L^2\frac{10\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_v n_t}} \ .$$

As $\theta \in (0,1]$ and $n_v \geqslant 100 \geqslant \mathrm{e}^{\frac{3}{2}}$, we have $\theta + \frac{2}{\theta} \leqslant \frac{3}{\theta} \leqslant \frac{2\log(n_v|\Lambda|)}{\theta}$, hence

$$\theta + \frac{2[1+\log(n_v|\Lambda|)]}{\theta} \leqslant \frac{4\log(n_v|\Lambda|)}{\theta} \ . \tag{41}$$

Therefore,

$$R_{2,3}(\theta) \leqslant \frac{4\log(n_v|\Lambda|)}{\theta n_v}\left(L^2\frac{\kappa C}{\lambda_m n_v} + L^2\frac{10\kappa\sqrt{\mathbb{E}[\widehat{y}]}}{\lambda_m\sqrt{n_v n_t}}\right) \ .$$

Since $\mathbb{E}[\widehat{y}] \leqslant \log(n_v|\Lambda|)$ by Claim 30,

$$\begin{aligned} R_{2,3}(\theta) &\leqslant 4\log(n_v|\Lambda|)\frac{L^2\kappa C}{\theta\lambda_m n_v^2} + 40L^2\kappa\frac{\log^{\frac{3}{2}}(n_v|\Lambda|)}{\theta\lambda_m n_v\sqrt{n_v n_t}} \\ &\leqslant \frac{4\theta^2}{\log(n_v|\Lambda|)}T_2(\theta) + \frac{40}{\sqrt{n_v}}T_3(\theta) \\ &\leqslant 0.8T_2(\theta) + 4T_3(\theta) \qquad \text{since } n_v \geqslant 100 \text{ and } |\Lambda| \geqslant 2 \\ &\leqslant 4.8 \times \max\{T_1, T_2, T_3\} \ . \end{aligned}$$

B.4.4 Bound on $R_{2,4}(\theta) = \frac{1}{n_v}\left(\theta + \frac{2\left[1+\log(n_v|\Lambda|)\right]+\log^2(n_v|\Lambda|)}{\theta}\right)\mathbb{E}\left[\widehat{\delta}^2\big(\widehat{w}_A, n_v\big)\right]$

By Equation (29) in Lemma 28 with $b = L\sqrt{\frac{\kappa C}{\lambda_m}}$ and $c = 5L^2\frac{\kappa\sqrt{\widehat{y}}}{\lambda_m\sqrt{n_t}}$, we have

$$\delta^2(\widehat{w}_A, n_v) \leqslant L^2\frac{\kappa C}{\lambda_m n_v^2} + L^2\frac{10\kappa\sqrt{\widehat{y}}}{\lambda_m n_v\sqrt{n_t}} \ . \tag{42}$$

Since $\theta \in [0,1]$, $n_v \geqslant 100$ and $|\Lambda| \geqslant 2$, we have $\log(n_v|\Lambda|) \geqslant \log(200) \geqslant 5$ and

$$\begin{aligned} \theta + \frac{2\big[1+\log(n_v|\Lambda|)\big]}{\theta} &\leqslant \frac{4\log(n_v|\Lambda|)}{\theta} \qquad \text{by Equation (41)} \\ &\leqslant \frac{4\log^2(n_v|\Lambda|)}{5\theta} \ . \end{aligned}$$

Hence, by Equation (42),

$$R_{2,4}(\theta) \leqslant \frac{1.8\log^2(n_v|\Lambda|)}{\theta n_v}\left[L^2\frac{\kappa C}{\lambda_m n_v^2} + L^2\frac{10\kappa\sqrt{\mathbb{E}[\widehat{y}]}}{\lambda_m n_v\sqrt{n_t}}\right] \ .$$

Since $\mathbb{E}[\widehat{y}] \leqslant \log(n_v|\Lambda|)$,

$$\begin{aligned} R_{2,4}(\theta) &\leqslant 1.8\log^2(n_v|\Lambda|)\frac{L^2\kappa C}{\theta\lambda_m n_v^3} + 18L^2\kappa\frac{\log^{\frac{5}{2}}(n_v|\Lambda|)}{\theta\lambda_m n_v^2\sqrt{n_t}} \\ &\leqslant \frac{1.8\theta^2}{n_v}T_2(\theta) + 18\frac{\log(n_v|\Lambda|)}{n_v}T_3(\theta) \ . \end{aligned}$$

