HAL open science

Aggregated hold out for sparse linear regression with a robust loss function

Guillaume Maillard

To cite this version:

Guillaume Maillard. Aggregated hold out for sparse linear regression with a robust loss function. 2020. hal-02485694v1

HAL Id: hal-02485694

https://hal.science/hal-02485694v1

Preprint submitted on 26 Feb 2020 (v1), last revised 28 Nov 2022 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Aggregated hold out for sparse linear regression with a robust loss function

Guillaume Maillard
guillaume.maillard@u-psud.fr
Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France.

February 26, 2020

Abstract

Sparse linear regression methods generally have a free hyperparameter which controls the amount of sparsity, and is subject to a bias-variance tradeoff. This article considers the use of Aggregated hold-out to aggregate over values of this hyperparameter, in the context of linear regression with the Huber loss function. Aggregated hold-out (Agghoo) is a procedure which averages estimators selected by hold-out (cross-validation with a single split). In the theoretical part of the article, it is proved that Agghoo satisfies a non-asymptotic oracle inequality when it is applied to sparse estimators which are parametrized by their zero-norm. In particular, this includes a variant of the Lasso introduced by Zou, Hastié and Tibshirani [39]. Simulations are used to compare Agghoo with cross-validation. They show that Agghoo performs better than CV when the intrinsic dimension is high and when there are confounders correlated with the predictive covariates.

1 Introduction

From the statistical learning point of view, linear regression is a risk-minimization problem wherein the aim is to minimize the average *prediction error* $\phi(Y - \theta^T X)$ on a new, independent data-point (X, Y), as measured by a *loss function* ϕ. When $\phi(x) = x^2$, this yields classical least-squares regression; however, Lipschitz-continuous loss functions have better robustness properties and are therefore preferred in the presence of heavy-tailed noise. In general, substracting the risk of the (distribution-dependent) optimal predictor yields a measure of performance for estimators, called the *excess risk*, which coïncides with the L^2 norm in the least-squares case.

In the high-dimensional setting, where $X \in \mathbb{R}^d$ with potentially $d >> n$, minimizing the risk over all θs is impossible; some assumptions must be made. A popular approach is to suppose that only a small number k_* of covariates are relevant to the prediction of Y, so that θ may be sought among the *sparse* vectors with less than k_* non-zero components. Estimators which target such problems include the Lasso and its variants, LARS, stagewise regression and the classical greedy procedures of stepwise regression. In the robust setting, variants

of the Lasso with robust loss functions have been investigated by a number of authors [14, 23, 7, 33].

Such methods generally introduce a free hyperparameter which regulates the "sparsity" of the estimator; sometimes this is directly the number of non-zero components, as in stepwise procedures, sometimes not, as in the Lasso. In any case, the user is left with the problem of calibrating this hyperparameter.

Several goals are conceivable for a hyperparameter selection method, such as support recovery or estimation of a "true" underlying regression coefficient. From a prediction perspective, hyperparameters should be chosen so as to minimize the risk, and a good method should approach this minimum. As a consequence, the proposed data-driven choice of hyperparameter should allow the estimator to attain all known convergence rates without any a priori knowledge, effectively adapting to the difficulty of the problem.

For the Lasso and some variants, such as the fused Lasso, Zou, Wang, Tibshirani and coauthors have proposed [39] and investigated [32, 28] a method based on Mallow's C_p and estimation of the "degrees of freedom of the Lasso". However, consistency of this method has only been proven [32] in an asymptotic where the dimension is fixed while n grows, hence not the setting considered here. Moreover, the method depends on specific properties of the Lasso, and may not be readily applicable to other sparse regression procedures.

A much more widely applicable procedure is to choose the hyperparameter by cross-validation. For the Lasso, this approach has been recommended by Tibshirani [26], van de Geer and Lederer [29] and Greenshtein [10], among many others. More generally, cross-validation is the default method for calibrating hyperparameters in practice. For exemple, R implementations of the elastic net (package glmnet), LARS (package lars) and the huberized lasso (package hqreg) all incorporate a cross-validation subroutine to automatically choose the hyperparameter.

Theoretically, cross-validation has been shown to perform well in a variety of settings [1]. For cross-validation with one split, also known as the hold-out, and for a bagged variant of v-fold cross-validation [16], some general oracle inequalities are available in least squares regression [18, Corollary 8.8] [36] [16]. However, they rely on uniform boundedness assumptions on the estimators which may not hold in high-dimensional linear regression. For the more popular V-fold procedure, results are only available in specific settings. Of particular interest here is the article [21] which proves oracle inequalities for linear model selection in least squares regression, since linear model selection is very similar to sparse regression (the main difference being that in sparse regression, the "models" are not fixed a priori but depend on the data). This suggests that similar results could hold for sparse regression.

However, in the case of the Lasso at least, no such theoretical guarantees exist, to the best of my knowledge. Some oracle inequalities [16, 20] and also fast rates [11, Theorem 1] have been obtained, but only under very strong assumptions: [16] assumes that X is log-concave, [20] that X is a gaussian vector, and [11, Theorem 1] assumes that there is a true model and that the variance-covariance matrix is diagonal dominant. In contrast, there are also theorems [6, 8] [11, Theorem 2] which make much weaker distributional assumptions but only prove convergence of the risk at the "slow" rate $\mathcal{O}(\sqrt{\frac{s_* \log p}{n}})$ or slower. Though this rate is minimax [6], a hyperparameter selection method should

adapt also to the favorable cases where the Lasso converges faster; these results do not show that CV has this property.

Thus, the theoretical justification for the use of standard CV in sparse regression is somewhat lacking. In fact, two of the articles mentioned above do not study standard CV applied to the Lasso but introduce a variant; a bagged CV in [16] and the aggregation of two hold-out predictors in [6]. In practice too, there is reason to consider alternatives to CV-based hyperparameter selection in sparse regression: sparse estimators are unstable, and selecting only one estimator can result in arbitrarily ignoring certain variables among a correlated group with similar predictive power [37]. For the Lasso, these difficulties have motivated researchers to introduce several aggregation schemes, such as the Bolasso [3], stability selection [19], the lasso-zero [9] and the random lasso [34], which are shown to have some better properties than the standard Lasso.

Since aggregating the Lasso seems to be advantageous, it seems logical to consider aggregation rather than cross-validation to handle the free hyperparameters. In this article, I consider the application to sparse regression of the aggregated hold-out procedure. Aggregated hold-out (agghoo) is a general aggregation method which mixes cross-validation with bagging. It is an alternative to cross-validation, with a comparable level of generality. In a previous article with Sylvain Arlot and Matthieu Lerasle [17], we formally defined and studied Agghoo, and showed empirically that it can improve on cross-validation when calibrating the level of regularization for kernel regression. Though we came up with the name and the general mathematical definition, Agghoo has already appeared in the applied litterature in combination with sparse regression procedures [12], among others [31], under the name "CV + averaging" in this case.

In the present article, the aim is to study the application of Agghoo to sparse regression with a robust loss function. Theoretically, assuming an $L^\infty - L^2$ norm inequality to hold on the set of sparse linear predictors, it is proven that Agghoo satisfies an asymptotically optimal oracle inequality. This result applies also to cross-validation with one split (the so-called hold-out), yielding a new oracle inequality which allows norms of the sparse linear predictors to grow polynomially with the sample size. Empirically, Agghoo is compared to cross-validation in a number of simulations, which investigate the impact of correlations in the design matrix and sparsity of the ground truth on the performance of aggregated hold-out and cross-validation. Agghoo appears to perform better than cross-validation when the number of non-zero coefficients to be estimated is not much smaller than the sample size. The presence of confounders correlated to the predictive variables also favours Agghoo relative to cross-validation.

2 Setting and Definitions

The problem of non-parametric regression is to infer a predictor $t : \mathcal{X} \to \mathbb{R}$ from a dataset $(X_i, Y_i)_{1 \le i \le n}$ of pairs, where $X_i \in \mathcal{X}$ and $Y_i \in \mathbb{R}$. The pairs will be assumed to be i.i.d, with joint distribution P. The prediction error made at a point $(x, y) \in \mathcal{X} \times \mathbb{R}$ is measured using a non-negative function of the residual $\phi(y - t(x))$. The global performance of a predictor is assessed on a new, independent data point (X, Y) drawn from the same distribution P using the risk $\mathcal{L}(t) = \mathbb{E}[\phi(Y - t(X))]$. The optimal predictors s are characterized by $s(x) \in \operatorname{argmin}_u \mathbb{E}[\phi(Y - u)|X = x]$ a.s. The risk of any optimal predictor is (in general)

a non-zero quantity which characterizes the intrinsic amount of "noise" in Y unaccounted for by the knowledge of X. A predictor t can be compared with this benchmark by using the *excess risk* $\ell(s,t) = \mathcal{L}(t) - \mathcal{L}(s)$. Taking $\phi(x) = x^2$ yields the usual least-squares regression, where $s(x) = \mathbb{E}[Y|X = x]$ and $\ell(s,t) = \|s - t\|^2_{L^2(X)}$. However, the least-squares approach is known to suffer from a lack of robustness. For this reason, in the field of robust statistics, a number of alternative loss functions are used. One popular choice was introduced by Huber [13].

Definition 2.1 *Let $c > 0$. Huber's loss function is $\phi_c(u) = \frac{u^2}{2}\mathbb{I}_{|u|\leq c} + c\left(|u| - \frac{c}{2}\right)\mathbb{I}_{|u|>c}$.*

When $c \to +\infty$, ϕ_c converges to the least-squares loss. When $c \to 0$, $\frac{1}{c}\phi_c$ converges to the absolute value loss $x \to |x|$ of median regression. Thus, the c parameter allows a trade-off between robustness and approximation of the least squares loss.

The rest of the article will focus on sparse linear regression with the loss function ϕ_c. Thus, notations s, $\ell(s,t)$ and $\mathcal{L}$ are to be understood with respect to ϕ_c.

2.1 Sparse linear regression

With finite data, it is impossible to solve the optimization problem $\min \mathcal{L}(t)$ over the set of all predictors t. Some modeling assumptions must be made to make the problem tractable. A popular approach is to build a finite set of features $(\psi_j(X))_{1\leq j\leq d}$ and consider predictors that are linear in these features: $\exists \theta \in \mathbb{R}^d, \forall x \in \mathcal{X}, t(x) = \sum_{j=1}^d \theta_j \psi_j(x)$. This is equivalent to replacing $X \in \mathcal{X}$ with $\tilde{X} = (\psi_j(X))_{1\leq j\leq d} \in \mathbb{R}^d$ and regressing Y on $\tilde{X}$. For theoretical purposes, it is thus equivalent to assume that $\mathcal{X} = \mathbb{R}^d$ for some d and predictors are linear: $t(x) = \theta^T x$.

As the aim is to reduce the average prediction error $\mathcal{L}(t)$, a logical way to choose θ is by *empirical risk minimization*:

$$\hat{\theta} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \phi_c(Y_i - \theta^T X_i).$$

Empirical risk minimization works well when $d << n$ but will lead to overfitting in large dimensions [30]. Sparse regression attempts instead to locate a "good" subset of variables in order to optimize risk for a given model dimension. Lasso penalization [26] is now a standard method of achieving sparsity. The specific version of the Lasso which we consider here is given by the following Definition.

Definition 2.2 *Let $n \in \mathbb{N}$ and let $D_n = (X_i, Y_i)_{1\leq i\leq n}$ be a dataset such that $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$ for all $i \in [|1;n|]$ and some $d \in \mathbb{N}$. Let ϕ_c be the huber loss defined in Definition 2.1. For any $\lambda > 0$, let*

$$\hat{\mathcal{C}}(\lambda) = \operatorname*{argmin}_{(q,\theta)\in\mathbb{R}^{d+1}:\|\theta\|_1\leq n^\alpha} \frac{1}{n}\sum_{i=1}^n \phi_c\big(Y_i - q - \theta^T X_i\big) + \lambda\|\theta\|_1 \quad \textit{and}$$

$$(\hat{q}(\lambda), \hat{\theta}(\lambda)) \in \operatorname*{argmin}_{(q,\theta)\in\hat{\mathcal{C}}(\lambda)} \Big|q + < \theta, \frac{1}{n}\sum_{i=1}^n X_i >\Big|. \tag{1}$$

Now let

$$\mathcal{A}(\lambda)(D_n) : x \to \hat{q}(\lambda) + \hat{\theta}(\lambda)^T x.$$

The restriction $\|\theta\|_1 \leq n^\alpha$ ensures that the solution does not become too large when the design matrix is ill-conditioned. It can be seen that the effect of this restriction is, potentially, to truncate the Lasso solution path at the value of λ at which the bound is attained (i.e, $\hat{\theta}(\cdot)$ becomes constant for smaller values of λ). Without this, the ℓ^1 norm of Lasso solutions is upper bounded by the minimal ℓ^1 norm $\hat{c}$ of an empirical risk minimizer on the whole set of variables. Hence, if $\hat{c} \leq n^\alpha$, Definition 2.2 coïncides with the huberized lasso along the whole regularization path. In the least squares case, [11] discuss conditions under which $\|\hat{c}\|_{L^4} \leq n^{\frac{1}{4}}$, which ensures that for $\alpha > \frac{1}{4}$, $\hat{c} < n^\alpha$ with high probability. It seems reasonable to expect a similar result to hold true in the case of the huberized lasso. However, rather than make further technical assumptions on the design to make sure of this, it seems simpler to introduce this slight modification to the standard definition of the huberized lasso, which may even be statistically beneficial, since it diminishes the variance by restricting the hypothesis space.

A suitable choice of α should guarantee that an optimal excess risk $\mathbb{E}\left[\ell(s, q + \langle\theta, \cdot\rangle)\right]$ can be obtained for some θ such that $\|\theta\|_1 \leq n^\alpha$. For example, if the features X form an orthonormal set and $Y \in L^2$, then the least-squares optimal coefficient θ_* belongs to $\left\{\theta : \|\theta\|_2 \leq \sqrt{\mathbb{E}[Y^2]}\right\}$. Assume that only sparse predictors θ^T, with less than n non-zero components, are considered. Since for such θ, $\|\theta\|_1 \leq \sqrt{n}\,\|\theta\|_2$, it is reasonable to restrict the optimization to the set $\{\theta : \|\theta\|_1 \leq n^\alpha\}$ for some $\alpha > \frac{1}{2}$.

The intercept q is left unpenalized in definition 2.2, as is usually the case in practice [38]. Equation (1) is a tiebreaking rule which is required for the proof to work.

2.2 Hyperparameter tuning

The zero-norm of a vector θ is the integer $\|\theta\|_0 = |\{i : \theta_i \neq 0\}|$. Many sparse estimators, such as best subset or forward stagewise, are directly parametrized by their desired zero-norm, which must be chosen by the practitioner. It controls the "complexity" of the estimator, and hence the bias-variance tradeoff. In the case of the standard Lasso (Definition 2.2 with $\phi(x) = x^2$), Zou, Hastie and Tibshirani [39] showed that $\left\|\hat{\theta}(\lambda)\right\|_0$ is an unbiased estimator of the "degrees of freedom" of the estimator $\mathcal{A}(\lambda)$. As a consequence, [39] suggests reparametrizing the lasso by its zero-norm. Applying their definition to the present setting yields the following.

