

Rejoinder on: Minimal penalties and the slope heuristics: a survey

Sylvain Arlot¹

I am grateful to the Editor-in-Chief for arranging such an informative and stimulating discussion of my survey paper (Arlot, 2019), and to all discussants for their inspiring comments. The discussion nicely complements the paper on various aspects. Let me add a few comments on these, following approximately the order of the paper.

For the sake of clarity, references to the survey paper are made with bolded numbers.

R.1. Slope heuristics when there is some model bias (Sections 2 and 8.2)

Gilles Celeux and Christine Keribin ask whether it is possible to justify the slope heuristics when *all* models are biased, assuming that the bias is constant over the large models. The answer is positive. For instance, modifying slightly the proof of Theorem 1 in Section 2, we obtain the following result.

Theorem R.1. *In the framework described in Section 2.1 of the survey, assume that \mathcal{M} is finite and that*

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n). \quad (\text{HG})$$

Let $m_1 \in \mathcal{M}$ be such that

$$B := \min_{m \in \mathcal{M}} \left\{ \frac{1}{n} \|(I_n - \Pi_m)F\|^2 \right\} = \frac{1}{n} \|(I_n - \Pi_{m_1})F\|^2 \quad (\text{R.1})$$

and assume that $m_2 \in \mathcal{M}$ exists such that $D_{m_2} \leq D_{m_1}/20$. Recall that for every $C \geq 0$, $\hat{m}(C)$ is defined by Eq. (10) of the survey. Then, for every $\gamma \geq 0$, some $n_0(\gamma)$ exists such that if $n \geq n_0(\gamma)$, with probability at least $1 - 4\text{card}(\mathcal{M})n^{-\gamma}$, the following inequalities hold simultaneously:

$$\forall C \leq (1 - \eta_n^-) \sigma^2, \quad D_{\hat{m}(C)} \geq \frac{9D_{m_1}}{10}, \quad (\text{R.2})$$

$$\text{and } \forall C \geq (1 + \eta_n^+) \sigma^2, \quad D_{\hat{m}(C)} \leq \frac{D_{m_1}}{10}, \quad (\text{R.3})$$

¹ Université Paris-Saclay, Univ. Paris-Sud, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France. E-mail: sylvain.arlot@u-psud.fr

where

$$\begin{aligned}\sigma^2 \eta_n^- &= \frac{n}{D_{m_1}} (57\sigma\sqrt{B} + 41\sigma^2) \sqrt{\frac{\gamma \log(n)}{n}}, \\ \sigma^2 \eta_n^+ &= \frac{n}{D_{m_1}} \left[20(B' - B) + (114\sigma\sqrt{B'} + 82\sigma^2) \sqrt{\frac{\gamma \log(n)}{n}} \right], \\ \text{and } B' &:= \mathcal{B}\left(\frac{D_{m_1}}{20}\right) = \inf_{m \in \mathcal{M} / D_m \leq D_{m_1}/20} \left\{ \frac{1}{n} \|(I_n - \Pi_m)F\|^2 \right\}.\end{aligned}$$

On the same event, Eq. (17) and (18) of the survey also hold true.

Proof. We adapt the proof of Theorem 1 of the survey. We keep unchanged the definition of $n_0(\gamma)$, step 1 and step 2.1 of the proof, as well as step 3 which still applies for proving Eq. (17)–(18). Only steps 2.2 and 2.3 need to be modified, as detailed below. Since we use it repeatedly in the proof, let us recall Eq. (25) of the survey, which holds true on the event Ω_x considered in the proof of Theorem 1:

$$\begin{aligned}\forall m \in \mathcal{M}, \quad |G_C(m) - \text{crit}_C(m)| &\leq 2\sigma^2 \left(\sqrt{\frac{x}{n}} + \frac{x}{n} \right) + \frac{2\sigma\sqrt{2x}}{n} \|(I_n - \Pi_m)F\| \quad (25) \\ \text{where } \text{crit}_C(m) &:= \frac{1}{n} \left[\|(I_n - \Pi_m)F\|^2 + (C - \sigma^2)D_m \right]. \quad (\text{Eq. (11)})\end{aligned}$$