Since $n_v \geqslant 100$ and $|\Lambda| \leqslant \mathrm{e}^{\sqrt{n_v}}$, we have $\frac{\log(n_v|\Lambda|)}{n_v} \leqslant \frac{\log(n_v)}{n_v} + \frac{\log(\mathrm{e}^{\sqrt{n_v}})}{n_v} \leqslant \frac{\log(100)}{100} + \frac{1}{10} \leqslant 0.15$ and so

$$\begin{aligned} R_{2,4}(\theta) &\leqslant 0.018\,T_2(\theta) + 2.7\,T_3(\theta) \\ &\leqslant 2.8 \times \max\big\{T_1(\theta)\,,\,T_2(\theta)\,,\,T_3(\theta)\big\} \ . \end{aligned}$$

B.4.5 Conclusion

Summing up the above inequalities, we get that for every $\theta \in (0,1]$,

$$\begin{aligned} R_2(\theta) &= R_{2,1}(\theta) + R_{2,2}(\theta) + R_{2,3}(\theta) + R_{2,4}(\theta) \\ &\leqslant 292\max\big\{T_1(\theta)\,,\,T_2(\theta)\,,\,T_3(\theta)\big\} \ . \end{aligned}$$

Equation (10) in Theorem 17 thus yields

$$(1-\theta)\mathbb{E}\big[\ell(s,\widehat{f}^{\mathrm{ag}}_{\mathcal{T}})\big] \leqslant (1+\theta)\mathbb{E}\Big[\min_{\lambda\in\Lambda}\ell(s,\mathcal{A}_\lambda(D_{n_t}))\Big] + 292\,\max\big\{T_1(\theta)\,,\,T_2(\theta)\,,\,T_3(\theta)\big\}$$

which proves Theorem 11 with $b_1 = 292(\nu\vee L)^2\kappa C$ and $b_2 = 292\,L(\nu\vee L)\kappa$. ■

Appendix C. Proof of Proposition 10 and Corollary 12

Let us start by two useful lemmas.

Lemma 33 *If ψ is a convex, Lipschitz-continuous, and even function, and if Y is a random variable with a non-atomic distribution, then the function*

$$R : u \mapsto \mathbb{E}\big[\psi(u-Y)\big]$$

is convex and differentiable with derivative $R'(u) = \mathbb{E}[\psi'(u-Y)]$. Moreover, if Y is symmetric around q, that is, $(q-Y)\sim(Y-q)$, then R reaches a minimum at q.

Proof First, remark that R is convex by convexity of ψ. Let $u\in\mathbb{R}$. For $h\neq 0$, let $k(h,Y) = \frac{\psi(u+h-Y)-\psi(u-Y)}{h}$. Let A be the set on which ψ is non-differentiable. Since ψ is convex, A is at most countable. By definition, $k(h,Y)\underset{h\to 0}{\longrightarrow}\psi'(u-Y)$ whenever $u-Y\notin A$, that is to say $Y\notin u-A$. Since Y is non-atomic, $\mathbb{P}(Y\notin u-A)=1$. Moreover, since ψ is Lipschitz, there exists a constant L such that $\forall h\neq 0, |k(h,Y)|\leqslant L$. Therefore, by the dominated convergence theorem,

$$\frac{R(u+h)-R(u)}{h} = \mathbb{E}[k(h,Y)] \underset{h\to 0}{\longrightarrow} \mathbb{E}[\psi'(u-Y)]\ .$$

Thus, R is differentiable and for all $u\in\mathbb{R}$, $R'(u) = \mathbb{E}[\psi'(u-Y)]$.

Moreover, we have

$$\begin{aligned} R'(q) &= \mathbb{E}[\psi'(q-Y)] \\ &= -\mathbb{E}[\psi'(Y-q)] \qquad \text{since } \psi'(-x) = -\psi'(x) \text{ on } \mathbb{R}\backslash A \\ &= -\mathbb{E}[\psi'(q-Y)] \qquad \text{since } (Y-q)\sim(q-Y)\ , \end{aligned}$$

which implies that $R'(q)=0$. Hence, R reaches a minimum at q since R is convex. ■

Lemma 34 *Let $G:\mathbb{R}\to\mathbb{R}$ be a differentiable convex function that reaches a minimum at $u_*\in\mathbb{R}$. If there exists $\varepsilon,\delta>0$ such that*

$$\forall u\in[u_*-\delta\,,\,u_*+\delta]\,,\qquad |G'(u)|\geqslant\varepsilon|u-u_*|\ , \tag{43}$$

then for all $(u,v)\in\mathbb{R}^2$,

$$(u-v)^2 \leqslant \left[\frac{4}{\varepsilon}\vee\left(\frac{4}{\varepsilon\delta}|u-v|\right)\right]\big[G(u)+G(v)-2G(u_*)\big]\ .$$