Definition 2.3 *For any dataset D_n, let $(\hat{q}, \hat{\theta})$ be given by Definition 2.2, equation* (1) *. Let $M \in \mathbb{N}$ and $(\lambda_m)_{1 \leq m \leq M}$ be the finite decreasing sequence at which the sets $\{i : \hat{\theta}(\lambda)_i \neq 0\}$ change. Let $\lambda_0 = +\infty$. For any $k \in \mathbb{N}$ let*

$$\hat{m}_k^{last} = \max\left\{m \in \mathbb{N} | \left\|\hat{\theta}(\lambda_m)\right\|_0 = k\right\},$$

with the convention $\max \emptyset = 0$. *Let then*

$$\mathcal{A}_k(D_n) = \mathcal{A}\left(\lambda_{\hat{m}_k^{last}}\right)(D_n). \tag{2}$$

More generally, consider sequences $(\mathcal{A}_k)_{k\in\mathbb{N}}$ of linear regression estimators $\mathcal{A}_k : D_n \to \left(x \to \hat{q}_k(D_n) + \langle \hat{\theta}_k(D_n), x\rangle\right)$, such that the following hypothesis holds.

Hypothesis 2.1 *For any $n \in \mathbb{N}$, let $D_n \sim P^{\otimes n}$ denote a dataset of size n. Assume that*

1. *Almost surely, for all $k \in [|1;n|]$, $\left\|\hat{\theta}_k(D_n)\right\|_0 \leq k$.*
2. *There exist L, α such that $\forall n \in \mathbb{N}, \mathbb{E}\left[\sup_{1\leq k\leq n}\left\|\hat{\theta}_k(D_n)\right\|_1\right] \leq Ln^\alpha$.*
3. *For all $k \in [|1;n|]$, $\hat{q}_k(D_n) \in \operatorname{argmin}_{q\in\hat{Q}\left(D_n,\hat{\theta}_k(D_n)\right)}\left|q + \langle\hat{\theta}_k(D_n), \frac{1}{n}\sum_{i=1}^n X_i\rangle\right|$, where $\hat{Q}(D_n,\theta) = \operatorname{argmin}_{q\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n \phi_c\left(Y_i - \langle\theta, X_i\rangle - q\right)$.*

These hypotheses hold for the reparametrized Lasso given by definition 2.2 and 2.3, by construction.

Moreover, Condition 1. is naturally satisfied by such sparse regression methods as forward stepwise and best subset. Condition 2 can be enforced by restricting the set of θs over which the optimization is conducted, similarly to Definition 2.2. Condition 3 states that the intercept q is chosen by empirical risk minimization, with a specific tie-breaking rule in case the minimum is not unique.

2.3 Aggregated hold out applied to the zero-norm parameter

The tuning of the zero-norm k is important to ensure good prediction performance by optimizing the bias-variance tradeoff. For the Lasso and other methods based on empirical risk minimization, such as forward stepwise, there is little interest in considering values of $k > n$, since n non-zero coefficients suffice for perfect interpolation of the (X_i, Y_i) and yield an empirical risk of 0. Practicioners may also want to impose additional limitations on the zero-norm in order to reduce the computational load or improve interpretability. For this reason, we consider the problem of selecting the zero-norm among the K_n first values, where $K_n \leq n$. This article investigates the use of Agghoo in this context, as an alternative to cross-validation. Agghoo is a general hyperparameter aggregation method which was defined in [17], in a general statistical learning context. Let us briefly recall its definition in the present setting. For a more detailed introductory discussion of this procedure, we refer the reader to [17]. To simplify notations, fix a collection $(\hat{q}_k, \hat{\theta}_k)_{1\leq k\leq K}$ of linear regression estimators. First, we need to define *hold-out* selection of the zero-norm parameter.

Definition 2.4 *Let $D_n = (X_i, Y_i)_{1\leq i\leq n}$ be a dataset. For any $T \subset \{1,\dots,n\}$, denote $D_n^T = (X_i, Y_i)_{i\in T}$. Let then*

$$\hat{k}_T(D_n) = \min \operatorname*{argmin}_{1\leq k\leq K} \frac{1}{|T^c|}\sum_{i\notin T}\phi_c\left(Y_i - \hat{q}_k(D_n^T) - \langle\hat{\theta}_k(D_n^T), X_i\rangle\right).$$

Using the hyperparameter $\hat{k}_T(D_n)$ together with the dataset D_n^T to train a linear regressor yields the hold-out predictor

$$\widehat{f}_T^{\mathrm{ho}}(D_n) : x \to \hat{q}_{\hat{k}_T(D_n)}(D_n^T) + \langle\hat{\theta}_{\hat{k}_T(D_n)}(D_n^T), x\rangle.$$

Aggregation of hold-out predictors is performed in the following manner.

Definition 2.5 *Let $\mathcal{T} \subset \mathcal{P}(\{1, \ldots, n\})$. Let:*

$$\hat{\theta}^{ag}_{\mathcal{T}} = \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} \hat{\theta}_{\hat{k}_T(D_n)}(D^T_n)$$
$$\hat{q}^{ag}_{\mathcal{T}} = \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} \hat{q}_{\hat{k}_T(D_n)}(D^T_n).$$

The Agghoo predictor is the linear regressor:

$$\widehat{f}^{\mathrm{ag}}_{\mathcal{T}}(D_n) : x \to \hat{q}^{ag}_{\mathcal{T}} + \langle \hat{\theta}^{ag}_{\mathcal{T}}, x \rangle.$$

Thus, Agghoo also yields a linear predictor, which means that it can be efficiently evaluated on new data. If the $\hat{\theta}_{\hat{k}_T(D_n)}$ have similar support, $\hat{\theta}^{ag}_{\mathcal{T}}$ will also be sparse: this will happen if the hold-out reliably identifies a true model. On the other hand, if the supports have little overlap, the Agghoo coefficient will lose sparsity, but it can be expected to be more stable and to perform better.

The linear regressors $x \to \hat{q}_{\hat{k}_T(D_n)}(D^T_n) + \langle \hat{\theta}_{\hat{k}_T(D_n)}(D^T_n), x \rangle$ aggregated by Agghoo are only trained on part of the data. This subsampling (typically) decreases the performance of each individual estimator, but combined with aggregation, it may stabilize an unstable procedure and improve its performance, similarly to bagging.

An alternative would be to *retrain* each regressor on the whole data-set D_n, yielding the following procedure, which we call "Aggregated cross-validation" (Agcv).

Definition 2.6 *Let $\mathcal{T} \subset \mathcal{P}(\{1, \ldots, n\})$. Let:*

$$\hat{\theta}^{acv}_{\mathcal{T}} = \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} \hat{\theta}_{\hat{k}_T(D_n)}(D_n)$$
$$\hat{q}^{acv}_{\mathcal{T}} = \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} \hat{q}_{\hat{k}_T(D_n)}(D_n).$$

The Agcv predictor is the linear regressor:

$$\widehat{f}^{\mathrm{acv}}_{\mathcal{T}}(D_n) : x \to \hat{q}^{acv}_{\mathcal{T}} + \langle \hat{\theta}^{acv}_{\mathcal{T}}, x \rangle.$$

Agghoo is easier to study theoretically than Agcv due to the conditional independence: $\left(\hat{\theta}_k(D^T_n)\right)_{1 \le k \le n_t} \perp\!\!\!\perp \hat{k}_T(D_n) \,\Big|\, D^T_n$. For this reason, the theoretical section will focus on Agghoo, while in the simulation study, both Agghoo and Agcv will be considered.

3 Theoretical results

Let $n \in \mathbb{N}$ and $D_n = (X_i, Y_i)_{1 \le i \le n}$ denote an i.i.d dataset with common distribution P. In this section, we make the following assumption on $\mathcal{T}$: there is an integer $n_t < n$ such that

$$\begin{aligned} &\mathcal{T} \subset \{T \subset \{1, \ldots, n\} : |T| = n_t\} \\ &\mathcal{T} \text{ independent from } D_n \ . \end{aligned} \tag{3}$$

Independence of $\mathcal{T}$ from D_n ensures that for $T \in \mathcal{T}$, D_n^T is also iid with distribution P. The assumption that $\mathcal{T}$ contain sets of equal size ensures that the pairs $\hat{q}_{\hat{k}_T(D_n)}(D_n^T), \hat{\theta}_{\hat{k}_T(D_n)}(D_n^T)$ are equidistributed for $T \in \mathcal{T}$. Most of the data partitioning procedures used for cross-validation satisfy hypothesis (3), including leave-p-out, V-fold cross-validation (with $n - n_t = n_v = n/V$) and Monte-Carlo cross-validation [1].

In the following, we will use the notion of support of a random variable, for which we introduce the following definition.

Definition 3.1 *Let X be a random variable belonging to $\mathbb{R}^d$ for some $d \in \mathbb{N}$. Then the support of X is*

$$supp(X) = \{x \in \mathbb{R}^d : \forall \varepsilon > 0, \mathbb{P}\left(\|x - X\| \leq \varepsilon\right) > 0.\}$$

The support is closed and has full measure: $\mathbb{P}(X \in supp(X)) = 1$.

When Agghoo is used on a collection $(\mathcal{A}_k)_{1 \leq k \leq K}$ of linear regression estimators satisfying Hypothesis (2.1), such as the Lasso parametrized by the number of non-zero coefficients, as in Definition 2.3, the following Theorem applies.

Theorem 3.2 *Let $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ be random variables with joint distribution P. Let $D_n = (X_i, Y_i)_{1 \leq i \leq n} \sim P^{\otimes n}$ be a dataset of size n. Let $n_v = n - n_t$, where n_t is given by assumption* (3).

Assume that for a bayes regressor s there exists $\eta > 0$ such that

$$\forall x, \mathbb{P}\left[|Y - s(x)| \leq \frac{c}{2} \,\middle|\, X = x\right] \geq \eta. \tag{4}$$

Let $\bar{X} = X - \mathbb{E}[X]$ and let $supp(\bar{X})$ be its support (in the sense of Definition 3.1). Let $R = \sup_{x \in supp(\bar{X})} \|x\|_\infty$. For any $K \in \{1, \ldots, n_t\}$, let

$$\kappa(K) = \sup_{\theta \neq 0, \|\theta\|_0 \leq 2K} \frac{\left\|\langle \bar{X}, \theta \rangle\right\|_{L^\infty}}{\left\|\langle \bar{X}, \theta \rangle\right\|_{L^2}}. \tag{5}$$

If $b_0 > 1$ and $K \in \{1, \ldots, n_t\}$ are such that

$$\kappa(K) \leq \frac{\eta}{8} \sqrt{\frac{n_v}{8 b_0 \log n_t}}, \tag{6}$$

applying Agghoo to a collection $(\mathcal{A}_k)_{1 \leq k \leq K}$ of linear regression estimators which satisfies hypothesis (2.1) *yields the following oracle inequality.*

For any $\theta \in \left[\frac{1}{\sqrt{b_0}}; 1\right]$,

$$(1-\theta)\mathbb{E}\left[\ell(s, \widehat{f}_{\mathcal{T}}^{\text{ag}})\right] \leq (1+\theta)\mathbb{E}\left[\min_{1 \leq k \leq K} \ell(s, \mathcal{A}_k(D_{n_t}))\right] + 24\theta b_0 \frac{c \log n_t}{\eta n_v} \left[c + \frac{cK}{n_t^{\theta^2 b_0}} + \frac{16KLR}{n_t^{\theta^2 b_0 - \alpha}}\right]. \tag{7}$$

Theorem 3.2 is proved in appendix A. It is, to the best of my knowledge, the first theoretical guarantee on hyperparameter selection for the huberized Lasso. Theorem 3.2 compares the excess risk of Agghoo to that of the best linear predictor in the collection $\mathcal{A}_k(D_{n_t})$, trained on a subset of the data of

size n_t. That n_t appears in the oracle instead of n is a limitation, but it is logical, since estimators aggregated by Agghoo are only trained on samples of size n_t. Typically, the excess risk increases at most by a constant factor when a dataset of size n is replaced by a subset of size τn, and this constant tends to 1 as $\tau \to 1$. This allows to take n_v of order n ($n_v = (1-\tau)n$), while losing only a constant factor in the oracle term.

Taking $|\mathcal{T}| = 1$ in Theorem 3.2 yields an oracle inequality for the hold-out, which is also cross-validation with one split. Compared to previously known oracle inequalities for the hold-out, Theorem 3.2 distinguishes itself by only requiring some polynomial upper bound on $\|\theta\|_1$ and $\|\bar{X}\|_\infty$, instead of a uniform upper bound on some norm independent of n. Indeed, the prevailing approach to proving oracle inequalities for the hold-out (applied to the Lasso by Lecué [16]) uses a *margin assumption* which requires a uniform upper bound on the loss function [16, Assumption (A)], leading to a bound on $\langle \hat{\theta}_k, \bar{X} \rangle$. Theorem 3.2 relaxes this constraint by exploiting an $L^\infty - L^2$ norm inequality (equation(6)).

In order to fulfill its purpose, Theorem 3.2 should imply that Agghoo performs as well as the best of the sparse estimators $\mathcal{A}_k(D_{n_t})$, at least asymptotically. Here, we are interested in the high-dimensional, non-parametric case where the dimension grows with the amount of data n as a power of n. More precisely, consider a sequence of problems $(Y, \psi_n(X_0))$ where $X_0 \in \mathcal{X}$ and $\psi_n : \mathcal{X} \to \mathbb{R}^{d_n}$, where $d_n > n^\beta$ for some $\beta > 0$. Assume that $R = 1$, which can be achieved by renormalizing ψ_n - as long as $\|\psi_n(X_0)\|_{\infty, L^\infty}$ grows at most polynomially in n, this simply yields an increase in L and α. If additionally, equation (6) holds with $b_0 > \alpha + 1$, choosing $\theta \in \left(\sqrt{\frac{1+\alpha}{b_0}}; 1\right)$ yields a remainder term of order $\mathcal{O}(\frac{\log n}{n})$ in equation (7). By comparison, in the least squares setting the minimax excess risk for sparse regression with k_n predictive covariates among a total of d_n is of order $\frac{k_n \log(\frac{d_n}{k_n})}{n}$ [35]. For large c, the huber loss approximates the least squares loss, so it is reasonable to expect this lower bound to apply also in huber regression. Assuming that the minimax is attained, the remainder term of equation (7) is negligible compared to the oracle whenever $k_n \to +\infty$, i.e when the problem is non-parametric.

Now if for any n, assumption (6) holds with $X = \psi_n(X_0)$ and $b_0 = b_{1,n} \to +\infty$, then Theorem 3.2 yields an asymptotically optimal oracle inequality. More precisely, applying Theorem 3.2 with $b_0 = b_{0,n} = b_{1,n} \wedge \sqrt{\frac{k_n}{1+\alpha}}$ (for which assumption (6) also holds) and $\theta = \theta_n = \sqrt{\frac{1+\alpha}{b_{0,n}}} \wedge 1$ yields a bounded term in the square brackets of equation (7). Moreover, $\theta_n b_{0,n} \leq \sqrt{k_n}$. This implies that

$$\limsup_{n \to +\infty} \frac{\mathbb{E}\left[\ell(s, \widehat{f}^{\mathrm{ag}}_{\mathcal{T}})\right]}{\mathbb{E}\left[\min_{1 \leq k \leq K_{n_t}} \ell(s, \mathcal{A}_k(D_{n_t}))\right]} \leq 1$$

since $\frac{\sqrt{k_n} \log n}{n} = o\left(\mathbb{E}\left[\min_{1 \leq k \leq K_{n_t}} \ell(s, \mathcal{A}_k(D_{n_t}))\right]\right)$.