Step 2.2': lower bound on $D_{\hat{m}(C)}$ when C is too small (proof of Eq. (R.2)) Let $C \in [0, \sigma^2)$. Since $\hat{m}(C)$ minimizes $G_C(m)$ over $m \in \mathcal{M}$, it is sufficient to prove that if $C \leq (1 - \eta_n^-)\sigma^2$,

$$G_C(m_1) < \inf_{m \in \mathcal{M}, D_m < 9D_{m_1}/10} \{G_C(m)\}. \quad (R.4)$$

On the one hand, by Eq. (25) and (R.1),

$$\begin{aligned}G_C(m_1) &\leq \text{crit}_C(m_1) + 2\sigma^2 \left(\sqrt{\frac{x}{n}} + \frac{x}{n} \right) + 2\sigma\sqrt{\frac{2x}{n}}\sqrt{B} \\ &= B + (C - \sigma^2)\frac{D_{m_1}}{n} + 2\sigma^2 \left(\sqrt{\frac{x}{n}} + \frac{x}{n} \right) + 2\sigma\sqrt{\frac{2x}{n}}\sqrt{B}.\end{aligned} \quad (R.5)$$

On the other hand, by Eq. (25), for any $m \in \mathcal{M}$ such that $D_m < 9D_{m_1}/10$,

$$\begin{aligned}G_C(m) &\geq \frac{(C - \sigma^2)D_m}{n} - 2\sigma^2 \left(\sqrt{\frac{x}{n}} + \frac{x}{n} \right) + \frac{1}{n} \|(I_n - \Pi_m)F\|^2 - \frac{2\sigma\sqrt{2x}}{n} \|(I_n - \Pi_m)F\| \\ &> \frac{9D_{m_1}}{10n} (C - \sigma^2) - 2\sigma^2 \left(\sqrt{\frac{x}{n}} + \frac{x}{n} \right) + B - 2\sigma\sqrt{\frac{2x}{n}}\sqrt{B} - \frac{2\sigma^2 x}{n},\end{aligned} \quad (R.6)$$

using that $u - 2\alpha\sqrt{u} \geq B - 2\alpha\sqrt{B} - \alpha^2$ when $u \geq B \geq 0$, with

$$\alpha = \sigma\sqrt{\frac{2x}{n}} \quad \text{and} \quad u = \frac{1}{n} \|(I_n - \Pi_m)F\|^2 \geq B \quad \text{by Eq. (R.1).}$$

To conclude, the upper bound in Eq. (R.5) is smaller than the lower bound in Eq. (R.6) when

$$C \leq \sigma^2 - \frac{20n}{D_{m_1}} \left(2\sigma^2 \sqrt{\frac{x}{n}} + 3\sigma^2 \frac{x}{n} + 2\sigma \sqrt{\frac{2x}{n}} \sqrt{B} \right) =: \tilde{C}'_1(x). \quad (\text{R.7})$$

Taking $x = \gamma \log(n)$, for $n \geq n_0(\gamma)$, we have $\tilde{C}'_1(x) \geq \sigma^2(1 - \eta_n^-)$ hence Eq. (R.2).

Step 2.3': upper bound on $D_{\hat{m}(C)}$ when C is large enough (proof of Eq. (R.3)) Let $C > \sigma^2$. Similarly to the proof of Eq. (R.2), it is sufficient to prove that if $C \geq (1 + \eta_n^+) \sigma^2$,

$$G_C(m_2) < \inf_{m \in \mathcal{M}, D_m > D_{m_1}/10} \{G_C(m)\} \quad (\text{R.8})$$

where $m_2 \in \operatorname{argmin}_{m \in \mathcal{M} / D_m \leq D_{m_1}/20} \{\|(I_n - \Pi_m)F\|^2\}$ exists by assumption. On the one hand, by Eq. (25),

$$\begin{aligned} G_C(m_2) &\leq \operatorname{crit}_C(m_2) + 2\sigma^2 \left(\sqrt{\frac{x}{n}} + \frac{x}{n} \right) + \frac{2\sigma\sqrt{2x}}{n} \|(I_n - \Pi_{m_2})F\| \\ &\leq B' + \frac{(C - \sigma^2)D_{m_1}}{20n} + 2\sigma^2 \left(\sqrt{\frac{x}{n}} + \frac{x}{n} \right) + 2\sigma \sqrt{\frac{2x}{n}} \sqrt{B'}. \end{aligned} \quad (\text{R.9})$$