Proof By integrating Equation (43),

$$\forall u \in [u_* - \delta\,,\, u_* + \delta]\,, \qquad G(u) - G(u_*) \geqslant \frac{\varepsilon}{2}(u - u_*)^2 \ . \tag{44}$$

Let

$$h(u) = \frac{1}{\delta}\big[G(u_* + \delta) - G(u_*)\big](u - u_*) \ . \tag{45}$$

By convexity of G, for any $u \geqslant u_* + \delta$, $G(u) - G(u_*) \geqslant h(u)$. Hence by Equation (44) with $u = u_* + \delta$ and Equation (45),

$$\forall u \geqslant u_* + \delta\,, \qquad G(u) - G(u_*) \geqslant \frac{1}{\delta}\frac{\varepsilon}{2}\delta^2(u - u_*) = \frac{\varepsilon\delta}{2}(u - u_*) \ .$$

The same argument applies to the convex function $x \mapsto G(-x)$ with minimum $-u_*$, which yields that

$$\forall u \in \mathbb{R} \quad \text{such that} \quad |u - u_*| \geqslant \delta\,, \qquad G(u) - G(u_*) \geqslant \frac{\varepsilon\delta}{2}|u - u_*| \ . \tag{46}$$

Let $(u, v) \in \mathbb{R}^2$. Assume without loss of generality that $|u - u_*| \geqslant |v - u_*|$. If $|u - u_*| \leqslant \delta$ then by Equation (44),

$$\begin{aligned}(u - v)^2 &\leqslant 2(u - u_*)^2 + 2(v - u_*)^2 \\ &\leqslant \frac{4}{\varepsilon}\big[G(u) + G(v) - 2G(u_*)\big] \ .\end{aligned}$$

Otherwise, by Equation (46),

$$\begin{aligned}(u - v)^2 &\leqslant |u - v|\big(|u - u_*| + |v - u_*|\big) \\ &\leqslant 2|u - v| \times |u - u_*| \\ &\leqslant \frac{4}{\varepsilon\delta}|u - v|\big[G(u) - G(u_*)\big] \\ &\leqslant \frac{4}{\varepsilon\delta}|u - v|\big[G(u) + G(v) - 2G(u_*)\big] \ .\end{aligned}$$

■

C.1 Proof of Proposition 10

Now, we can prove Proposition 10. Let $R_x : u \mapsto \int |u - y| \mathrm{d}F_x(y)$. By Lemma 33 with $\psi = |\cdot|$, for all $v \in \mathbb{R}$,

$$\begin{aligned}R_x'(v) &= \int (-\mathbb{I}_{v-y\leqslant 0} + \mathbb{I}_{v-y\geqslant 0})\,\mathrm{d}F_x(y) \\ &= F_x(v) - \big[1 - F_x(v)\big] \\ &= 2\Big[F_x(v) - F_x\big(s(x)\big)\Big]\end{aligned}$$

since by definition, $F_x(s(x)) = \frac{1}{2}$. Hence by hypothesis (5),

$$\forall u \in \big[s(x) - b(x)\,,\, s(x) + b(x)\big]\,, \qquad \big|R'_x(u)\big| \geqslant 2a(x)\big|u - s(x)\big| \ .$$

Therefore by Lemma 34 with $G = R_x$, $\delta = b(x)$ and $\varepsilon = 2a(x)$, we obtain that for all $x \in \mathcal{X}$ and $(u, v) \in \mathbb{R}^2$,

$$\begin{aligned}(u-v)^2 &\leqslant \left(\frac{4}{2a(x)} \vee \frac{4|u-v|}{2a(x)b(x)}\right)\Big[R_x(u) + R_x(v) - 2R_x\big(s(x)\big)\Big] \\ &\leqslant \left[\frac{2}{a_m} \vee \left(\frac{2}{\mu_m}|u-v|\right)\right]\Big[R_x(u) + R_x(v) - 2R_x\big(s(x)\big)\Big] \ .\end{aligned}$$

Since g is the function $(u, y) \mapsto |u - y|$, it follows by taking $x = X$ that

$$\big[g(u,Y) - g(v,Y)\big]^2 \leqslant (u-v)^2 \leqslant \left[\frac{2}{a_m} \vee \left(\frac{2}{\mu_m}|u-v|\right)\right]\big[\ell_X(u) + \ell_X(v)\big] \ ,$$

which implies hypothesis $SC_{\frac{2}{a_m},\frac{2}{\mu_m}}$. ■

C.2 Proof of Corollary 12

Corollary 12 is a consequence of Theorem 11. Let us check that its assumptions are satisfied.