To summarize, Theorem 3.2 proves an asymptotically optimal oracle inequality whenever

- Equation (4) holds.
- $\|\|\psi_n(X_0)\|_\infty\|_{L^\infty}$ grows at most polynomially in n.

- The problem is non-parametric, i.e the risk converges slower than $\frac{\log n}{n}$.
- For any $b_0 > 1$, equation (6) holds for all n large enough.

Equation (4) is specific to the Huber loss: it requires the conditional distribution of the residual $Y - s(x)$ to put sufficient mass in a region where the huber function is quadratic. If $Y = s_0(X_0) + \sigma\varepsilon$, with ε independent of X_0 and if ψ_n is injective, then $s(X) = s_0(X_0)$ and η depends only on c, σ and the distribution of ε. In particular, it is constant with respect to n. Moreover, if the Huber parameter c is proportional to σ, then η is independent also of σ and the remainder term of equation (7) is proportional to σ^2, as in least-squares regression. Injectivity of ψ_n will typically hold for nonparametric function bases (trigonometric, splines of degree greater than 1) as soon as n is large enough.

The norm inequality (Equation (6)) requires more clarification. It is worth giving some background on such hypotheses, which are relatively classical in the model selection litterature. They were introduced by Birgé and Massart in the context of least-squares density estimation [4, Section 3.1], where loosely speaking, it is assumed that $\kappa(K) = \mathcal{O}(\sqrt{K})$. A similar assumption was made by Arlot and Lerasle [2, Section 3.3, hypothesis (H1)] to prove oracle inequalities for cross-validation, also in least squares density estimation. In the regression setting, [22] proves an oracle inequality for cross-validation based on the assumption that the models have a "strongly localized basis", which implies in particular that $\kappa(K) = \mathcal{O}(\sqrt{K})$ when the model collection consists of all $\left\{\langle\theta, (\bar{X}_i)_{i\in I}\rangle : \theta \in \mathbb{R}^{|I|}\right\}$ for $I \subset [|1; n|]$.

These assumptions have been shown to hold for several standard model collections. In particular, in the regression setting, [25, Lemma 7] implies that linear models m consisting of piecewise-polynomial functions on an interval partition $(I_i)_{1\leq i\leq d}$ satisfy $\|\cdot\|_{L^\infty(X)} \leq C\sqrt{d}\,\|\cdot\|_{L^2(X)}$, provided that $\min_i \mathbb{P}(X \in I_i) \geq \frac{a}{d}$ for some constant a and that the distribution of X has a lower-bounded density with respect to the Lebesgue measure.

The assumption (6) differs from its analogs in the model selection litterature in two ways: first, because of the sparse variable selection setting, the "models" which give rise to $\kappa(K)$ are the $\left\{\langle\theta, (\bar{X}_i)_{i\in I}\rangle : \theta \in \mathbb{R}^{|I|}\right\}$ for $I \subset [|1; d|]$ of cardinality $|I| = K$. Second, because an additional intercept term is included, the feature vector X has to be replaced by its centered version $\bar{X}$ in the definition of $\kappa(K)$.

We give below two simple examples where the hypotheses of Theorem 6 hold. In the case where the variables are independent and binary valued, we have the following.

Corollary 3.3 *Let $X = (X_1, \ldots, X_p)$ where the X_i are independent Bernoulli random variables with parameters $p_i \in [0; 1]$ Then*

$$\kappa(K) \leq \sqrt{\frac{2K}{\min_{1\leq i\leq p} p_i(1-p_i)}}.$$

Corollary 3.3 is proved in appendix B.1. In the setting it describes, Assumption (6) is equivalent to choosing K of order $\frac{n_t}{\log n_t}$ or less, provided that the classes are well-balanced ($p_i \in [\varepsilon, 1-\varepsilon]$).

Despite the fact that Theorem 3.2 is formulated in the setting of sparse linear regression, it can also be applied to other regression problems, such as adaptive

piecewise constant regression. In that case, the equivalent of the zero-norm of a vector is the number of discontinuities of a piecewise constant function, as can be seen from the following definition.

Definition 3.4 *Let $(I_j)_{1\leq j\leq d}$ denote a partition of $\mathbb{R}$ into disjoint intervals, indexed such that $\forall j, \sup I_j = \inf I_{j+1}$. For any $u \in \mathbb{R}^d$, let $t_u = \sum_j u_j \mathbb{I}_{I_j}$. Then the number of jumps of the piecewise constant function t_u is*

$$k(u) = |\{j \in [|1;d|] : u_{j+1} \neq u_j\}|,$$

and we say that t_u has k jumps *if and only if $k(u) = k$. Let $(j_r(u))_{0\leq r\leq k(u)}$ denote the ordered sequence of jump indices, i.e $(j_r(u))_{1\leq r\leq k(u)}$ is increasing, $j_0(u) = 0$ and*

$$\{j_r(u) | 1 \leq r \leq k(u)\} = \{j \in [|1;d|] : u_{j+1} \neq u_j\}.$$

For any $r \leq k(u)$, let $A_r(u) = \cup_{i=j_{r-1}(u)+1}^{j_r(u)} I_i$ be the largest intervals on which t_u is constant. Let $D_n = (U_i, Y_i)_{1\leq i\leq n}$ denote a dataset, with $U_i \in \mathbb{R}$ and $Y_i \in \mathbb{R}$.

Let now $(\hat{u}_k)_{0\leq k\leq d-1}$ denote a sequence of estimators such that $\hat{u}_k(D_n)$ has k jumps, and such that its coefficients $\hat{u}_{k,j}$ are obtained by empirical risk minimization on the minimal partition $(A_r(\hat{u}_k))_{1\leq r\leq k}$, i.e $\hat{u}_k \in \hat{C}_t(D_n, \hat{u}_k)$ where

$$\hat{C}_t(D_n, u) = \Big\{u' \text{ s.t. } k(u') = k \text{ and } \forall r \in [|1;k|], j_r(u') = j_r(u) \\ \text{and } u'_{j_r(u')} \in \operatorname*{argmin}_{q\in\mathbb{R}} \sum_{i:X_i\in A_r(u)} \phi_c(Y_i - q)\Big\},$$

Assume also that the following tie-breaking rule applies:

$$\hat{u}_k \in \operatorname*{argmin}_{u\in\hat{C}_t(D_n,\hat{u}_k)} \left|\sum_{j=1}^{d} P_n(I_j)u_j\right|. \tag{8}$$

Estimators $\hat{u}_k$ which meet definition 3.4 can be obtained by a variety of model selection methods, including wavelet thresholding, empirical risk minimization over the set $\{u : k(u) = k\}$ [15] and the fused lasso [27] or total variation penalties [24, 5] (if the penalty is used only for estimating the change points).

Applied to such estimators, Theorem 3.2 allows to prove the following.

Proposition 3.5 *Consider the problem of tuning k so as to minimize the risk of the one-dimensional regression problem: $\mathbb{E}[\phi_c(Y - t_u(U))]$, where U is a random variable with distribution P. Assume that*

$$\forall x, \mathbb{P}\left[|Y - t_{u_*}(x)| \leq \frac{c}{2} \,\middle|\, X = x\right] \geq \eta, \tag{9}$$

where $u_* \in \operatorname{argmin}_{u\in\mathbb{R}^d} \mathbb{E}[\phi_c(Y - t_u(X))]$. Using the notations of definition 3.4, assume that:

$$\min_{1\leq j\leq d} P(I_j) \geq \frac{1536 \log^2 n_t}{\eta^2 n_v}. \tag{10}$$

Then, assuming that $n_t \geq n_v$,

$$\left(1 - \frac{1}{\sqrt{\log n_t}}\right) \mathbb{E}\left[\ell\big(t_{u_*}, \widehat{f}^{\mathrm{ag}}_{\mathcal{T}}\big)\right] \leq \left(1 + \frac{1}{\sqrt{\log n_t}}\right) \mathbb{E}\left[\inf_{1\leq k\leq d} \ell\big(t_{u_*}, t_{\hat{u}_k(D_{n_t})}\big)\right] + 72\frac{c\log^{\frac{3}{2}} n_t}{\eta n_v}\left[c + \frac{3c}{n_t} + \frac{4\mathbb{E}[|Y|]}{n_t}\right]. \tag{11}$$

Proposition 3.5 is proved in appendix B.2. The specific setting of Proposition 3.5 allows to state a fairly explicit oracle inequality for Agghoo under few conditions. Assuming as before that n_t and n_v are both of order n, the remainder term in equation (3.5) is of order $\frac{\log^{\frac{3}{2}} n}{n}$. This is negligible compared to the minimax rates achievable under regularity assumptions (eg. Hölder or Besov balls), which are of order $n^{-\alpha}$ with $\alpha \in (0; \frac{2}{3}]$. Hence, Proposition 3.5 shows that Agghoo adapts to the unknown level of regularity, achieving the correct convergence rate.

Three assumptions are made to obtain this oracle inequality. The distributional assumption on the residuals (equation (9)) is identical to the one made in Theorem 3.2, and has already been discussed. The moment condition $\mathbb{E}[|Y|] < +\infty$, without which inequality (11) becomes vacuous, is self-explanatory.

Finally, hypothesis (10) requires that the intervals of the partition contain at least $cst \times \log^2 n_t$ points, on average. This is a mild requirement, since partitions finer than this cannot be expected to perform well anyway due to high variance of the empirical average.

Though hypothesis (10) involves the unknown distribution P, Bernstein's inequality shows that if $\min_{1\leq j\leq d} P_{n_t}(I_j) \geq \frac{C\log^2 n_t}{\eta^2 n_v}$, where $C > 1536$ and P_{n_t} denotes the empirical measure on a sample of size n_t, then equation (10) holds with high probability. Thus, provided a lower bound on η is known, it is possible to guarante empirically that (10) holds, with a high degree of confidence.

3.1 Effect of V

The upper bound given by Theorem 3.2 only depends on $\mathcal{T}$ through n_v and n_t. The purpose of this section is to show that for a given value of n_v, increasing $V = |\mathcal{T}|$ always decreases the risk. This is proved in the case of monte carlo subset generation defined below.

Definition 3.6 *For $\tau \in \left[\frac{1}{n}; 1\right]$ and $V \in \mathbb{N}^*$, let $\mathcal{T}^{mc}_{\tau,V}$ be generated independently of the data D_n by drawing V elements independently and uniformly in the set*

$$\{T \subset [|1; n|] : |T| = \lfloor \tau n \rfloor\}.$$

For fixed τ, the excess risk of Agghoo is a non-increasing function of V.

Proposition 3.7 *Let $U \leq V$ be two non-zero integers. Let $\tau \in \left[\frac{1}{n}; 1\right]$. Then:*

$$\mathbb{E}\left[\ell(s, \widehat{f}^{\mathrm{ag}}_{\mathcal{T}^{mc}_{\tau,V}})\right] \leq \mathbb{E}\left[\ell(s, \widehat{f}^{\mathrm{ag}}_{\mathcal{T}^{mc}_{\tau,U}})\right].$$

Proof Let $(T_i)_{i=1,\dots,V} = \mathcal{T}^{mc}_{\tau,U}$. Let $\mathcal{I} = \{I \subset [|1;V|] : |I| = U\}$. Then

$$\begin{aligned}
\widehat{f}^{\text{ag}}_{\mathcal{T}^{mc}_{\tau,U}} &= \sum_{i=1}^{V} \frac{1}{V} \widehat{f}^{\text{ho}}_{T_i} \\
&= \sum_{i=1}^{V} \frac{\binom{V-1}{U-1}}{U\binom{V}{U}} \widehat{f}^{\text{ho}}_{T_i} \\
&= \frac{1}{U} \sum_{i=1}^{V} \frac{\sum_{I\in\mathcal{I}} \mathbb{I}_{i\in I}}{|\mathcal{I}|} \widehat{f}^{\text{ho}}_{T_i} \\
&= \frac{1}{|\mathcal{I}|} \sum_{I\in\mathcal{I}} \frac{1}{U} \sum_{i\in I} \widehat{f}^{\text{ho}}_{T_i}.
\end{aligned}$$

It follows by convexity of $f \mapsto \ell(s,f)$ that

$$\mathbb{E}\left[\ell(s, \widehat{f}^{\text{ag}}_{\mathcal{T}^{mc}_{\tau,U}})\right] \leq \frac{1}{|\mathcal{I}|} \sum_{I\in\mathcal{I}} \mathbb{E}\left[\ell(s, \frac{1}{U}\sum_{i\in I} \widehat{f}^{\text{ho}}_{T_i})\right].$$

For any $I \in \mathcal{I}$, $(T_i)_{i\in I} \sim \mathcal{T}^{mc}_{\tau,U}$ and is independent of D_n, therefore $\frac{1}{U}\sum_{i\in I} \widehat{f}^{\text{ho}}_{T_i} \sim \widehat{f}^{\text{ag}}_{\mathcal{T}^{mc}_{\tau,U}}$. This yields the result. ■

It can be seen from the proof that the proposition also holds for Agcv. Therefore, increasing V always decreases the risk of both methods. On the other hand, no such theoretical result is known for CV, even though increasing the number of CV splits (for given τ) almost always improves performance in practice.

4 Simulation study

This section focuses on hyperparameter selection for the Lasso with huber loss, either using a fixed grid or using the reparametrization from Definition 2.3. The methods considered for this task are Aggregated hold-out given by Definition 2.5, Aggregated cross-validation given by Definition 2.6 and standard cross-validation. In all cases, the subsamples are generated independently from the data and uniformly among subsets of a given size τn, as in Definition 3.6. Thus, all three methods share the same two hyperparameters: τ, the fraction of data used for training the Lasso, and V, the number of subsets used by the method.

For the huberized Lasso with a fixed grid, the hqreg_raw function from the R package hqreg is used with a fixed grid designed to emulate the default choice: a geometrically decreasing sequence of length 100, with maximum value λ_{max} and minimum value $\lambda_{min} = 0.05\lambda_{max}$. The fixed value of λ_{max} is obtained by averaging the (data-dependent) default value chosen by hqreg_raw over 10 independent datasets. To compute the reparametrization given by Definition 2.3, I implemented the LARS-based algorithm described by Rosset and Zhu [23], which allows to compute the whole regularization path.

I.i.d training samples of size $n = 100$ are generated according to a distribution (X, Y), where $X \in \mathbb{R}^{1000}$ and $Y = w_*^T X + \varepsilon$, with ε independent from X. To illustrate the robustness of the estimators, Cauchy noise is used:

$\varepsilon \sim \text{Cauchy}(0, \sigma)$. The performance of Agghoo and cross-validation may depend on the presence of correlations between the covariates X and the sparsity of the ground truth w_*. To investigate these effects, three parametric families of distribution are considered for X, in sections 4.1, 4.2 and 4.3.

The risk of each method is evaluated on an independent training set of size 500, and results are averaged over 1000 repetitions of the simulation. More precisely, 1000 training sets D_j of size $n = 100$ are generated, along with 1000 test sets $(X'_{i,j}, Y'_{i,j})_{1 \le i \le 500}$, each of size 500. For each simulation j and any learning rule $\mathcal{A}_{\tau,V}$ among the six obtained by combining Agghoo, monte carlo CV and AGCV with either a fixed grid or the zero-norm parametrization, the average excess risk

$$\hat{R}_j(\mathcal{A}, \tau, V) = \frac{1}{500} \sum_{i=1}^{500} \left[\phi_c \left(Y'_{i,j} - \mathcal{A}_{\tau,V}(D_j)(X'_{i,j}) \right) - \phi_c \left(Y'_{i,j} - s(X'_{i,j}) \right) \right]$$

is computed on the test set for all values of $V \in \{1, 2, 5, 10\}$ and $\tau \in \left\{ \frac{i}{10} : 1 \le i \le 9 \right\}$.