On the other hand, as in step 2.2', using Eq. (25) and Eq. (R.1), we get that for any $m \in \mathcal{M}$ such that $D_m > D_{m_1}/10$,

$$G_C(m) > \frac{D_{m_1}}{10n} (C - \sigma^2) - 2\sigma^2 \left(\sqrt{\frac{x}{n}} + \frac{x}{n} \right) + B - 2\sigma \sqrt{\frac{2x}{n}} \sqrt{B} - \frac{2\sigma^2 x}{n}. \quad (\text{R.10})$$

To conclude, the upper bound in Eq. (R.9) is smaller than the lower bound in Eq. (R.10) when

$$C \geq \sigma^2 + \frac{20n}{D_{m_1}} \left[B' - B + 2\sigma \sqrt{\frac{2x}{n}} (\sqrt{B} + \sqrt{B'}) + 2\sigma^2 \left(2\sqrt{\frac{x}{n}} + \frac{3x}{n} \right) \right] =: \tilde{C}'_2(x). \quad (\text{R.11})$$

Taking $x = \gamma \log(n)$, for $n \geq n_0(\gamma)$, we have $\tilde{C}'_2(x) \leq \sigma^2(1 + \eta_n^+)$ since $B \leq B'$, hence Eq. (R.3). \square

Theorem R.1 can also be generalized to sub-Gaussian noise, following Remark 1 in Section 2.5. Note that an equivalent of Eq. (15) —with an additional term depending on B — could certainly be obtained under the assumptions of Theorem R.1.

Assume that $D_{m_1} > \frac{10}{17} \max_{m \in \mathcal{M}} D_m$; for instance, this assumption holds true when the models are nested since one can always take $m_1 \in \operatorname{argmax}_{m \in \mathcal{M}} D_m$. Then,

$$\frac{9D_{m_1}}{10} - \frac{D_{m_1}}{10} > \max_{m \in \mathcal{M}} \{D_m\} - \frac{9D_{m_1}}{10},$$

so Theorem R.1 shows that $\hat{C}_{\text{window}}(\eta)$ with $\eta = \max\{\eta_n^-, \eta_n^+\}$ is close to σ^2 ; a formal statement can be obtained, similarly to Proposition 3 in Section 6.1.

Theorem R.1 shows one possible generalization of Theorem 1, that allows all models to be biased ($B > 0$). The particular case $\Pi_{m_1} = I_n$ —hence $B = 0$ and $D_{m_1} = n$ —yields Theorem 1, up to numerical constants¹. When $B > 0$ and $D_{m_1} < n$, η_n^- and η_n^+ are slightly larger than in Theorem 1—hence the estimation of σ^2 by the slope heuristics could be less precise—, but this loss is at most by a multiplicative factor of order

$$\frac{n}{D_{m_1}} \max \left\{ 1, \sqrt{\frac{B'}{\sigma^2}} \right\}.$$

For instance, if σ^2 stays constant while $\mathcal{M} = \mathcal{M}_n$, $m_1 = m_{1,n}$, $B = B_n$ and $B' = B'_n$ are such that

$$D_{m_{1,n}} \gg \sqrt{n \log(n)}, \quad \frac{B'_n}{\sigma^2} \in \mathcal{O}(1) \quad \text{and} \quad B'_n - B_n \ll \frac{\sigma^2 D_{m_{1,n}}}{n}$$

as n tends to infinity, then

$$\lim_{n \rightarrow +\infty} \max \{ \eta_n^-, \eta_n^+ \} = 0,$$

hence the slope heuristics yields a consistent estimation of σ^2 .