C.2.1 Compatibility Hypothesis $Comp_1(c_0^{eps}, c_\varepsilon^{eps})$

Fix $x \in \mathcal{X}$ and let p_x, F_x be respectively the density and the cumulative distribution function of Y given $X = x$. By assumption, p_x is symmetric. Recall that the contrast function here is $\gamma(t, (x, y)) = c_0^{eps}(t(x), y) = |t(x) - y|$, so any conditional median is a possible value for $s(x)$, and we can take $s(x)$ equal to the center of symmetry. Let

$$R_{\varepsilon,x} : u \mapsto \int_y c_\varepsilon^{eps}(u, y)p_x(y)\mathrm{d}y = \int \psi_\varepsilon(u - y)p_x(y)\mathrm{d}y \ , \tag{47}$$

where $\psi_\varepsilon(z) = (|z| - \varepsilon)_+$ for any $z \in \mathbb{R}$. Lemma 33 applies, since p_x is symmetric by assumption and ψ_ε is even, convex and 1-Lipschitz.

Hence for any $\varepsilon \geqslant 0$, $R_{\varepsilon,x}$ has a minimum at $s(x)$ and is differentiable, with

$$\begin{aligned}R'_{\varepsilon,x}(u) &= \int \psi'_\varepsilon(u - y)p_x(y)\mathrm{d}y = \int \left[-\mathbb{I}_{u-y\leqslant-\varepsilon} + \mathbb{I}_{u-y\geqslant\varepsilon}\right]p_x(y)\mathrm{d}y \\ &= F_x(u - \varepsilon) - \big[1 - F_x(u + \varepsilon)\big] \ .\end{aligned}$$

Therefore, for any $\varepsilon \geqslant 0$ and $u \in \mathbb{R}$,

$$R'_{\varepsilon,x}(u) - R'_{0,x}(u) = \int_0^\varepsilon \big[-p_x(u - t) + p_x(u + t)\big]\mathrm{d}t \ . \tag{48}$$

Now, assume that $u \geqslant s(x)$. By symmetry of p_x around $s(x)$, for all $t \geqslant 0$,

$$\begin{aligned}p_x(u - t) &= p_x\Big(s(x) + \big[u - s(x) - t\big]\Big) \\ &= p_x\big[s(x) + |u - s(x) - t|\big] \ .\end{aligned} \tag{49}$$

Since p_x is unimodal, its mode is $s(x)$ and p_x is nonincreasing on $[s(x), +\infty)$. It follows from Equation (49) that for all $u \geqslant s(x)$ and $t \geqslant 0$,

$$\begin{aligned} p_x(u-t) &\geqslant p_x\big(s(x) + |u - s(x)| + t\big) \\ &= p_x(u+t) \ . \end{aligned} \tag{50}$$

Therefore, by Equations (48) and (50), for all $u \geqslant s(x)$ and $\varepsilon \geqslant 0$, $R'_{\varepsilon,x}(u) \leqslant R'_{0,x}(u)$. By integration, this implies that for all $u \geqslant s(x)$,

$$R_{\varepsilon,x}(u) - R_{\varepsilon,x}\big(s(x)\big) \leqslant R_{0,x}(u) - R_{0,x}\big(s(x)\big) \ . \tag{51}$$

By Equation (47) and symmetry of p_x, $R_{\varepsilon,x}$ and $R_{0,x}$ are symmetric around $s(x)$, hence inequality (51) is also valid when $u \leqslant s(x)$. Choosing $x = X$, $u = t(X)$ and taking an expectation, we get $\mathcal{L}_{c_\varepsilon^{eps}}(t) - \mathcal{L}_{c_\varepsilon^{eps}}(s) \leqslant \mathcal{L}_{c_0^{eps}}(t) - \mathcal{L}_{c_0^{eps}}(s)$ which proves $Comp_1(c_0^{eps}, c_\varepsilon^{eps})$.

C.2.2 Hypothesis $SC_{4\sigma,8}$

We first compute a lower bound on $R_{0,x}$.

Let $q_{x,\frac{1}{4}} = \sup\{y \,:\, F_x(y) \leqslant \frac{1}{4}\}$ and $q_{x,\frac{3}{4}} = \inf\{y \,:\, F_x(y) \geqslant \frac{3}{4}\}$. By continuity of F_x, $F_x(q_{x,\frac{1}{4}}) = \frac{1}{4}$ and $F_x(q_{x,\frac{3}{4}}) = \frac{3}{4}$. Let $\sigma(x) = q_{x,\frac{3}{4}} - q_{x,\frac{1}{4}}$, which is the smallest determination of the interquartile range. By symmetry of p_x around $s(x)$, $\frac{1}{2}\big[q_{x,\frac{1}{4}} + q_{x,\frac{3}{4}}\big] = s(x)$, therefore $q_{x,\frac{3}{4}} = s(x) + \frac{\sigma(x)}{2}$ and $q_{x,\frac{1}{4}} = s(x) - \frac{\sigma(x)}{2}$.