4.1 Experimental setup 1

X is generated using the formula $X_i = \frac{1}{\|u\|_2} \sum_{j=1}^{d} u_{i-j} Z_j$, where Z_j are independent standard Gaussian random variables, $u_i = \mathbb{I}_{|i| \le cor} e^{-\frac{2.33^2 i^2}{2cor^2}}$ and $cor \in \mathbb{N}$ is a parameter regulating the strength of the correlations. The regression coefficient has a support of size $r = 3 * k$ drawn at random from $[|1; 1000|]$, and is defined by $w_{*,j} = u_{*,g(j)}$, where g is a uniform random permutation, $u_{*,j} = b$ if $1 \le j \le k$ and $u_{*,j} = \frac{b}{4}$ if $2k + 1 \le j \le 3k$, with b calibrated so that $\|Xw_*\|_{L^2} = 1$. The noise parameter is $\sigma = 0.08$, while the huber loss parameter c is set to 2 – a suboptimal choice in this setting, but convenient for computing the huberized Lasso regularization path.

Choice of τ parameter For all methods, in most cases the optimal value of τ is 0.8 or 0.9, similarly to what was observed in the rkhs case, where $\tau = 0.8$ was recommended. Table 1 displays the quantity

$$\hat{G}(\mathcal{A}, \tau, V) = \frac{\text{Mean}\left[(\hat{R}_j(\mathcal{A}, \tau, V) - \hat{R}_j(\mathcal{A}, \tau_*, V))_{1 \le j \le 1000} \right]}{\text{Sd}\left[(\hat{R}_j(\mathcal{A}, \tau, V) - \hat{R}_j(\mathcal{A}, \tau_*, V))_{1 \le j \le 1000} \right]},$$

where Sd denotes the (empirical) standard deviation and τ_* the optimal choice of τ, $\tau_* = \text{argmin}_{\tau \in \{0.1, \ldots, 0.9\}} \text{Mean}\left[(\hat{R}_j(\mathcal{A}, \tau, V))_{1 \le j \le 1000} \right]$. Thus, values of $\hat{G}(\mathcal{A}, \tau, V)$ bigger than a few units suggest that τ is suboptimal to a statistically significant degree. When $\tau_* = 0.9$, $\hat{G}(\mathcal{A}, 0.8, V)$ is displayed in black on table 1. When $\tau_* = 0.8$, $\hat{G}(\mathcal{A}, 0.9, V)$ is displayed in blue on table 1. Exceptions where $\tau_* \notin \{0.8, 0.9\}$ are highlighted in red, with the value $\min\left(\hat{G}(\mathcal{A}, 0.8, V), \hat{G}(\mathcal{A}, 0.9, V) \right)$.

Most of the exceptions $\tau_* \notin \{0.8, 0.9\}$ occur on the column $r = 150$, $cor = 1$, while most of the others are of low statistical significance, with values less than 1.1 on the fourth column ($r = 60$ and $cor = 1$). Thus, table 1 confirms the claim that $\tau_* \in \{0.8, 0.9\}$ for all methods, in most cases. For grid agghoo, $0-$norm agghoo, grid agcv and $V \ge 5$, $\tau_* \in \{0.8, 0.9\}$ for all simulations. Comparing

			r = 150		r = 60		r = 24	
	method	V	15	1	15	1	15	1
1	grid agghoo	1	2.2	2.7	3.0	2.7	0.5	5.6
2	grid agghoo	2	2.5	2.1	3.1	1.4	1.0	7.9
3	grid agghoo	5	2.5	6.8	3.5	0.6	0.6	11.9
4	grid agghoo	10	0.7	7.2	3.7	1.1	4.5	16.7
5	grid cv	1	1.0	3.9	1.6	0.1	1.2	1.5
6	grid cv	2	0.8	5.0	2.6	0.5	1.4	1.1
7	grid cv	5	1.4	2.8	1.5	0.8	0.5	3.7
8	grid cv	10	2.0	2.6	2.9	1.1	1.6	5.9
9	grid agcv	1	1.0	3.9	1.6	0.1	1.2	1.5
10	grid agcv	2	0.3	2.0	1.4	1.9	0.3	0.8
11	grid agcv	5	0.3	2.2	0.5	0.7	0.5	1.1
12	grid agcv	10	0.5	0.4	0.0	0.3	0.8	1.0
13	0−norm agghoo	1	1.3	4.1	2.0	0.3	0.5	5.6
14	0−norm agghoo	2	3.0	1.4	3.2	1.3	1.9	9.2
15	0−norm agghoo	5	4.0	6.7	5.1	3.3	4.0	13.7
16	0−norm agghoo	10	4.6	7.3	7.0	3.7	5.2	18.5
17	0−norm cv	1	4.3	9.4	4.3	1.1	2.0	3.9
18	0−norm cv	2	1.9	7.2	1.8	4.4	4.8	2.7
19	0−norm cv	5	2.7	5.3	2.4	3.3	1.5	0.7
20	0−norm cv	10	6.1	4.6	5.4	3.5	0.6	0.1
21	0−norm agcv	1	4.3	9.4	4.3	1.1	2.0	3.9
22	0−norm agcv	2	1.9	5.8	2.4	4.5	5.9	3.5
23	0−norm agcv	5	2.1	1.9	1.0	4.0	5.7	3.7
24	0−norm agcv	10	4.5	1.0	3.3	3.6	7.3	3.9

Table 1: $\hat{G}(\mathcal{A}, \tau, V)$ for sub-optimal $\tau \in \{0.8, 0.9\}$ and various distributions. Colours show optimal τ_*: blue for $\tau_* = 0.8$, black for 0.9, red when $\tau_* \notin \{0.8, 0.9\}$.

now $\tau = 0.8$ and $\tau = 0.9$, grid agghoo and 0−norm agghoo with $V \geq 5$ show a clear pattern: $\tau = 0.9$ is better or as good as $\tau = 0.8$ in all cases except $r = 150, cor = 1$ where $\tau = 0.8$ is significantly better. For other methods, results are not so clear and the difference in risk between the two values of τ is often insignificant.

Choice of V For all methods considered, performance is expected to improve when V is increased, but by how much? If the performance increase is too slight, it may not be worth the additional computational cost. In figure 1, the mean excess risk for the optimal value of τ is displayed as a function of V, with error bars corresponding to one standard deviation. The scale used for the vertical axis in each graph is the average excess risk of the oracle with respect to the fixed grid over the λ parameter. Quantifying performance as a percentage of the oracle risk, when $cor = 15$, Agghoo improves by roughly 20% from $V = 1$ to $V = 2$, by roughly 10% from $V = 2$ to $V = 5$ and by a few percent more from $V = 5$ to $V = 10$. CV with the standard grid behaves similarly in these two simulations, while CV with the zero-norm parametrization shows much less improvement when V is increased. Thus, taking $V \geq 5$ is advantageous,

but there are clearly diminishing returns to choosing V much larger than this. For CV with the zero-norm parametrization, $V = 2$ seems sufficient in these simulations .

Comparison between methods From figure 1, it appears that grid agcv is a very poor choice, being worse than both grid agghoo and grid cv for all values of V when $r = 150, cor = 15$, and being the worst of all the methods for $V \geq 2$ when $r = 24$, as well as highly unstable, as the size of the error bars clearly shows.

Interestingly, $0-$norm agcv behaves much better, being the second best method when $cor = 1$, and very close to the best when $r = 24$ and $cor = 15$.

Generally speaking, of the two types of parametrization of the Lasso, the zero-norm parametrization appears to perform better than the standard grid when correlations are small ($cor = 1$), while the performance is significantly worse when $r = 150$ and $cor = 15$.

Comparing now Agghoo and CV, Agghoo appears to be better than CV when $V \geq 2$ in situations where r is larger ($r = 150$). This seems to hold for both the standard parametrization (grid agghoo) and the zero-norm one ($0-$norm agghoo). The relation is reversed for small r, with CV performing better than Agghoo for all values of V when $r = 24$.

Further studies The previous simulations suggest that Agghoo performs better than CV in the case of high intrinsic dimension. However, the effect of correlations is unclear. Experimental setup 1 mixes different types of correlations: correlations between predictive variables, correlations between predictive and non-predictive variables, and correlations among non-predictive variables. It is possible that one type of correlation favours Agghoo while another favours CV.

To gain a more accurate idea of when Agghoo is advantageous over CV, two more settings are studied, considering separately correlations among predictive variables, and between predictive and non-predictive variables. Since previous simulations showed that $\tau = 0.8, 0.9$ and $V = 10$ were the optimal parameters, only those parameters will be considered in the following.

Since the choice of lasso parametrization did not seem to affect the relative performance of Agghoo and CV, we only consider the standard parametrization, as it is more popular and also easier to use in our simulations. Agcv is not considered either, since it was discovered to be unreliable in previous simulations.

4.2 Experimental setup 2: correlations between predictive and noise variables

Let r be the number of predictive variables and let each predictive covariate have s "noise" covariates which are correlated with it at level $\rho = 0.8$. Assume that $rs \leq d$, where d is the total number of variables. Let $(Z_i^0)_{1\leq i\leq r}$, $(Z_{i,j})_{1\leq i\leq r, 1\leq j\leq s}$ and $(W_k)_{1\leq k\leq d-rs}$ be independent standard gaussian variables. For any $j \in [|0 : r-1|]$ and any $i \in [|1; s|]$, let $Z_{jr+i} = \sqrt{0.8}Y_j^0 + \sqrt{0.2}Z_{i,j}$ and for $rs < i \leq d$, let $X_i = W_{i-rs}$. For the regression coefficient, choose $w_* = \frac{3*u}{\|Xu\|_{L_2}}$, where $u = (\mathbb{I}_{r|(j-1)}\mathbb{I}_{j\leq rs})_{1\leq j\leq d}$. Let then Y be distributed conditionnally on X as Cauchy$(\langle w_*, X\rangle, 0.3)$. The loss function used here is ϕ_c with $c = 2$.

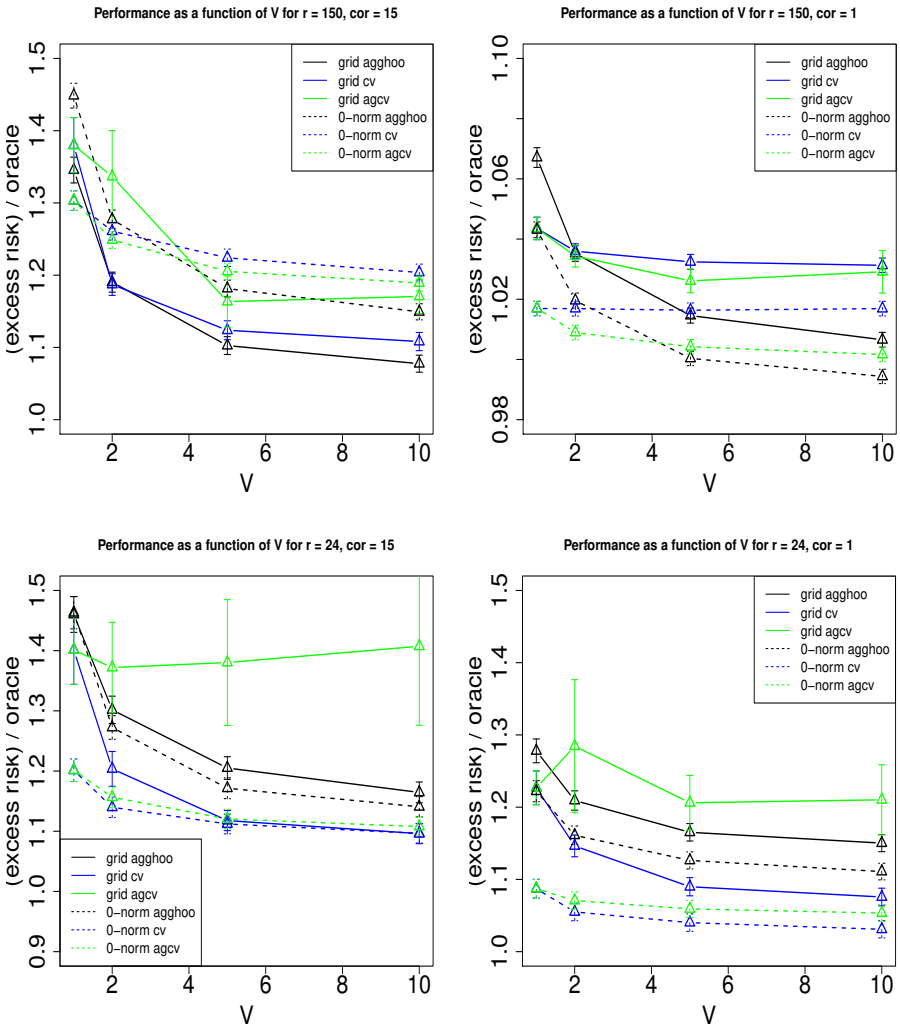

Figure 1: Performance relative to the oracle, as a function of V

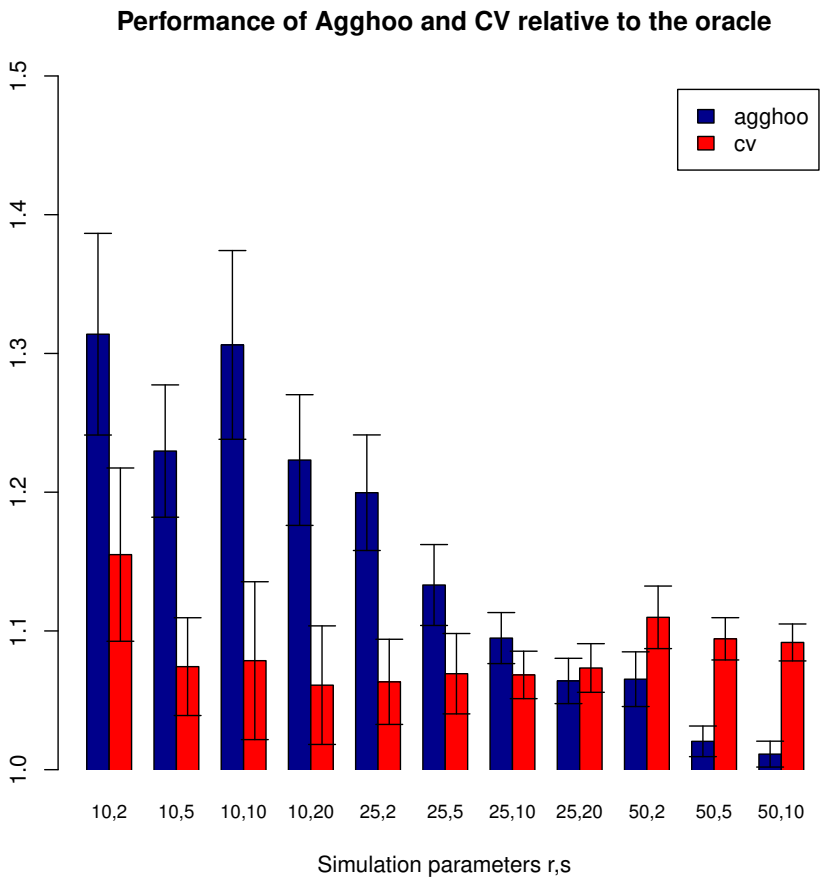

Figure 2: Relative risk in experimental setup 2 (section 4.2)

Results Figure 4.2 shows a bar plot of the average excess risk of CV and Agghoo as a fraction of the average risk of the oracle. 90 % error bars were estimated using asymptotic theory. Parameters used for Agghoo and CV were $\tau = 0.9$ and $V = 10$ ($\tau = 0.8$ yields similar result).