Therefore, contrary to what is suggested at the end of Section 5.2.2 and in Section 8.2, having a model with a small approximation error in $(S_m)_{m \in \mathcal{M}}$ is not an “unavoidable assumption” for the slope heuristics to work. Could it be relaxed even more² than what is done by Theorem R.1 above?

R.2. Related procedures (Section 6)

R.2.1. Elbow heuristics (Section 6.4)

Some *elbow-heuristics* procedure is present in the simulation experiments of Servane Gey—under the name $\hat{m}_{plateau}$ —, for the calibration of CART in binary classification. I notice that it performs decently, but a “jump” formulation of the slope heuristics seems to outperform it, at least in the four settings considered.

R.2.2. Scree test and related methods (Section 6.5)

David Donoho and Matan Gavish provide an interesting insight into the *scree test* for choosing the number of factors in factor analysis, in the light of random-matrix theory. Their assessment of several eigenvalue-thresholding procedures points out several important ideas and questions.

Universal singular value thresholding (USVT; Chatterjee, 2015) is another procedure related to minimal-penalty algorithms, not mentioned in the survey. In the setting considered by David Donoho and Matan Gavish, USVT suggests to take a threshold λ “just above” the minimal threshold $\lambda_{\min} = 2\sqrt{n}\sigma$, which is suboptimal since Gavish and Donoho (2014) show that the optimal threshold is $(4/\sqrt{3})\sqrt{n}\sigma = (2/\sqrt{3})\lambda_{\min}$. As in the slope heuristics, there is here a factor

¹ We can even recover the same constants as in Theorem 1 by using Eq. (R.11) in the proof of Theorem R.1.

² Following Gilles Celeux’s footnote, having jumped onto the Moon, I now hope that some researchers will be able to explore Mars.

$2/\sqrt{3} \approx 1.15 > 1$ between the minimal and optimal values of the tuning parameter (the constant C in front of the penalty for the slope heuristics, the threshold λ here). Therefore, I consider these results as perfectly consistent with what is known about minimal-penalty algorithms.

A remarkable difference with the slope heuristics is that this factor is not equal to 2, and it is not universal: in the white-noise model, it depends on the aspect ratio m/n of the matrix, and in general it depends on the noise distribution. From the practical point of view, as detailed in Section R.4.7, a universal relationship between the minimal threshold and a good (but suboptimal) threshold—such as USVT—would be useful when the noise distribution is unknown, unless a fully data-driven optimal procedure is available.

Note finally that Sylvain Sardy points out another good competitor of the scree test: quantile universal thresholding (QUT; [Giacobino et al., 2017](#)).

R.3. Practical aspects of minimal-penalty algorithms (Section 7)

Émilie Devijver judiciously underlines the need for theoretical analyses of minimal-penalty algorithms *exactly as they are used*, which can differ from the general-purpose algorithms defined in Sections 2–3. This raises several theoretical open problems that are missing in the survey, such as theoretical grounds for the nested minimal-penalty algorithm (Section 7.3), or for choosing a penalty shape (Sections 7.4 and 7.6; see also Section R.3.2 below).

R.3.1. Definitions of \hat{C} (Section 7.1)

The discussions emphasize that choosing between \hat{C}_{jump} and \hat{C}_{slope} is not an easy task: Gilles Celeux argues for \hat{C}_{slope} —thinking especially of a collection of nested models—, while Émilie Lebarbier argues for \hat{C}_{jump} for change-point detection—given that both \hat{C}_{jump} and \hat{C}_{slope} have drawbacks in that framework, but the ones of \hat{C}_{jump} are easier to handle—, and Émilie Devijver suggests to visualize both \hat{C}_{jump} and \hat{C}_{slope} when $(S_m)_{m \in \mathcal{M}}$ is a union of subcollections! Overall, this confirms that one should not use a single definition of \hat{C} blindly, especially when using it for the first time in a given framework.

Servane Gey uses a variant of \hat{C}_{jump} —called \hat{m}_{jump} in her contribution— which is not presented in the survey: it is a nice alternative to \hat{C}_{window} for merging very close jumps.