For any $u \in \Big[s(x) - \frac{\sigma(x)}{2}\,,\, s(x) + \frac{\sigma(x)}{2}\Big]$, by symmetry of p_x around $s(x)$,

$$\begin{aligned} \big|F_x(u) - F_x\big(s(x)\big)\big| &= \int_{s(x)}^{s(x)+|u-s(x)|} p_x(v)\mathrm{d}v \\ &= |u - s(x)|\frac{1}{|u-s(x)|}\int_{s(x)}^{s(x)+|u-s(x)|} p_x(v)\mathrm{d}v \ . \end{aligned}$$

Since p_x is nonincreasing on $[s(x), +\infty)$ and $|u - s(x)| \leqslant \frac{\sigma(x)}{2}$,

$$\begin{aligned} \big|F_x(u) - F_x\big(s(x)\big)\big| &\geqslant |u - s(x)|\frac{2}{\sigma(x)}\int_{s(x)}^{s(x)+\frac{\sigma(x)}{2}} p_x(v)\mathrm{d}v \\ &= |u - s(x)|\frac{2}{\sigma(x)}\Big[F_x\big(q_{x,\frac{3}{4}}\big) - F_x\big(s(x)\big)\Big] \\ &= \frac{|u - s(x)|}{2\sigma(x)} \ . \end{aligned}$$

Hence, by Proposition 10 with $a(x) = \frac{1}{2\sigma(x)}$ and $b(x) = \frac{\sigma(x)}{2}$, (g, X, Y) satisfies hypothesis $SC_{4\sigma,8}$.

C.2.3 Conclusion

To conclude, we apply Theorem 11 with $\kappa = 1$, $C = 1$, $L = 1$ (since c_0^{eps} and c_ε^{eps} are 1-Lipschitz), $\rho = 4\sigma$ and $\nu = 8$. Since constants b_1, b_2 of Theorem 11 only depend on

κ, L, C, ν and all these parameters have now received explicit values, the constants b_1, b_2 are now absolute. ■

Appendix D. Classification: Proof of Theorem 13

In the proof of Theorem 17, we use convexity of the risk to show that the risk of the average is less than the average of the risk. A property of this type also holds in the setting of classification, with the average replaced by the majority vote.

Proposition 35 *In the classification framework—see Example 1—, let $(\widehat{f}_i)_{1\leqslant i\leqslant V}$ denote a finite family of functions $\mathcal{X}\to\mathcal{Y}$ and let $\widehat{f}^{\mathrm{mv}}$ be some majority vote rule, defined by*

$$\forall x\in\mathcal{X}\,,\qquad \widehat{f}^{\mathrm{mv}}(x)\in\operatorname*{argmax}_{y\in\mathcal{Y}}|\{i\in[V]:\widehat{f}_i(x)=m\}|\ .$$

Then, we have

$$\ell(s,\widehat{f}^{\mathrm{mv}})\leqslant\frac{M}{V}\sum_{i=1}^{V}\ell(s,\widehat{f}_i)\qquad\text{and}\qquad\mathcal{L}(\widehat{f}^{\mathrm{mv}})\leqslant\frac{2}{V}\sum_{i=1}^{V}\mathcal{L}(\widehat{f}_i)\ .$$

Proof For any $y\in\mathcal{Y}$, define $\eta_y: x\mapsto\mathbb{P}(Y=y|X=x)$. Then, for any $f\in\mathbb{S}$, we have $\mathcal{L}(f)=\mathbb{E}[1-\eta_{f(X)}(X)]$ hence $s(X)\in\operatorname{argmax}_{y\in\mathcal{Y}}\eta_y(X)$ and

$$\ell(s,f)=\mathbb{E}\Big[\max_{y\in\mathcal{Y}}\eta_y(X)-\eta_{f(X)}(X)\Big]=\mathbb{E}\big[\eta_{s(X)}(X)-\eta_{f(X)}(X)\big]\ .$$

We now fix some $x\in\mathcal{X}$ and define $\mathcal{C}_x(y)=\{i\in[V]:\widehat{f}_i(x)=y\}$ and $C_x=\max_{y\in\mathcal{Y}}|\mathcal{C}_x(y)|$. Since $C_xM\geqslant\sum_{y\in\mathcal{Y}}|\mathcal{C}_x(y)|=V$, we get $C_x\geqslant V/M$. On the other hand, by definition of $\widehat{f}^{\mathrm{mv}}$,

$$\frac{1}{V}\sum_{i=1}^{V}\underbrace{\big[\eta_{s(x)}(x)-\eta_{\widehat{f}_i(x)}(x)\big]}_{\geqslant 0}\geqslant\frac{C_x}{V}\big[\eta_{s(x)}(x)-\eta_{\widehat{f}^{\mathrm{mv}}(x)}(x)\big]\geqslant\frac{1}{M}\big[\eta_{s(x)}(x)-\eta_{\widehat{f}^{\mathrm{mv}}(x)}(x)\big]\ .$$

Integrating over x (with respect to the distribution of X) yields the first bound.