Overall, Agghoo's risk relative to the oracle significantly decreases as the zero-norm of w_* increases from $r = 10$ to $r = 50$, as was observed in section 4.1 . For $r = 25$ and $r = 50$ separately, the risk relative to the oracle significantly decreases as s increases from 2 to 10. For $r = 10$, this trend is unclear due to the random errors.

In contrast, CV's performance relative to the oracle shows no clear trend either as a function of r or as as function of s, and could be constant when taking error bars into accounts.

As a result of these trends, Agghoo performs significantly worse than CV for $r = 10$ and significantly better when $r = 50$, especially when $s \geq 5$. When $r = 25$, CV performs significantly better than Agghoo for $s = 2$ and $s = 5$ and they perform similarly when $s = 10$ and $s = 20$.

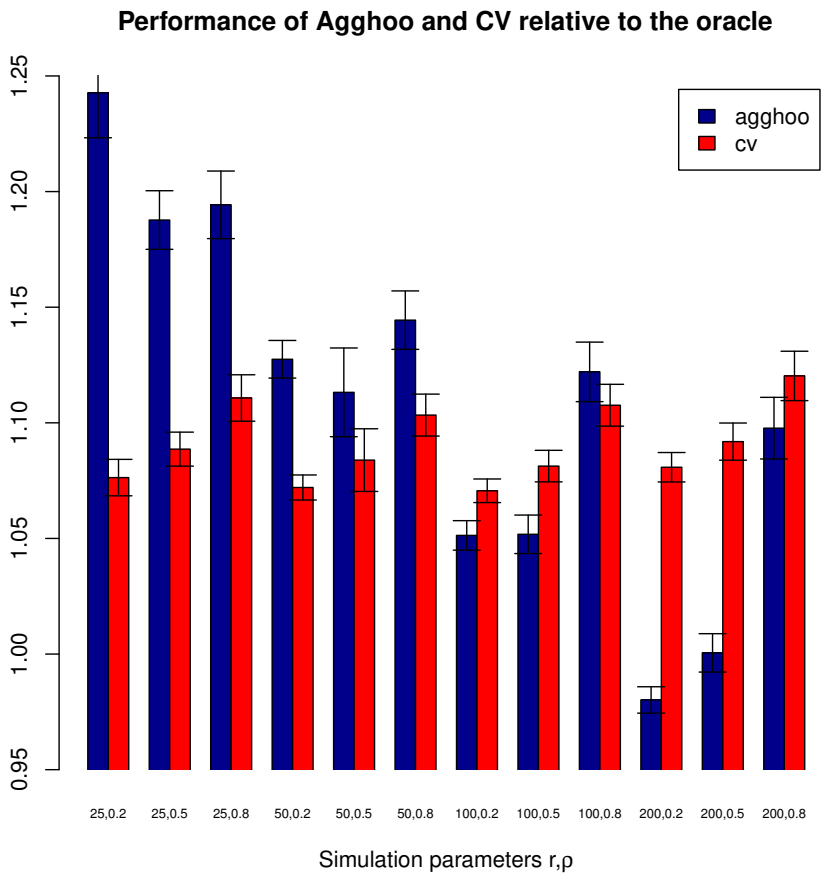

Figure 3: Relative risk in experimental setup 3 (section 4.3)

4.3 Experimental setup 3: correlations between predictive variables

We consider now predictive covariates which are correlated between them, and independent from the unpredictive covariates. As above, let r denote the number of predictive variables and $\rho > 0$ be the level of correlations. Let Z_0, $(Z_i)_{1\leq i\leq r}$ and $(W_i)_{1\leq i\leq d-r}$ be standard Gaussian random variables. The random variable X is then defined by $X_i = \sqrt{\rho}Z_0 + \sqrt{1-\rho}Z_i$ for $1 \leq i \leq r$ and $X_i = W_{i-r}$ for $r+1 \leq i \leq d$. As in section 4.2, the regression coefficient w_* is a constant vector of the form $\frac{3*u}{\|Xu\|_{L^2}}$, where this time $u = (\mathbb{I}_{1\leq i\leq r})_{1\leq i\leq d}$.

Y is distributed conditionnally on X as Cauchy$(\langle X, w_*\rangle, 0.3)$ and the loss function used is the Huber loss ϕ_2.

Results Figure 4.3 shows a barplot generated in the same way as in section 4.2. Parameters used for Agghoo and CV were $V = 10$ and $\tau = 0.8$, which is optimal in this case for both Agghoo and CV.

As in previous simulations, Agghoo's performance relative to the oracle improves significantly when the intrinsic dimension r grows from 25 to 200, for a given value of ρ. The decrease in relative risk is faster for small values of ρ. As a result, Agghoo performs best, relative to the oracle, when $\rho = 0.2$ for $r = 200$, whereas best performance seems to occur at $\rho = 0.5$ for smaller values of r, up to random errors.

For cross-validation, the relative risk seems more or less unaffected by the dimension r, but shows an increasing trend as a function of ρ for all values of r.

As a result, Agghoo performs better than CV for $r = 200$ and for $r = 100$ and $\rho = 0.2, 0.5$. For $r = 200$ and $\rho = 0.2$, Agghoo even performs significantly better than the oracle! This is possible, since the Agghoo regression coefficient $\hat{\theta}^{ag}_{\mathcal{T}}$ does not itself belong to the Lasso regularization path.

5 Conclusion

Aggregated hold-out (Agghoo) satisfies an oracle inequality (Theorem 3.2) in sparse linear regression with the huber loss. This oracle inequality is asymptotically optimal in the non-parametric case where the intrinsic dimension tends to $+\infty$ with the sample size n, provided that an $L^\infty(X) - L^2(X)$ norm inequality holds on the set of sparse linear predictors, where X is the random vector of covariates. When X is a vector of independent Bernoulli variables, this condition amounts to restricting the zero-norm of the coefficients to be less than a constant times $\frac{n}{\log n}$. Theorem 3.2 also applies to adaptive piecewise constant regression, yielding an oracle inequality in that setting (Proposition 3.5).

When Monte-Carlo subsampling is used (Definition 3.6), Agghoo has two parameters, τ and V. Theoretically, it is shown that Agghoo's performance always improves when V grows for a fixed τ. Simulations show a large improvement from $V = 1$ to $V = 5$ in some cases, but diminishing returns for $V > 5$. With respect to τ, simulations show that $\tau = 0.8$ or $\tau = 0.9$ is optimal or near optimal in most cases. In particular, a default choice of $V = 10$, $\tau = 0.8$ seems reasonable.

Compared to cross-validation with the same number of splits V, simulations show that Agghoo performs better when the intrinsic dimension r is large enough ($r = 150$ in section 4.1, $r = 50$ in section 4.2 and $r = 100$ in 4.3) for $n = 100$ observations and $d = 1000$ covariates. Correlations between predictive and non-predictive covariates, which increase the number of covariates correlated with the response Y, clearly favour Agghoo relative to CV and the oracle, whereas the effect of correlations between predictive covariates is ambiguous.

A Proof of Theorem 3.2

The idea is to apply [17, Theorem A.3] using suitable functions $(\hat{w}_{i,j})_{(i,j)\in\{1;2\}^2}$. Fix a dataset D_{n_t}, $K \in \{1, \dots, n_t\}$ and for any $k \in [|1; K|]^2$, let $\hat{t}_k = \mathcal{A}_k(D_{n_t}) : x \to \hat{q}_k(D_{n_t}) + \langle \hat{\theta}_k(D_{n_t}), x\rangle$. More precisely, to apply [17, Theorem A.3], one

must show inequalities of the form $H(w_1, w_2, (\hat{t}_k)_{1\leq k\leq K})$: for all $r \geq 2$,

$$\mathbb{E}\Big(\big|\phi_c(\hat{t}_k(X)-Y)-\phi_c(\hat{t}_l(X)-Y)-c_l^k\big|^r\Big) \leq k!\Big[w_1\big(\sqrt{\ell(s,\hat{t}_k)}\big)+w_1\big(\sqrt{\ell(s,\hat{t}_l)}\big)\Big]^2 \times \Big[w_2\big(\sqrt{\ell(s,\hat{t}_k)}\big)+w_2\big(\sqrt{\ell(s,\hat{t}_l)}\big)\Big]^{r-2}, \tag{12}$$

where w_1, w_2 are non-decreasing functions. Since ϕ_c is Lipschitz, it is enough to control $\left\|\hat{t}_k-\hat{t}_l\right\|_{L^\infty(X)}$ and $\left\|\hat{t}_k-\hat{t}_l\right\|_{L^2(X)}$ by functions of $\ell(s,\hat{t}_k)$ and $\ell(s,\hat{t}_l)$.

A.1 Controlling the supremum norm $\left\|\hat{t}_k-\hat{t}_l\right\|_{L^\infty(X)}$

First, let us bound the supremum norm by the L^2 norm.

Claim A.0.1 *For any $k \in \{1,\dots,K\}$, recall that $\hat{t}_k = \mathcal{A}_k(D_{n_t})$. Then:*

$$\forall (k,l) \in \{1,\dots,K\}^2, \left\|\hat{t}_k-\hat{t}_l\right\|_{L^\infty(X)} \leq \sqrt{2}\kappa(K)\left\|\hat{t}_k-\hat{t}_l\right\|_{L^2(X)} \ \ a.s. \ .$$

Proof Let X be independent from D_n and observe that for any k,

$$\hat{t}_k(X) = \hat{b}_k + \hat{\theta}_k^T(X-EX),$$

where $\hat{b}_k = \hat{q}_k + \hat{\theta}_k^T EX$ (using the notations of hypothesis 2.1). Hence,

$$\left\|\hat{t}_k(X)-\hat{t}_l(X)\right\|_{L^\infty} \leq |\hat{b}_k-\hat{b}_l| + \left\|(\hat{\theta}_k-\hat{\theta}_l)^T(X-EX)\right\|_{L^\infty}.$$

By hypothesis 2.1, $\left\|\hat{\theta}_k\right\|_0 = k$. Thus, if $K \geq \max(k,l)$, $\left\|\hat{\theta}_k-\hat{\theta}_l\right\|_0 \leq k+l \leq 2K$. The definition of κ (equation (5)) implies that

$$\begin{aligned}
\left\|\hat{t}_k(X)-\hat{t}_l(X)\right\|_{L^\infty} &\leq |\hat{b}_k-\hat{b}_l| + \kappa(K)\left\|(\hat{\theta}_k-\hat{\theta}_l)^T(X-EX)\right\|_{L^2} \\
&\leq \kappa(K)\left[|\hat{b}_k-\hat{b}_l| + \left\|(\hat{\theta}_k-\hat{\theta}_l)^T(X-EX)\right\|_{L^2}\right] \ (\kappa(K)\geq 1 \text{ by definition}) \\
&\leq \sqrt{2}\kappa(K)\sqrt{|\hat{b}_k-\hat{b}_l|^2 + \left\|(\hat{\theta}_k-\hat{\theta}_l)^T(X-EX)\right\|_{L^2}^2} \\
&= \sqrt{2}\kappa(K)\left\|\hat{t}_k(X)-\hat{t}_l(X)\right\|_{L^2}.
\end{aligned}$$

■

A uniform bound on the supremum norm is also required.

Definition A.1 *Let*

$$\hat{\beta} = \max_{(k,l)\in[|1;n_t|]^2}\left\|\hat{t}_k-\hat{t}_l\right\|_{L^\infty(X)}$$

$\mathbb{E}[\hat{\beta}]$ can be bounded as follows.

Claim A.1.1 *The $\hat{\beta}$ of definition A.1 is such that*

$$\mathbb{E}[\hat{\beta}] \leq 8LRn^\alpha$$

Proof Let $(k,l) \in [|1;n_t|]^2$. Defining $\widetilde{X}_i = X_i - \frac{1}{n}\sum_{i=1}^n X_i$ and changing variables in hypothesis 2.1 from (q,θ) to $\left(b = q+ < \theta, \frac{1}{n}\sum_{i=1}^n X_i >, \theta\right)$, we can rewrite $\hat{t}_k$ as

$$\hat{t}_k(x) = \hat{b}_k(D_n) + \hat{\theta}_k(D_n)^T \left(x - \frac{1}{n}\sum_{i=1}^n X_i \right)$$

where

$$\hat{b}_k(D_n) \in \underset{b \in \hat{Q}'\left(D_n, \hat{\theta}_k(D_n)\right)}{\operatorname{argmin}} |b|$$

$$\hat{Q}'(D_n,\theta) = \underset{b\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n \phi_c\left(Y_i - b - \theta^T \widetilde{X}_i\right).$$

Therefore, differentiating with respect to b,

$$\frac{1}{n}\sum_{i=1}^n \phi_c'\big(Y_i - \hat{b}_k - \hat{\theta}_k^T \widetilde{X}_i\big) = 0.$$

Assume by contradiction that

$$\exists b > 0, \forall i \in [|1;n_t|], \hat{b}_k + b + \hat{\theta}_k^T \widetilde{X}_i \leq \hat{b}_l + \hat{\theta}_l^T \widetilde{X}_i. \tag{13}$$

Let b be such that (13) holds. Then by monotony of ϕ_c', for all ε in $[0;\frac{b}{2}]$,

$$\begin{aligned}
0 &= \frac{1}{n}\sum_{i=1}^n \phi_c'\big(Y_i - \hat{b}_k - \hat{\theta}_k^T \widetilde{X}_i\big) \\
&\geq \frac{1}{n}\sum_{i=1}^n \phi_c'\big(Y_i - \hat{b}_k - \varepsilon - \hat{\theta}_k^T \widetilde{X}_i\big) \\
&\geq \frac{1}{n}\sum_{i=1}^n \phi_c'\big(Y_i - \hat{b}_k - \frac{b}{2} - \hat{\theta}_k^T \widetilde{X}_i\big) \\
&\geq \frac{1}{n}\sum_{i=1}^n \phi_c'\big(Y_i - \hat{b}_l + \frac{b}{2} - \hat{\theta}_l^T \widetilde{X}_i\big) \\
&\geq \frac{1}{n}\sum_{i=1}^n \phi_c'\big(Y_i - \hat{b}_l + \varepsilon - \hat{\theta}_l^T \widetilde{X}_i\big) \\
&\geq \frac{1}{n}\sum_{i=1}^n \phi_c'\big(Y_i - \hat{b}_l - \hat{\theta}_l^T \widetilde{X}_i\big) \\
&= 0.
\end{aligned}$$

It follows that

$$\forall \varepsilon \in [0;\frac{b}{2}], \frac{1}{n}\sum_{i=1}^n \phi_c'\big(Y_i - \hat{b}_k - \varepsilon - \hat{\theta}_k^T \widetilde{X}_i\big) = \frac{1}{n}\sum_{i=1}^n \phi_c'\big(Y_i - \hat{b}_l + \varepsilon - \hat{\theta}_l^T \widetilde{X}_i\big) = 0. \tag{14}$$

By integration, this implies that for all $\varepsilon \in [0;\frac{b}{2}]$,

$$(\hat{b}_k + \varepsilon) \in \hat{Q}'\left(D_n, \hat{\theta}_k(D_n)\right), \tag{15}$$

$$(\hat{b}_l - \varepsilon) \in \hat{Q}'\left(D_n, \hat{\theta}_l(D_n)\right). \tag{16}$$