Finally, in frameworks where both the jump and slope approaches have important drawbacks—for instance, change-point detection—, other representations of the empirical risk could be useful for estimating the minimal penalty. I mention this as an open problem, for which some inspiration might come from the suggest of David Donoho and Matan Gavish to use a histogram instead of a scree plot. In their setting, this suggest comes from modern random-matrix theory; for minimal-penalty estimation, existing and future probabilistic results might suggest other meaningful representations of the empirical risk.

R.3.2. Choice of a penalty shape (Sections 7.4 and 7.6)

The question of choosing a penalty shape, for instance between a simple and a complex one, is raised by several discussants. Bertrand Michel provides several examples where theory provides

“complex” penalty shapes —that is, depending on several unknown constants—, which makes their data-driven calibration difficult. Then, one may consider instead a “simplified” shape —depending only on one unknown constant—, but it is often not obvious that such a simplification improves the performance compared to the multivariate extension of the slope approach. Section 7.4 reports numerical experiments illustrating the difficulty of this choice, and the numerical experiments of Émilie Lebarbier for change-point detection provide another example where both options face practical issues.

Then, two strategies are proposed by the discussants. On the one hand, Émilie Devijver shows that a precise enough mathematical study —with upper *and* lower risk bounds— can help answering the question. On the other hand, Bertrand Michel asks for a data-driven procedure choosing among several penalty shapes —which is another (elaborate) estimator selection problem—, *with theoretical grounds*. Both options raise challenging mathematical open problems, and I think that solving any of them would have a significant practical impact.

The second strategy could be solved thanks to minimal-penalty algorithms, because, as written by Bertrand Michel, “it is very easy to detect when the method goes wrong”. This suggests an easy way of discarding a candidate penalty shape³. Nevertheless, Jean-Patrick Baudry shows that *validating* a penalty shape is a more difficult issue: according to some ongoing research about model-based clustering, the slope heuristics can fail —by underfitting— even when the empirical risk has a clear linear behaviour! This can be an issue about which practitioners using minimal-penalty algorithms should be warned. Fortunately, minimal-penalty algorithms cannot strongly overfit by construction, so this issue should only have a moderate impact on the risk, except in pathological settings.

R.4. Conjectures and open problems on minimal-penalty algorithms (Sections 8.3 and 8.5)

R.4.1. Need for theory with various risk functions

Yannick Baraud clearly shows that the least-squares risk —which is considered by most theoretical results about minimal-penalty algorithms— can be a bad choice, except in the Gaussian framework; Lucien Birgé also argues in this direction. This points out the strong need for theoretical results on minimal-penalty algorithms (and beyond) with various risk functions. Let me explain why.

Following Breiman (2001), two main points of view exist in statistical modeling. On the one hand, one can assume a stochastic data model and aim at estimating its parameters, in order to get an estimation of the full data-generation mechanism. With such a goal, the performance measure of any procedure should reflect the quality of this estimation. As a consequence, as explained by Yannick Baraud and Lucien Birgé, the Hellinger or total-variation distances between probability distributions should be used, while the least-squares risk can be misleading for non-Gaussian data in the regression framework.

On the other hand, one can consider the data-generation mechanism as a black box and stick to predictive accuracy (with respect to some application-dependent risk function) for evaluating

³ We implicitly assume here that the slope heuristics is valid, or at least that the minimal and optimal penalties are proportional. Otherwise, as in the setting of Section 3.3, a data-driven choice of the optimal-penalty shape seems to be too difficult a problem.

the performance of a statistical procedure. This can be seen as the “machine-learning” point of view, as opposed to the previous one which could be called “statistical”.

In both cases, the key observation is that one should use a performance measure —that is, a risk function, in the setting of Section 3.1— that is *adapted to the final goal*. If one wants to model precisely the full data-generation mechanism, the risk should be a distance between probability distributions (Hellinger or total variation, for instance). If the final goal is to build an efficient predictor, one should use a prediction risk with an appropriate cost function (such as the quadratic, absolute value, hinge, logistic or exponential loss in binary classification). And other goals are possible, depending on the task to be solved (for instance, estimating a specific part of the data-generation mechanism, as in semi-parametric statistics).