For the second bound, fix $x\in\mathcal{X}$ and define $\mathcal{C}_x(y)$ and C_x as above. Let $y\in\mathcal{Y}$ be such that $\widehat{f}^{\mathrm{mv}}(x)\neq y$. Since y occurs less often than $\widehat{f}^{\mathrm{mv}}(x)$ among $\widehat{f}_1(x),\dots,\widehat{f}_V(x)$, we have $|\mathcal{C}_x(y)|\leqslant V/2$. Therefore,

$$\frac{1}{V}\sum_{i=1}^{V}\mathbb{I}_{\{\widehat{f}_i(x)\neq y\}}=\frac{V-|\mathcal{C}_x(y)|}{V}\geqslant\frac{1}{2}\ .$$

Thus,

$$\widehat{f}^{\mathrm{mv}}(x)\neq y\qquad\text{implies}\qquad\frac{1}{V}\sum_{i=1}^{V}\mathbb{I}_{\{\widehat{f}_i(x)\neq y\}}\geqslant\frac{1}{2}\ .$$

Hence, for any $y \in \mathcal{Y}$,

$$\mathbb{I}_{\{\widehat{f}^{\mathrm{mv}}(x)\neq y\}} \leqslant \frac{2}{V}\sum_{i=1}^{V} \mathbb{I}_{\{\widehat{f}_i(x)\neq y\}} \ .$$

Taking expectations with respect to (x, y) yields $\mathcal{L}(\widehat{f}^{\mathrm{mv}}) \leqslant 2V^{-1}\sum_{i=1}^{V}\mathcal{L}(\widehat{f}_i)$. ■

We can now proceed with the proof of Theorem 13. It relies on a result by Massart (2007, Equation 8.60, which is itself a consequence of Corollary 8.8), which holds true as soon as

$$\forall t \in \mathbb{S}, \qquad \mathrm{Var}\big(\mathbb{I}_{\{t(X)\neq Y\}} - \mathbb{I}_{\{s(X)\neq Y\}}\big) \leqslant \Big[w\big(\sqrt{\ell(s,t)}\big)\Big]^2 \tag{52}$$

for some nonnegative and nondecreasing continuous function w on $\mathbb{R}^+$, such that $x \mapsto w(x)/x$ is nonincreasing on $(0, +\infty)$ and $w(1) \geqslant 1$.

Let us first prove that assumption (52) holds true. On the one hand, since $\mathcal{Y} = \{0, 1\}$, for any $t \in \mathbb{S}$,

$$\begin{aligned}\mathrm{Var}\big(\mathbb{I}_{\{t(X)\neq Y\}} - \mathbb{I}_{\{s(X)\neq Y\}}\big) &\leqslant \mathbb{E}\Big[\big(\mathbb{I}_{\{t(X)\neq Y\}} - \mathbb{I}_{\{s(X)\neq Y\}}\big)^2\Big] \\ &= \mathbb{E}\big[\mathbb{I}_{\{t(X)\neq s(X)\}}\big] = \mathbb{E}\big[|t(X) - s(X)|\big] \ .\end{aligned} \tag{53}$$

On the other hand, since we consider binary classification with the 0–1 loss, for any $t \in \mathbb{S}$ and $h > 0$,

$$\begin{aligned}\ell(s,t) &= \mathbb{E}\big[|2\eta(X) - 1| \cdot |t(X) - s(X)|\big] \qquad \text{by Devroye et al. (1996, Theorem 2.2)} \\ &\geqslant h\mathbb{E}\big[|t(X) - s(X)|\mathbb{I}_{\{|2\eta(X)-1|\geqslant h\}}\big] \\ &\geqslant h\mathbb{E}\big[|t(X) - s(X)| - \mathbb{I}_{\{|2\eta(X)-1|<h\}}\big] \qquad \text{since } \|t - s\|_\infty \leqslant 1 \\ &\geqslant h\mathbb{E}\big[|t(X) - s(X)|\big] - rh^{\beta+1} \qquad \text{by (MA).}\end{aligned}$$