If $\hat{b}_l > 0$, then for small enough ε, (16) contradicts the minimality of $|\hat{b}_l|$. On the other hand, if $\hat{b}_l \leq 0$, then averaging (13) over $i \in \{1, \ldots, n\}$ yields

$$\hat{b}_k + b \leq \hat{b}_l \leq 0.$$

Then for $\varepsilon \in [0; \frac{b}{2}]$, (15) contradicts the minimality of $|\hat{b}_k|$. Thus, (13) leads to a contradiction. Let i be such that $\hat{b}_k + \hat{\theta}_k^T \widetilde{X}_i \geq \hat{b}_l + \hat{\theta}_l^T \widetilde{X}_i$. Then

$$\hat{b}_l - \hat{b}_k \leq \left(\hat{\theta}_k - \hat{\theta}_l\right)^T \widetilde{X}_i \leq \max_{i=1,\ldots,n_t} \left|\left(\hat{\theta}_k - \hat{\theta}_l\right)^T \widetilde{X}_i\right|.$$

Exchanging k and l yields

$$|\hat{b}_l - \hat{b}_k| \leq \max_{1 \leq i \leq n_t} \left|\left(\hat{\theta}_k - \hat{\theta}_l\right)^T \widetilde{X}_i\right|.$$

Therefore, for any k, l,

$$\begin{aligned}
\left\|\hat{t}_k - \hat{t}_l\right\|_{L^\infty}(X) &\leq |\hat{b}_l - \hat{b}_k| + \sup_{x \in \mathrm{supp}(X)} \left|(\hat{\theta}_k - \hat{\theta}_l)^T (x - \frac{1}{n}\sum_{i=1}^n X_i)\right| \\
&\leq 2 \sup_{x \in \mathrm{supp}(X)} \left|(\hat{\theta}_k - \hat{\theta}_l)^T (x - \frac{1}{n}\sum_{i=1}^n X_i)\right| \\
&\leq 2\left\|\hat{\theta}_k - \hat{\theta}_l\right\|_1 \sup_{(x,y) \in \mathrm{supp}(X)^2} \|x - y\|_\infty \\
&\leq 4\left\|\hat{\theta}_k - \hat{\theta}_l\right\|_1 \sup_{x \in \mathrm{supp}(X)} \|x - EX\|_\infty \\
&\leq 8 \sup_{1 \leq k \leq n_t} \left\|\hat{\theta}_k\right\| \sup_{x \in \mathrm{supp}(X)} \|x - EX\|_\infty .
\end{aligned}$$

Thus, by definition A.1, $\hat{\beta} \leq 8 \sup_{1 \leq k \leq n_t} \left\|\hat{\theta}_k\right\| \sup_{x \in \mathrm{supp}(\bar{X})} \|x\|_\infty$.

Hence, by hypothesis 2.1,

$$\mathbb{E}[\hat{\beta}] \leq 8LRn^\alpha.$$

■

A.2 Proving hypotheses $H\left(\hat{w}_{i,1}, \hat{w}_{i,2}, (\hat{t}_k)_{1 \leq k \leq K}\right)$

The following lemma will be useful.

Lemma A.2 *Let r, s, x be positive real numbers. Let*

$$I_{r,s}(x) = \{v \in \mathbb{R}_+ : v \leq (r \vee s\sqrt{v})x\}$$

and $h_{r,s}(x) = (\sqrt{r}x) \vee sx^2$. Then for all $x, y \geq 0$,

$$\sup I_{r,s}(x + y) \leq (h_{r,s}(\sqrt{x}) + h_{r,s}(\sqrt{y}))^2.$$

Proof Let $v \in I_{r,s}(x)$. Then if $s\sqrt{v} \leq r$, by definition of $I_{r,s}(x)$, $v \leq rx$. Otherwise $s\sqrt{v} > r$. Therefore by definition of $I_{r,s}(x)$,

$$\begin{aligned} v &\leq s\sqrt{v}x \\ v &\leq s^2x^2. \end{aligned}$$

In all cases, $v \leq (rx) \vee s^2x^2$. Therefore,

$$\sqrt{\sup I_{r,s}(x+y)} \leq \left(\sqrt{r(x+y)}\right) \vee s(x+y) \leq \left(\sqrt{rx}\right) \vee sx + \left(\sqrt{ry}\right) \vee sy,$$

using the elementary inequalities:

$$\begin{aligned} &\forall (x,y,a,b) \in \mathbb{R}_+^3, \\ &\sqrt{x+y} \leq \sqrt{x} + \sqrt{y} \\ &(x+y) \vee (a+b) \leq x \vee a + y \vee b \end{aligned}$$

■

We now relate the L^2 norm to the excess risk in the following Proposition.

Proposition A.3 *Let $(X,Y) \in \mathcal{X} \times \mathbb{R}$ be random variables. Let ϕ_c be the Huber loss with parameter $c > 0$. Assume that there exists $\eta > 0$ such that almost everywhere,*

$$\mathbb{P}\left(|Y - s(X)| \leq \frac{c}{2}|X\right) \geq \eta.$$

Then for any measurable functions $(f_1, f_2) : \mathcal{X} \to \mathbb{R}^2$,

$$c^2 \left\|f_1 - f_2\right\|_{L^2(X)}^2 \leq \frac{4c}{\eta}\left[c \vee 2\left\|f_1 - f_2\right\|_{L^\infty(X)}\right]\left[\ell(s,f_1) + \ell(s,f_2)\right]. \tag{17}$$

Proof Recall that

$$\phi_c(x) = \frac{x^2}{2}\mathbb{I}_{|x|\leq c} + c\big(|x| - \frac{c}{2}\big)\mathbb{I}_{|x|>c}.$$

In the rest of the proof, for any $x \in \mathcal{X}$, let $\ell_x(u) = \mathbb{E}\left[\phi_c(Y-u) - \phi_c(Y)|X = x\right]$. Let $s : x \mapsto \operatorname{argmin}_{u\in\mathbb{R}} \ell_x(u)$; s is a risk minimizer. Then $\phi_c'(x) = \operatorname{sgn}(x)(|x| \wedge c)$ and $\phi_c''(x) = \mathbb{I}_{|x|\leq c}$. By differentiating under the expectation, for any u such that $|u - s(x)| \leq \frac{c}{2}$,

$$\begin{aligned} \ell_x''(u) &= \partial_u^2 \mathbb{E}\left[\phi_c(Y-u) - \phi_c(Y)|X = x\right] \\ &= \mathbb{E}\left[\phi_c''(Y-u)|X = x\right] \\ &= \mathbb{P}\left[|Y-u| \leq c|X = x\right] \\ &\geq \mathbb{P}\left[|Y - s(x)| + |u - s(x)| \leq c|X = x\right] \\ &\geq \mathbb{P}\left[|Y - s(x)| \leq \frac{c}{2}|X = x\right] \\ &\geq \eta \end{aligned}$$

Since $s(x)$ is a local minimum, it follows that, for any $u \in \left[s(x) - \frac{c}{2}; s(x) + \frac{c}{2}\right]$,

$$\begin{aligned}|\ell'_x(u)| &= \left|\ell'_x(u) - \ell'_x(s(x))\right| \\ &= \left|\int_{s(x)}^{u} \ell''_x(t)dt\right| \\ &\geq \eta|u - s(x)|\end{aligned}$$

By [17, Lemma C.2], it follows that for any $(u, v) \in \mathbb{R}^2$,

$$(u - v)^2 \leq \left(\frac{4}{\eta} \vee \left(\frac{8}{c\eta}|u - v|\right)\right)(\ell_x(u) + \ell_x(v) - 2\ell_x(s(x)))\,. \tag{18}$$

Now using equation (18) with $u = f_1(X)$, $v = f_2(X)$, $x = X$ and taking expectations, we have:

$$c^2 \|f_1 - f_2\|^2_{L^2(X)} \leq \frac{4c}{\eta}\left[c \vee 2\|f_1 - f_2\|_{L^\infty(X)}\right][\ell(s, f_1) + \ell(s, f_2)]\,. \tag{19}$$

■

We are now ready to obtain functions $(\hat{w}_{i,j})_{(i,j)\in\{1;2\}^2}$ such that $H\left(\hat{w}_{i,1}, \hat{w}_{i,2}, (\hat{t}_k)_{1\leq k\leq K}\right)$ holds. In the following, fix $K \in [|1; n_t|]$ and write $\kappa = \kappa(K)$ for short.

By Proposition A.3, for all $(k, l) \in [|1; K|]^2$,

$$c^2 \left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^2(X)} \leq \frac{4c}{\eta}\left[c \vee (2\left\|\hat{t}_k - \hat{t}_l\right\|_{L^\infty(X)})\right]\left[\ell(s, \hat{t}_k) + \ell(s, \hat{t}_l)\right]. \tag{20}$$

Hence, by claim A.0.1, for all $(k, l) \in [|1; K|]^2$,

$$c^2 \left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^2(X)} \leq \frac{4c}{\eta}\left[c \vee (2\sqrt{2}\kappa\left\|\hat{t}_k - \hat{t}_l\right\|_{L^2(X)})\right]\left[\ell(s, \hat{t}_k) + \ell(s, \hat{t}_l)\right] \tag{21}$$

By lemma A.2 with $v = c^2 \left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^2(X)}$, $r = \frac{4c^2}{\eta}$ and $s = \frac{8\sqrt{2}\kappa}{\eta}$,

$$c^2 \left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^2(X)} \leq \left(\hat{w}_A(\sqrt{\ell(s, \hat{t}_k)}) + \hat{w}_A(\sqrt{\ell(s, \hat{t}_l)})\right)^2, \tag{22}$$

where

$$\hat{w}_A(x) = \left(\frac{2c}{\sqrt{\eta}}x\right) \vee \frac{8\sqrt{2}\kappa}{\eta}x^2. \tag{23}$$

Now,

$$\begin{aligned}c^2 \left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^\infty(X)} &\leq 2c^2\kappa^2 \left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^2(X)} \\ &\leq \frac{8c\kappa^2}{\eta}\left[c \vee 2\left\|\hat{t}_k - \hat{t}_l\right\|_{L^\infty(X)}\right]\left[\ell(s, \hat{t}_k) + \ell(s, \hat{t}_l)\right] \text{ by Proposition } A.3.\end{aligned}$$

By lemma A.2 with $v = c^2 \left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^\infty(X)}$, $s = \frac{16\kappa^2}{\eta}$, $r = \frac{8c^2\kappa^2}{\eta}$,

$$c^2 \left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^\infty(X)} \leq \left(\hat{w}_{2,2}(\sqrt{\ell(s, \hat{t}_k)}) + \hat{w}_{2,2}(\sqrt{\ell(s, \hat{t}_k)})\right)^2 \tag{24}$$

where

$$\hat{w}_{2,2}(x) = \frac{2c\sqrt{2}\kappa}{\sqrt{\eta}} x \vee \frac{16\kappa^2}{\eta} x^2 = \sqrt{2}\kappa \hat{w}_A. \tag{25}$$

Because the Huber loss ϕ_c is $c-$Lipschitz,

$$\forall u, v \in \mathbb{R}^d, |\phi_c(Y-u) - \phi_c(Y-v)| \leq c|u-v|$$

Therefore by (22) and (24),

$$\begin{aligned}
\mathbb{E}\left[\left(\phi_c(Y-\hat{t}_k(X)) - \phi_c(Y-\hat{t}_l(X))\right)^k\right] &\leq \left(c^2 \left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^2(X)}\right)\left(c\left\|\hat{t}_k - \hat{t}_l\right\|_{L^\infty(X)}\right)^{k-2} \\
&\leq \left(\hat{w}_A(\sqrt{\ell(s,\hat{t}_k)}) + \hat{w}_A(\sqrt{\ell(s,\hat{t}_l)})\right)^2 \\
&\quad \times \left(\sqrt{2}\kappa\hat{w}_A(\sqrt{\ell(s,\hat{t}_k)}) + \sqrt{2}\kappa\hat{w}_A(\sqrt{\ell(s,\hat{t}_l)})\right)^{k-2},
\end{aligned}$$

which proves $H\left(\hat{w}_A, \sqrt{2}\kappa\hat{w}_A, (\hat{t}_k)_{1\leq k\leq K}\right)$. Now going back to equation (17), by Definition A.1,

$$\begin{aligned}
c^2 \left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^2(X)} &\leq \frac{4c}{\eta}\left[c \vee 2\hat{\beta}\right]\left[\ell(s,\hat{t}_k) + \ell(s,\hat{t}_l)\right] \\
&\leq \left(\hat{w}_B(\sqrt{\ell(s,\hat{t}_k)}) + \hat{w}_B(\sqrt{\ell(s,\hat{t}_l)})\right)^2
\end{aligned} \tag{26}$$

where

$$\hat{w}_B(x) = \left(\frac{2c}{\sqrt{\eta}} \vee 2\sqrt{\frac{2c\hat{\beta}}{\eta}}\right) x. \tag{27}$$

Moreover,

$$\begin{aligned}
c^2 \left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^\infty(X)} &\leq 2\kappa^2 c^2 \left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^2(X)} \\
&\leq 2\kappa^2 \left(\hat{w}_B(\sqrt{\ell(s,\hat{t}_k)}) + \hat{w}_B(\sqrt{\ell(s,\hat{t}_l)})\right)^2 \text{ by (26)}
\end{aligned}$$

Therefore, by (26),

$$\begin{aligned}
P|\gamma(\hat{t}_k) - \gamma(\hat{t}_l)|^k &\leq \mathbb{E}\left[\left(\phi_c(Y-\hat{t}_k(X)) - \phi_c(Y-\hat{t}_l(X))\right)^2\right]\left\|\phi_c(Y-\hat{t}_k(X)) - \phi_c(Y-\hat{t}_l(X))\right\|^{k-2}_\infty \\
&\leq \left(c^2 \left\|\hat{t}_k - \hat{t}_l\right\|^2_{L^2(X)}\right)\left(c\left\|\hat{t}_k - \hat{t}_l\right\|_{L^\infty(X)}\right)^{k-2} \\
&\leq \left(\hat{w}_B(\sqrt{\ell(s,\hat{t}_k)}) + \hat{w}_B(\sqrt{\ell(s,\hat{t}_l)})\right)^2 \\
&\quad \times \left(\sqrt{2}\kappa\hat{w}_B(\sqrt{\ell(s,\hat{t}_k)}) + \sqrt{2}\kappa\hat{w}_B(\sqrt{\ell(s,\hat{t}_l)})\right)^{k-2},
\end{aligned}$$

which proves $H\left(\hat{w}_B, \sqrt{2}\kappa\hat{w}_B, (\hat{t}_k)_{1\leq k\leq K}\right)$.

A.3 Conclusion of the proof

We have proved that $H\left(\hat{w}_B, \sqrt{2}\kappa\hat{w}_B, (\hat{t}_k)_{1\le k\le K}\right)$ and $H\left(\hat{w}_A, \sqrt{2}\kappa\hat{w}_A, (\hat{t}_k)_{1\le k\le K}\right)$ hold, where $\hat{w}_B$ is given by equation (27) and $\hat{w}_A$ is defined by equation (23). It remains to apply [17, Theorem A.3] and to express the remainder term as a simple function of $c, n_v, n_t, \kappa, L, R, K$ and α. We recall here the definition of the operator δ used in the statement of that theorem.