On the contrary, choosing a risk function because it is the only one for which theoretical results are available is not satisfactory at all. So, there must be a strong effort by theoreticians to provide mathematical guarantees for various performance measures, adapted to practical needs. In particular, in addition to the open problems already mentioned in the survey, an important open problem would be to provide minimal-penalty algorithms with theoretical guarantees in terms of Hellinger or total-variation distance. Lucien Birgé points out a more specific interesting open problem that is related to the above question: could a minimal-penalty algorithm (or an observable phase transition, as suggested in Section 8.5) be used for the fine tuning of the penalty appearing inside ρ -estimators (Baraud et al., 2017; Baraud and Birgé, 2018)?

Up to now, the most advanced results about minimal-penalty algorithms with general risk functions are the ones of Saumard (2010, Chapters 7–8) on minimum-contrast estimators with a regular contrast and the associated risk. In the discussion, Adrien Saumard proposes an approach for 0–1 classification that might work more generally for bounded M-estimation, hence leading to further improvements in this direction. Nevertheless, obtaining results for the Hellinger and total-variation distances requires new ideas, since these risk functions are not the expectation of a contrast.

R.4.2. Time-series analysis

I would like to thank Adrien Saumard for pointing out an important open problem, which is to develop minimal-penalty algorithms for time-series analysis, for which cross-validation and resampling procedures can be difficult to use for estimator selection. A subsection on this topic was indeed missing in Section 8.3.

In addition to the references provided by Adrien Saumard and a few results mentioned in the survey that deal with dependent data (Lerasle, 2011; Garivier and Lerasle, 2011), a framework in which theoretical results can be obtained with dependent data is the one of Section 2 with assumption (HG) replaced by $\varepsilon \sim \mathcal{N}(0, \Sigma)$ for some general covariance matrix Σ . We can always write $\Sigma = A^\top A$ with $A \in \mathcal{M}_n(\mathbb{R})$ symmetric, so that $\varepsilon = A\xi$ with $\xi \sim \mathcal{N}(0, I_n)$. Then, by Eq. (5)–(6), we obtain the expectations of the risk and empirical risk:

$$\mathbb{E} \left[\frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right] = \frac{1}{n} \left\| (I_n - \Pi_m) F \right\|^2 + \frac{\text{tr}(\Sigma \Pi_m)}{n},$$

and

$$\mathbb{E} \left[\frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 \right] = \frac{1}{n} \left\| (I_n - \Pi_m) F \right\|^2 + \frac{\text{tr}(\Sigma)}{n} - \frac{\text{tr}(\Sigma \Pi_m)}{n},$$

so that

$$\text{pen}_{\text{opt}}^{\Sigma}(m) := \text{pen}_{\text{opt},0}^{\Sigma}(m) + \frac{\text{tr}(\Sigma)}{n} = \frac{2\text{tr}(\Sigma\Pi_m)}{n}$$

is an optimal penalty, while

$$\text{pen}_{\text{min}}^{\Sigma}(m) = \frac{\text{tr}(\Sigma\Pi_m)}{n}$$

is a minimal penalty. A remarkable consequence is that

$$\text{pen}_{\text{opt}}^{\Sigma}(m) = 2\text{pen}_{\text{min}}^{\Sigma}(m)$$

holds true whatever the noise dependence structure! Furthermore, the proof of Theorem 1 can certainly be extended to this setting, with

$$\hat{m}(C) \in \underset{m \in \mathcal{M}}{\text{argmin}} \left\{ \frac{1}{n} \left\| \hat{F}_m - Y \right\|^2 + C \frac{\text{tr}(\Sigma\Pi_m)}{n} \right\},$$

D_m replaced by $\mathcal{C}_m = \text{tr}(\Sigma\Pi_m)$, a phase transition at $C = 1$, and an optimal oracle inequality for $C = 2$. The key argument is to get an equivalent of Eq. (23)–(24) in the dependent case, which can follow from standard Gaussian concentration results (Arlot and Bach, 2011, Propositions 4 and 6) under appropriate assumptions on Σ and $(\Pi_m)_{m \in \mathcal{M}}$. We leave this problem as an exercise to the reader. It seems reasonable to conjecture that the results of Arlot and Bach (2011) on linear estimators can also be extended to dependent data, under appropriate conditions on Σ and the matrices $(A_m)_{m \in \mathcal{M}}$.