This lower bound is maximized by taking

$$h = h_* := \left(\frac{\mathbb{E}\big[|t(X) - s(X)|\big]}{r(\beta+1)}\right)^{\frac{1}{\beta}} \ ,$$

which belongs to $[0, 1]$ since $r \geqslant 1$ and $\mathbb{E}\big[|t(X) - s(X)|\big] \leqslant 1$. Thus, we obtain

$$\ell(s,t) \geqslant h_*\frac{\beta}{\beta+1}\mathbb{E}\big[|t(X) - s(X)|\big] = \frac{\beta}{(\beta+1)^{(\beta+1)/\beta}r^{1/\beta}}\mathbb{E}\big[|t(X) - s(X)|\big]^{(\beta+1)/\beta} \ .$$

Therefore, Equation (53) leads to

$$\mathrm{Var}\big(\mathbb{I}_{\{t(X)\neq Y\}} - \mathbb{I}_{\{s(X)\neq Y\}}\big) \leqslant \mathbb{E}\big[|t(X) - s(X)|\big] \leqslant \frac{\beta+1}{\beta^{\beta/(\beta+1)}}r^{\frac{1}{\beta+1}}\ell(s,t)^{\frac{\beta}{\beta+1}} \ .$$

By Lemma 36 below, $\frac{\beta+1}{\beta^{\beta/(\beta+1)}} \leqslant 2$; hence, defining $r_1 = 2r^{\frac{1}{\beta+1}}$, Equation (52) holds true with $w(u) = \sqrt{r_1}u^{\frac{\beta}{\beta+1}}$, which statisfies the required conditions. So, by Massart (2007, Equation 8.60), for any $\theta \in (0, 1)$,

$$\mathbb{E}\big[\ell\big(s, \widehat{f}_T^{\mathrm{ho}}\big) \,|\, D_n^T\big] \leqslant \frac{1+\theta}{1-\theta}\inf_{m\in\mathcal{M}}\Big\{\ell\big(s, \mathcal{A}_m(D_n^T)\big)\Big\} + \frac{\delta_*^2}{1-\theta}\left[2\theta + \log\big(\mathrm{e}|\mathcal{M}|\big)\left(\frac{1}{3} + \theta^{-1}\right)\right]$$

where δ_* is the positive solution of the fixed-point equation $w(\delta_*) = \sqrt{n_v}\delta_*^2$, that is,

$$\delta_*^2 = \left(\frac{r_1}{n_v}\right)^{\frac{\beta+1}{\beta+2}} \ .$$

Taking expectations with respect to the training data D_n^T, we obtain

$$\mathbb{E}\big[\ell(s, \widehat{f}_T^{\mathrm{ho}})\big] \leqslant \frac{1+\theta}{1-\theta}\mathbb{E}\left[\inf_{m\in\mathcal{M}} \ell\big(s, \mathcal{A}_m(D_n^T)\big)\right] + \frac{2r^{\frac{1}{\beta+2}}}{1-\theta}\frac{2\theta + \log\big(\mathrm{e}|\mathcal{M}|\big)\left(\frac{1}{3}+\theta^{-1}\right)}{n_v^{\frac{\beta+1}{\beta+2}}} \ .$$

Under assumption (2), $\mathbb{E}\big[\ell(s, \widehat{f}_T^{\mathrm{ho}})\big]$ and $\mathbb{E}\big[\mathcal{L}(\widehat{f}_T^{\mathrm{ho}})\big]$ do not depend on $T \in \mathcal{T}$ (they only depend on T through its cardinality n_t). Now, by Proposition 35 applied to $(\widehat{f}_T^{\mathrm{ho}})_{T\in\mathcal{T}}$,

$$\begin{aligned}\mathbb{E}\big[\ell(s, \widehat{f}_{\mathcal{T}}^{\mathrm{mv}})\big] &\leqslant 2\mathbb{E}\big[\ell(s, \widehat{f}_{T_1}^{\mathrm{ho}})\big] \\ &\leqslant 2\frac{1+\theta}{1-\theta}\mathbb{E}\left[\inf_{m\in\mathcal{M}} \ell\big(s, \mathcal{A}_m(D_n^T)\big)\right] + \frac{4r^{\frac{1}{\beta+2}}}{1-\theta}\frac{2\theta + \log\big(\mathrm{e}|\mathcal{M}|\big)\left(\frac{1}{3}+\theta^{-1}\right)}{n_v^{\frac{\beta+1}{\beta+2}}} \ .\end{aligned}$$

Taking $\theta = 1/5$ leads to the result. ■

The proof of Theorem 13 makes use of the following lemma.