Definition A.4 *For any function $h : \mathbb{R}_+ \mapsto \mathbb{R}_+$ and any $\xi > 0$, let*

$$\delta(h, \xi) = \inf\{x \in \mathbb{R}_+ : \forall u \ge x, h(u) \le \xi u^2\}.$$

The following lemma will facilitate the computation of $\delta(\hat{w}_A, \cdot)$.

Lemma A.5 *Let $r > 0, s > 0$ and $h_{r,s}(x) = (\sqrt{r}x) \vee sx^2$. Then $\delta(h_{r,s}, \xi) < \infty$ if and only if $\xi \ge s$ and then $\delta(h_{r,s}, \xi) = \frac{\sqrt{r}}{\xi}$.*

Proof To find $\delta(h_{r,s}, \xi)$, notice that given the definition of $\delta(h_{r,s}, \xi)$, the condition $s \le \xi$ is obviously necessary for the infimum to be finite. Assume now that $\xi \ge s$. For any $u \ge \frac{\sqrt{r}}{\xi}$, then $\xi u^2 \ge \sqrt{r}u$ as well as $\xi u^2 \ge su^2$ (since we assumed $\xi \ge s$), therefore $\xi u^2 \ge h_{r,s}(u)$. Thus by definition, $\delta(h_{r,s}, \xi) \le \frac{\sqrt{r}}{\xi}$ (in particular, $\delta(h_{r,s}, \xi)$ is finite). Furthermore, by definition of $\delta(h_{r,s}, \xi)$, $\sqrt{r}\delta(h_{r,s}, \xi) \le \xi\delta(h_{r,s}, \xi)^2$, that is $\delta(h_{r,s}, \xi) \ge \frac{\sqrt{r}}{\xi}$. ■

The following claim can now be proved.

Claim A.5.1 *If $K \in [|1; n_t|]$ and $b > 1$ are such that*

$$\kappa(K) \le \frac{\eta}{8}\sqrt{\frac{n_v}{8b\log K}}, \tag{28}$$

then applying Agghoo to the collection $(\mathcal{A}_k)_{1\le k\le K}$ yields the following oracle inequality.

$$(1-\theta)\mathbb{E}[\ell(s, \widehat{f}^{\mathrm{ag}}_{\mathcal{T}})] \le (1+\theta)\mathbb{E}[\min_{1\le k\le K} \ell(s, \hat{t}_k)] + 24\theta b\frac{c\log K}{\eta n_v}\left[c + \frac{c + 2LRn_t^{\alpha}}{K^{\theta^2 b - 1}}\right].$$

Proof Theorem [17, Theorem A.3] applies with $\hat{w}_{1,1} = \hat{w}_B, \hat{w}_{1,2} = \sqrt{2}\kappa\hat{w}_B, \hat{w}_{2,1} = \hat{w}_A, \hat{w}_{2,2} = \sqrt{2}\kappa\hat{w}_A$, $x = (\theta^2 b - 1)\log K$ and it remains to bound the remainder terms $(R_{2,i})_{1\le i\le 4}$. Now assume that equation (28) holds.

Bound on $R_{2,1}(\theta) = \sqrt{2}\theta\mathbb{E}\left[\delta^2\left(\hat{w}_A, \frac{\theta}{2}\sqrt{\frac{n_v}{\theta^2 b\log K}}\right)\right]$

By (23), we can apply lemma A.5 with $s = \frac{8\sqrt{2}\kappa}{\eta}$, $r = \frac{4c^2}{\eta}$ and $\xi = \frac{1}{2}\sqrt{\frac{n_v}{b\log K}}$. By (28),

$$s = \frac{8}{\eta}\sqrt{2}\kappa \le \sqrt{\frac{n_v}{4b\log K}} = \xi.$$

It follows by lemma A.5 that

$$\delta\left(\hat{w}_A, \sqrt{\frac{n_v}{4b\log K}}\right) = \frac{2c}{\sqrt{\eta}}\sqrt{\frac{4b\log K}{n_v}}.$$

Hence,

$$R_{2,1}(\theta) \le \sqrt{2}\theta\frac{4c^2}{\eta}\frac{4b\log K}{n_v} \le 23\theta b\frac{c^2\log K}{\eta n_v} \tag{29}$$

Bound on $R_{2,2}(\theta) = \frac{\theta^2}{2}\mathbb{E}\left[\delta^2\left(\sqrt{2}\kappa\hat{w}_A, \frac{\theta^2}{4}\frac{n_v}{\theta^2 b\log K}\right)\right]$

By (23), we can apply lemma A.5 with $s = \frac{16\kappa^2}{\eta}$, $r = \frac{8c^2\kappa^2}{\eta}$ and $\xi = \frac{n_v}{4b\log K}$. By (28) and since $\eta \leq 1$,

$$s = \frac{16\kappa^2}{\eta} \leq \frac{16}{\eta}\frac{\eta^2}{64}\frac{n_v}{4b\log K} \leq \frac{n_v}{4b\log K} = \xi.$$

Therefore

$$\begin{aligned}\delta\left(\sqrt{2}\kappa\hat{w}_A, \frac{\theta^2}{4}\frac{n_v}{\theta^2 b\log K}\right) &\leq \frac{2c\sqrt{2}\kappa}{\sqrt{\eta}}\frac{4b\log K}{n_v} \text{ by lemma A.5} \\ &\leq \frac{c}{2}\sqrt{\frac{\eta b\log K}{n_v}} \text{ by (28).}\end{aligned}$$

Hence, since $\theta, \eta \in [0;1]$,

$$R_{2,2}(\theta) \leq \frac{\theta^2}{2}\frac{c^2}{4}\frac{\eta b\log K}{n_v} \leq \frac{\theta b}{8}\frac{c^2\log K}{n_v}. \tag{30}$$

Bound on $R_{2,3}(\theta) = \frac{1}{K^{\theta^2 b-1}}\left(\theta + \frac{2\left[1+\log(K)\right]}{\theta}\right)\mathbb{E}\left[\delta^2\big(\hat{w}_B, \sqrt{n_v}\big)\right]$

By (27), $x \to \frac{\hat{w}_B(x)}{x}$ is constant and, in particular, non-increasing. Therefore, $\delta\big(\hat{w}_B, \sqrt{n_v}\big)$ is the unique nonnegative solution to the equation

$$\hat{w}_B(x) = \sqrt{n_v}x^2 \iff \left(\frac{2c}{\sqrt{\eta}} \vee 2\sqrt{\frac{2c\hat{\beta}}{\eta}}\right)x = \sqrt{n_v}x^2.$$

It follows that

$$\delta\big(\hat{w}_B, \sqrt{n_v}\big) = \left(\frac{2c}{\sqrt{\eta}} \vee 2\sqrt{\frac{2c\hat{\beta}}{\eta}}\right)\frac{1}{\sqrt{n_v}}. \tag{31}$$

and

$$\begin{aligned}\delta^2\big(\hat{w}_B, \sqrt{n_v}\big) &= \left(\frac{4c^2}{\eta} \vee \frac{8c\hat{\beta}}{\eta}\right)\frac{1}{n_v} \\ &= 4\frac{c(c \vee 2\hat{\beta})}{\eta n_v}.\end{aligned} \tag{32}$$

As $n_t \geq 3$, we can assume that $K \geq 3$, hence $\log K \geq 1$ and

$$\theta + \frac{2(1+\log K)}{\theta} \leq \frac{5\log K}{\theta}.$$

Since $\theta \geq \frac{1}{\sqrt{b}}$, $\frac{1}{\theta} \leq \theta b$ and therefore

$$\theta + \frac{2(1+\log K)}{\theta} \leq 5\theta b\log K. \tag{33}$$

By equations (32) and (33),

$$R_{2,3}(\theta) \leq 20\theta b\frac{c\log K}{\eta n_v}\frac{c + 2\mathbb{E}[\hat{\beta}]}{K^{\theta^2 b-1}} \tag{34}$$

Bound on $R_{2,4}(\theta) = \frac{1}{K^{\theta^2 b-1}}\left(\theta + \frac{2(1+\log K)+\log^2 K}{\theta}\right)\mathbb{E}\left[\delta^2\left(\sqrt{2}\kappa\hat{w}_B, n_v\right)\right]$

$\delta^2\left(\sqrt{2}\kappa\hat{w}_B, \sqrt{n_v}\right)$ is the unique nonnegative solution to the equation

$$\sqrt{2}\kappa\hat{w}_B(x) = \sqrt{n_v}x^2 \iff \left(\frac{2c}{\sqrt{\eta}} \vee 2\sqrt{\frac{2c\hat{\beta}}{\eta}}\right)\sqrt{2}\kappa x = n_v x^2,$$

which yields

$$\begin{aligned}\delta\left(\sqrt{2}\kappa\hat{w}_B, n_v\right) &= \left(\frac{2c}{\sqrt{\eta}} \vee 2\sqrt{\frac{2c\hat{\beta}}{\eta}}\right)\frac{\sqrt{2}\kappa}{n_v} \\ &\leq \frac{1}{8}\sqrt{\frac{\eta}{bn_v\log K}}\left(c \vee \sqrt{2c\hat{\beta}}\right) \text{ by (28) .}\end{aligned}$$

Since $\eta \leq 1$ and $b \geq 1$, it follows that

$$\delta^2\left(\sqrt{2}\kappa\hat{w}_B, n_v\right) \leq \frac{\eta}{64}\frac{c^2 \vee 2c\hat{\beta}}{bn_v\log K} \tag{35}$$

$$\tag{36}$$

By equation (33), and since $\frac{1}{\theta} \leq \theta b$,

$$\begin{aligned}\theta + \frac{2(1+\log K)+\log^2 K}{\theta} &\leq 5\theta b\log K + \theta b\log^2 K \\ &\leq 6\theta b\log^2 K \text{ since } K \geq 3.\end{aligned}$$

Therefore, since $b \geq 1$ and $\eta \in [0;1]$,

$$R_{2,4}(\theta) \leq \frac{6\theta\eta}{64}\frac{c\log K}{n_v}\frac{c+2\mathbb{E}[\hat{\beta}]}{K^{\theta^2 b-1}} \leq \frac{3\theta b}{32}\frac{c\log K}{\eta n_v}\frac{c+2\mathbb{E}[\hat{\beta}]}{K^{\theta^2 b-1}}. \tag{37}$$

Conclusion Summing up equations (29), (30), (34) and (37), [17, Theorem A.3] implies that assuming equation (28) holds for K, for all $\theta \in \left[\frac{1}{\sqrt{b}};1\right]$,

$$(1-\theta)\mathbb{E}[\ell(s,\widehat{f}^{\text{ag}}_{\mathcal{T}})] \leq (1+\theta)\mathbb{E}[\min_{1\leq k\leq K}\ell(s,\hat{t}_k)] + 24\theta b\frac{c\log K}{\eta n_v}\left[c + \frac{c+2\mathbb{E}[\hat{\beta}]}{K^{\theta^2 b-1}}\right]. \tag{38}$$

It follows by claim A.1.1 that

$$(1-\theta)\mathbb{E}[\ell(s,\widehat{f}^{\text{ag}}_{\mathcal{T}})] \leq (1+\theta)\mathbb{E}[\min_{1\leq k\leq K}\ell(s,\hat{t}_k)] + 24\theta b\frac{c\log K}{\eta n_v}\left[c + \frac{c+16LRn_t^{\alpha}}{K^{\theta^2 b-1}}\right].$$

■

Let K satisfy assumption (6) from Theorem 3.2. Take now $b = b_0\frac{\log n_t}{\log K}$. Then equation (28) holds for this value of b. Moreover, since $K \leq n_t$, $b_0 \geq b$

and therefore $\theta \in \left[\frac{1}{\sqrt{b_0}}; 1\right] \implies \theta \in \left[\frac{1}{\sqrt{b}}; 1\right]$. Thus, claim A.5.1 implies that for any $\theta \in \left[\frac{1}{\sqrt{b_0}}; 1\right]$,

$$(1-\theta)\mathbb{E}[\ell(s, \widehat{f}^{\text{ag}}_{\mathcal{T}})] \leq (1+\theta)\mathbb{E}[\min_{1\leq k\leq K} \ell(s, \hat{t}_k)] + 24\theta b_0 \frac{c \log n_t}{\eta n_v} \left[c + \frac{cK}{n_t^{\theta^2 b_0}} + \frac{16KLR}{n_t^{\theta^2 b_0 - \alpha}}\right].$$

This proves Theorem 3.2.

B Applications of Theorem 3.2

B.1 Proof of corollary 3.3

Let $\bar{x} \in \text{supp}(X - EX)$. Let $\theta \in \mathbb{R}^p$ be such that $\|\theta\|_0 \leq 2K$ and let $I = \{i : \theta_i \neq 0\}$. For all $i \in [|1; p|]$, $\bar{x}_i \in [-1; 1]$, hence

$$\begin{aligned}
\left|\sum_{i=1}^p \theta_i \bar{x}_i\right| &= \left|\sum_{i\in I} \theta_i \bar{x}_i\right| \\
&\leq \sqrt{\sum_{i\in I} \bar{x}_i^2} \sqrt{\sum_{i\in I} \theta_i^2} \\
&\leq \sqrt{2K} \sqrt{\sum_{i\in I} \theta_i^2}. \qquad (39)
\end{aligned}$$

On the other hand, let $\bar{X}_i = X_i - EX_i$. Then by independence of the X_i,

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{i=1}^p \theta_i \bar{X}_i\right)^2\right] &= \sum_{i=1}^p \theta_i^2 \text{Var}(X_i) \\
&= \sum_{i=1}^p \theta_i^2 p_i(1-p_i) \\
&\geq \min_{1\leq i\leq p} \{p_i(1-p_i)\} \times \sum_{i\in I} \theta_i^2. \qquad (40)
\end{aligned}$$

Combining inequalities (39) and (40) yields

$$\sup_{\bar{x}\in \text{supp}(X-EX)} \left|\sum_{i=1}^p \theta_i \bar{x}_i\right| \leq \sqrt{\frac{2K}{\min_{1\leq i\leq p} p_i(1-p_i)}} \sqrt{\mathbb{E}\left[\langle \theta, \bar{X}\rangle^2\right]}.$$

B.2 Proof of Proposition 3.5

We associate to the original piecewise constant regression problem an ordinary sparse regression problem. Define the linear operator

$$\begin{aligned}
S : \mathbb{R}^d &\to \mathbb{R}^d \\
(x_j)_{1\leq j\leq d} &\mapsto \left(\sum_{i=1}^j x_i\right)_{1\leq j\leq d}
\end{aligned}$$

and:

$$\begin{aligned}\Delta : \mathbb{R}^d &\to \mathbb{R}^{d-1}\\ (x_j)_{1\le j\le d} &\mapsto ((x_j - x_{j-1})_{j\ge 2})\,.\end{aligned}$$