Lemma 36 *For all $\beta > 0$, we have*

$$\frac{\beta+1}{\beta^{\beta/(\beta+1)}} \leqslant 2 \ .$$

Proof We first notice that

$$\begin{aligned}\log\left(\frac{\beta+1}{\beta^{\beta/(\beta+1)}}\right) &= \log(\beta+1) - \frac{\beta}{\beta+1}\log\beta \\ &= \log(\beta+1) - \frac{\beta}{\beta+1}\left[\log(\beta+1) + \log\left(\tfrac{\beta}{\beta+1}\right)\right] \\ &= \frac{\log(\beta+1)}{\beta+1} - \frac{\beta}{\beta+1}\log\left(\tfrac{\beta}{\beta+1}\right) \ .\end{aligned}$$

Defining $p = \frac{1}{\beta+1}$, this can be written

$$\log\left(\frac{\beta+1}{\beta^{\beta/(\beta+1)}}\right) = -p\log p - (1-p)\log(1-p) \ ,$$

which is the entropy of a Bernoulli distribution. The result follows from the fact that this entropy attains its maximal value $\log 2$ at $p = \frac{1}{2}$. ■

References

Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

Sylvain Arlot and Matthieu Lerasle. Choice of V for V-fold cross-validation in least-squares density estimation. *Journal of Machine Learning Research (JMLR)*, 17(208):1–50, 2016.

Viorel Barbu and Teodor Precupanu. *Convexity and optimization in Banach spaces*. Springer, 2012.

Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*. Springer Series in the Data Sciences. Springer, 2015.

Gérard Biau and Erwan Scornet. A random forest guided tour. *TEST*, 25(2):197–227, 2016.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: PS*, 9:323–375, 2005.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

Peter Bühlmann and Bin Yu. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002.

Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference*. Springer-Verlag, New York, second edition, 2002. A practical information-theoretic approach.

Olivier Catoni. *Statistical Learning Theory and Stochastic Optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer, Berlin, 2001. Lectures from the 31st Summer School on Probability Theory held in Saint-Flour, July 8 – 25, 2001, with a foreword by Jean Picard.

Luc P. Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

Thomas G. Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

Mona Eberts and Ingo Steinwart. Optimal regression rates for SVMs using Gaussian kernels. *Electron. J. Statist.*, 7:1–42, 2013.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1, part 2): 119–139, 1997. EuroCOLT '95.

Robin Genuer. Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24(3):543–562, 2012.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer New York, 2002.

Peter Hall and Andrew P. Robinson. Reducing variability of crossvalidation for smoothing-parameter choice. *Biometrika*, 96(1):175–186, January 2009.

Andres Hoyos-Idrobo, Yannick Schwartz, Gael Varoquaux, and Bertrand Thirion. Improving sparse recovery on structured images with bagged clustering. In *2015 International Workshop on Pattern Recognition in NeuroImaging*. IEEE, June 2015.

Yoonsuh Jung. Efficient tuning parameter selection by cross-validated score in high dimensional models. *International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering*, 10(1):19–25, 2016.

Yoonsuh Jung and Jianhua Hu. A K-fold averaging cross-validation procedure. *Journal of Nonparametric Statistics*, 27(2):167–179, 2015.

Guillaume Lecué. Suboptimality of Penalized Empirical Risk Minimization in Classification. In *COLT 2007*, volume 4539 of *Lecture Notes in Artificial Intelligence*. Springer, Berlin, 2007.

Guillaume Maillard. Aggregated hold out for sparse linear regression with a robust loss function, 2020a. arXiv:2002.11553.

Guillaume Maillard. *Hold-out and Aggregated hold-out*. PhD thesis, Université Paris-Saclay, September 2020b.

Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

Pascal Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

Ron Meir and Tong Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

Arkadi Nemirovski. *Topics in Non-parametric Statistics*, volume 1738 of *Lecture Notes in Math.* Springer, Berlin, 2000.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, and et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

Maya L. Petersen, Annette M. Molinaro, Sandra E. Sinisi, and Mark J. van der Laan. Cross-validated bagged learning. *Journal of Multivariate Analysis*, 98(9):1693–1704, October 2007.

Joseph Salmon and Arnak S. Dalalyan. Optimal aggregation of affine estimators. In *COLT - 24th Conference on Learning Theory - 2011*, Budapest, Hungary, Jul 2011.

Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Methodological*, 58(1):267–288, 1996.

Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Gaël Varoquaux, Pradeep Reddy Raamana, Denis A. Engemann, Andrés Hoyos-Idrobo, Yannick Schwartz, and Bertrand Thirion. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, 145:166–179, January 2017.

Yuhong Yang. Adaptive regression by mixing. *Journal of the American Statistical Association*, 96(454):574–588, 2001.