Fo any $k \in [|1;d|]$, let also $E_k = \cup_{j=k}^d I_j$. Then t_u has k jumps if and only if $\|\Delta(u)\|_0 = k$, moreover we have the representation:

$$t_u = u_1\mathbb{I} + \sum_{j=2}^d \Delta(u)_j \mathbb{I}_{E_j}. \tag{41}$$

Equivalently,

$$t_{S(u)} = \sum_{j=1}^d u_j \mathbb{I}_{E_j} \tag{42}$$

It follows that the original jump detection problem is equivalent to a sparse regression problem with covariate vector $X_i = \left(\mathbb{I}_{E_j}(U_i)\right)_{2\le j\le d}$, where the non-penalized intercept corresponds to the component u_1 of the original problem. The set of $X-$measurable functions coïncides with the set of linear functions on the finite set $\mathrm{supp}(X)$, hence a Bayes estimator is $s : x \mapsto u_{*,1} + \langle \Delta(u_*), x\rangle)$, where $u_* = \mathrm{argmin}_{u\in\mathbb{R}^d}\, \mathbb{E}\left[\phi_c(Y - t_u(X)\right]$. Let $\hat\theta_k = \Delta(\hat u_k)$ be the sparse estimator associated to $\hat u_k$. This allows to apply Theorem 3.2, provided that $(\hat\theta_k)_{1\le k\le d}$ satisfies also the two last points of hypothesis 2.1. Let us now prove this. Write $\hat\theta_k, \hat u_k$ for $\theta_k(\hat D_{n_t}), \hat u_k(D_{n_t})$ to simplify notation. Using the notation of Definition 3.4, we have:

$$\left\|\hat\theta_k\right\|_1 = \|\Delta(\hat u_k)\|_1 = \sum_{r=2}^k |\hat u_{k,j_r(\hat u_k)} - \hat u_{k,j_{r-1}(\hat u_k)}| \le 2\sum_{r=1}^k |\hat u_{k,j_r(\hat u_k)}|.$$

Hence, by the triangular inequality,

$$\left\|\hat\theta_k\right\|_1 \le 2\sum_{r=0}^k \min_{i:X_i\in A_r(\hat u_k)} |Y_i| + |Y_i - \hat u_{j_r(\hat u_k)}|.$$

Using the inequality $|x| \le c + \frac{1}{c}\phi_c(x)$ yields

$$\begin{aligned}\left\|\hat\theta_k\right\|_1 &\le 2\sum_{r=0}^k \min_{i:X_i\in A_r(\hat u_k)} |Y_i| + c + \frac{1}{c}\phi_c\left(Y_i - \hat u_{j_r(\hat u_k)}\right)\\ &\le 2\sum_{r=0}^k c + \sum_{i:X_i\in A_r(\hat u_k)} |Y_i| + \frac{1}{c}\phi_c\left(Y_i - \hat u_{j_r(\hat u_k)}\right).\end{aligned}$$

Hence, by Definition 3.4,

$$\begin{aligned}\left\|\hat\theta_k\right\|_1 &\le 2\sum_{r=0}^k c + \sum_{i:X_i\in A_r(\hat u_k)} |Y_i| + \frac{1}{c}\phi_c\left(Y_i\right)\\ &\le 2kc + 4\sum_{i=1}^{n_t} |Y_i|.\end{aligned}$$

It follows that

$$\mathbb{E}\left[\sup_{1\le k\le d-1}\left\|\hat{\theta}_k\right\|_1\right] \le 2dc + 4n_t\mathbb{E}[|Y|], \tag{43}$$

so since $d \le n_v \le n_t$, the second point of hypothesis 2.1 is satisfied with $\alpha = 1$ and $L = 2c + 4\mathbb{E}[|Y|]$. Let $u \in \mathbb{R}^d$. For any $y \in \mathbb{R}$, by convexity of ϕ_c,

$$\begin{aligned} y \in \operatorname*{argmin}_{q\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n \phi_c(Y_i - t_u(U_i) - q) &\iff \frac{1}{n}\sum_{i=1}^n \phi_c'\left(Y_i - y - t_u(U_i)\right) = 0 \\ &\iff \frac{1}{n}\sum_{i=1}^n \phi_c'\left(Y_i - y - u_1 - \sum_{j=2}^d \Delta(u)_j \mathbb{I}_{E_j}(U_i)\right) = 0 \\ &\iff u_1 + y \in \operatorname*{argmin}_x \frac{1}{n}\sum_{i=1}^n \phi_c\left(Y_i - x - \langle \Delta(u), X\rangle\right). \end{aligned}$$

Hence, using the notations of hypothesis 2.1, for any $u \in \mathbb{R}^d$,

$$\hat{Q}\big((X_i, Y_i)_{1\le i\le n}, \Delta(u)\big) = u_1 + \operatorname*{argmin}_{q\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n \phi_c(Y_i - t_u(U_i) - q). \tag{44}$$

By Definition 3.4, $\hat{u}_k \in \hat{C}_t(D_n, \hat{u}_k)$, therefore

$$\begin{aligned} \sum_{i=1}^n \phi_c'(Y_i - t_{\hat{u}_k}(U_i)) &= \sum_{r=1}^k \sum_{i:X_i\in A_r(\hat{u}_k)} \phi_c'\left(Y_i - \hat{u}_{k,j_r(\hat{u}_k)}\right) \\ &= 0, \end{aligned}$$

therefore $0 \in \operatorname{argmin}_{q\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n \phi_c(Y_i - t_{\hat{u}_k}(U_i) - q)$ by convexity of ϕ_c. Hence, by equation (44), $\hat{u}_{k,1} \in \hat{Q}\big((X_i, Y_i)_{1\le i\le n}, \hat{\theta}_k\big)$. Let now $u_1' \in \hat{Q}\big((X_i, Y_i)_{1\le i\le n}, \hat{\theta}_k\big)$, and let $y = u_1' - \hat{u}_{k,1}$. Assume by contradiction that $(\hat{u}_{k,j} + y)_{1\le j\le d} \notin \hat{C}_t(D_n, \hat{u}_k)$. Then there exists $l \in [|1; k|]$ such that

$$\sum_{i:X_i\in A_l(\hat{u}_k)} \phi_c\left(Y_i - \hat{u}_{k,j_l(\hat{u}_k)} - y\right) > \sum_{i:X_i\in A_l(\hat{u}_k)} \phi_c\left(Y_i - \hat{u}_{k,j_l(\hat{u}_k)}\right),$$

while for all $r \ne l$, since by assumption $\hat{u}_k \in \hat{C}_t(D_n, \hat{u}_k)$,

$$\sum_{i:X_i\in A_r(\hat{u}_k)} \phi_c\left(Y_i - \hat{u}_{k,j_r(\hat{u}_k)} - y\right) \ge \sum_{i:X_i\in A_r(\hat{u}_k)} \phi_c\left(Y_i - \hat{u}_{k,j_r(\hat{u}_k)}\right).$$

It follows that:

$$\begin{aligned} \sum_{i=1}^n \phi_c(Y_i - t_{\hat{u}_k}(U_i) - y) &= \sum_{r=1}^k \sum_{i:X_i\in A_r(\hat{u}_k)} \phi_c\left(Y_i - \hat{u}_{k,j_r(\hat{u}_k)} - y\right) \\ &> \sum_{r=1}^k \sum_{i:X_i\in A_r(\hat{u}_k)} \phi_c\left(Y_i - \hat{u}_{k,j_r(\hat{u}_k)}\right) \\ &= \sum_{i=1}^d \phi_c(Y_i - t_{\hat{u}_k}(U_i)). \end{aligned}$$

Yet by equation (44), $y \in \operatorname{argmin}_{q\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n \phi_c(Y_i - t_{\hat{u}_k}(U_i) - q)$, which yields a contradiction. Therefore, $(\hat{u}_{k,j} + y)_{1\le j\le d} \in \hat{C}_t(D_n, \hat{u}_k)$.

Thus,

$$\begin{aligned}
\left| u_1' + \langle \hat{\theta}_k, \frac{1}{n}\sum_{i=1}^n X_i \rangle \right| &= \left| \hat{u}_{k,1} + y + \frac{1}{n}\sum_{i=1}^n \sum_{j=2}^d \Delta(\hat{u}_k)_j \mathbb{I}_{E_j}(U_i) \right| \\
&= \left| y + \frac{1}{n}\sum_{i=1}^n t_{\hat{u}_k}(U_i) \right| \\
&= \left| \sum_{j=1}^d P_n(I_j)\,[\hat{u}_{k,j} + y] \right| \\
&\ge \left| \sum_{j=1}^d P_n(I_j)\hat{u}_{k,j} \right| \text{ by Definition 3.4, equation (8)} \\
&= \left| \frac{1}{n}\sum_{i=1}^n t_{\hat{u}_k}(U_i) \right| \\
&= \left| \hat{u}_{k,1} + \langle \hat{\theta}_k, \frac{1}{n}\sum_{i=1}^n X_i \rangle \right|.
\end{aligned}$$

This shows that the linear regressor $D_n \to x \to \hat{u}_{k,1} + \langle \Delta(\hat{u}_k), x\rangle$ satisfies also the last point of hypothesis 2.1.

It remains to check the assumptions of Theorem 3.2. We first remark that $R \le \sup_{x\in\mathbb{R}} \max_j |\mathbb{I}_{E_j}(x) - P(E_j)| \le 1$. We now bound $\kappa(K)$. For any $\theta \in \mathbb{R}^d$,

$$\begin{aligned}
\sum_{j=k}^d \theta_k \bar{X}_k &= \sum_{k=1}^d \theta_k \left(\mathbb{I}_{E_k}(U) - P(E_k)\right) \\
&= \sum_{k=1}^d \theta_k \sum_{j=1}^d \mathbb{I}_{j\ge k} \left(\mathbb{I}_{I_j}(U) - P(I_j)\right) \\
&= \sum_{j=1}^d \left(\sum_{k=1}^j \theta_k\right) \left(\mathbb{I}_{I_j}(U) - P(I_j)\right)
\end{aligned}$$

Thus, for any $\theta \in \mathbb{R}^d$, $\langle \bar{X}, \theta\rangle$ is in $\operatorname{span}\left(\mathbb{I}_{I_j}(U)\right)_{1\le j\le d}$ (constant functions also belong to this space). Now, for any $Z = \sum_{j=1}^d z_j \mathbb{I}_{I_j}(U) \in \operatorname{span}\left(\mathbb{I}_{I_j}(U)\right)_{1\le j\le d}$, by assumption (10) of Proposition 3.5,

$$\|Z\|_\infty \le \max_{1\le j\le d} |z_j| \le \sqrt{\sum_{j=1}^d z_j^2} \le \frac{1}{\sqrt{\min_{1\le j\le d} P(I_j)}} \sqrt{\sum_{j=1}^d z_j^2 P(I_j)} \le \sqrt{\frac{\eta^2 n_v}{1536 \log^2 n_t}} \mathbb{E}[Z^2]^{\frac{1}{2}}.$$

This yields $\kappa(d) \le \sqrt{\frac{\eta^2 n_v}{1536\log^2 n_t}}$. Let $b_0 = 3\log n_t$, then $\kappa(d) \le \frac{\eta}{8\log n_t}\sqrt{\frac{n_v}{24}} = \frac{\eta}{8}\sqrt{\frac{n_v}{8 b_0 \log n_t}}$. Finally, applying Theorem 3.2 with $K = d$, $R = 1$, $L = 2c + 4\mathbb{E}[|Y|]$, $\alpha = 1$ (by equation (43)), $b_0 = 3\log n_t$ and $\theta = \frac{1}{\sqrt{\log n_t}}$ yields equation (11), proving Proposition 3.5.

References

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

[2] 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.

[3] Francis Bach. Bolasso: Model consistent lasso estimation through the bootstrap. *Proceedings of the 25th international conference on Machine learning*, 33-40, 05 2008.

[4] Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 09 1998.

[5] Kevin Bleakley and Jean-Philippe Vert. The group fused lasso for multiple change-point detection. *Arxiv preprint arXiv:1106.4199*, 06 2011.

[6] Sourav Chatterjee and Jafar Jafarov. Prediction error of cross-validated lasso. 02 2015.

[7] X. Chen, Z. J. Wang, and M. J. McKeown. Asymptotic analysis of robust lassos in the presence of noise with large variance. *IEEE Transactions on Information Theory*, 56(10):5131–5149, Oct 2010.

[8] Denis Chetverikov, Zhipeng Liao, and Victor Chernozhukov. On cross-validated Lasso. *arXiv e-prints*, page arXiv:1605.02214, May 2016.

[9] Pascaline Descloux and Sylvain Sardy. Model selection with lasso-zero: adding straw to the haystack to better find needles. *arXiv e-prints*, page arXiv:1805.05133, May 2018.

[10] Eitan Greenshtein and Ya'Acov Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 12 2004.

[11] Darren Homrighausen and Daniel McDonald. Risk consistency of cross-validation with lasso-type procedures. *Statistica Sinica*, 27, 08 2013.

[12] 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.

[13] Peter J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964.

[14] Sophie Lambert-Lacroix and Laurent Zwald. Robust regression through the huber's criterion and adaptive lasso penalty. *Electron. J. Statist.*, 5:1015–1053, 2011.

[15] Marc Lavielle. Using penalized contrasts for the change-point problem. *Signal Processing*, 85(8):1501 – 1510, 2005.

[16] Guillaume Lecué and Charles Mitchell. Oracle inequalities for cross-validation type procedures. *Electron. J. Statist.*, 6:1803–1837, 2012.

[17] Guillaume Maillard, Sylvain Arlot, and Matthieu Lerasle. Aggregated Hold-Out. working paper or preprint, September 2019.

[18] 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.

[19] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society Series B*, 72:417–473, 09 2010.

[20] Léo Miolane and Andrea Montanari. The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *arXiv e-prints*, page arXiv:1811.01212, Nov 2018.

[21] Fabien Navarro and Adrien Saumard. Slope heuristics and V-Fold model selection in heteroscedastic regression using strongly localized bases. *ESAIM: Probability and Statistics*, 21, 2017.

[22] Fabien Navarro and Adrien Saumard. Slope heuristics and v-fold model selection in heteroscedastic regression using strongly localized bases. *ESAIM: Probability and Statistics*, 21:412–451, 2017.

[23] Saharon Rosset and Ji Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, 35(3):1012–1030, 2007.

[24] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259 – 268, 1992.

[25] Adrien Saumard. Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression. *Electron. J. Statist.*, 6:579–655, 2012.

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

[27] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, pages 91–108, 2005.

[28] Ryan J. Tibshirani and Jonathan Taylor. Degrees of freedom in lasso problems. *Ann. Statist.*, 40(2):1198–1232, 04 2012.

[29] Sara van de Geer and Johannes Lederer. The lasso, correlated design, and improved oracle inequalities. 07 2011.

[30] V. N. Vapnik. An overview of statistical learning theory. *Transactions on Neural Networks*, 10(5):988–999, sep 1999.

[31] 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.

[32] Hansheng Wang and Chenlei Leng. Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, 102(479):1039–1048, 2007.

[33] Hansheng Wang, Guodong Li, and Guohua Jiang. Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business and Economic Statistics*, 25(3):347–355, 2007.

[34] Sijian Wang, Bin Nan, Saharon Rosset, and Ji Zhu. Random lasso. *Ann. Appl. Stat.*, 5(1):468–485, 03 2011.

[35] Zhan Wang, Sandra Paterlini, Fuchang Gao, and Yuhong Yang. Adaptive minimax regression estimation over sparse ℓ_q-hulls. *Journal of Machine Learning Research*, 15:1675–1711, 2014.

[36] Marten Wegkamp. Model selection in nonparametric regression. *Ann. Statist.*, 31(1):252–273, 02 2003.

[37] Huan Xu, Constantine Caramanis, and Shie Mannor. Sparse algorithms are not stable: A no-free-lunch theorem. *IEEE transactions on pattern analysis and machine intelligence*, 34, 08 2011.

[38] Congrui Yi and Jian Huang. Semismooth newton coordinate descent algorithm for elastic-net penalized huber loss regression and quantile regression. *Journal of Computational and Graphical Statistics*, 26(3):547–557, 2017.

[39] Hui Zou, Trevor Hastie, and Robert Tibshirani. On the "degrees of freedom" of the lasso. *Annals of Statistics*, 35(5):2173–2192, 2007.