However, the above result provides theoretical grounds to a slope-heuristics algorithm that can be used *only if* $\text{tr}(\Sigma\Pi_m)$ is known (up to a multiplicative constant) for all $m \in \mathcal{M}$. Such an assumption is strong, but it might be satisfied in some specific cases, either exactly—for instance for linear inverse problems with independent homoscedastic noise—or approximately—with assumptions on the dependence structure.

Let us finally remark that the above arguments can also serve as a starting point for analyzing the heteroscedastic-regression setting mentioned in point 1 of the first paragraph of Section 8.3.7.

R.4.3. Supervised classification (Section 8.3.1)

Servane Gey's contribution complements Section 8.3.1 (and the end of Section 8.3.7) well, with some numerical experiments about the slope heuristics for tuning CART in binary classification. This should increase the motivation for working on this open problem, and Adrien Saumard actually provides us a well-motivated conjecture on this very topic.

R.4.4. Model-based clustering (Section 8.3.2)

Minimal penalties for clustering is an active field of research—for good reasons since cross-validation cannot directly be applied for clustering, as noticed by Adrien Saumard—and the discussion reflects it.

Christine Keribin provides a new example where the slope heuristics empirically works well for model-based clustering, despite the fact that all models are biased. I do not dare suggest as an

additional open problem the validation of the slope heuristics in this setting —with a significant bias for all models—, given the lack of theory available for slope heuristics and model-based clustering.

Two discussants show the difficulty of developing a well-founded and practical minimal-penalty algorithm for model-based clustering. Adrien Saumard points out some mathematical challenges that may explain the lack of theory in this framework. I hope his suggests will help solving the problem. Jean-Patrick Baudry reports that some extra-parameters of the algorithms used in practice for model-based clustering may induce some troubles for the slope heuristics, leading to underfitting.

Jean-Patrick Baudry’s discussion underlines a more general issue: when the (theoretical) estimators $\hat{s}_m(D_n)$ cannot be computed exactly in practice, one should not study the estimator-selection problem among $(\hat{s}_m)_{m \in \mathcal{M}}$ but the (more difficult) problem of choosing among the *practical approximations* to these estimators.

R.4.5. Large collections of models (Section 8.3.4)

Model selection among a “large” collection (using the terminology introduced in Section 4.7) appears at several places of the discussion.

For change-point detection, Émilie Lebarbier provides numerical experiments that support the conjecture of problem (a) in Section 8.3.4, with $C^* = \sigma^2$, which would be a useful result for reasons detailed in Section 8.3.4.

For orthonormal variable selection by hard thresholding, Sylvain Sardy suggests that quantile universal thresholding (QUT; [Giacobino et al., 2017](#)) can work well. I conjecture that minimal-penalty algorithms can also work well in such a setting with a large collection of models, because the richness of the collection makes the minimal penalty much larger, as shown by partial theoretical results and numerical experiments reported in Sections 4.7 and 8.3.4).

Nevertheless, a major difficulty for the analysis and the practical use of minimal penalties with large collections of models is that the minimal and optimal penalties certainly depend strongly on the noise distribution, which should be added to the conjectures formulated in Section 8.3.4. This is explained well by Lucien Birgé, who points out the role of large-deviation arguments for the analysis of minimal and optimal penalties in this context. David Donoho and Matan Gavish show that a similar phenomenon appears in a slightly different context (factor analysis), as detailed above in Section R.2.2. From the practical point of view, the key problem is that minimal-penalty algorithms can only be used if the relationship between the minimal and optimal penalty is known *before* observing the data. In particular, if this relationship depends on the (unknown) noise distribution, and if it cannot be easily estimated from the data, then minimal-penalty algorithms are useless. In such cases, the only hope would be to have a universal relationship between the minimal penalty and a suboptimal (but good) penalty, that is, a penalty satisfying an oracle inequality with a bounded leading constant. Hopefully, this seems to be possible in the settings respectively mentioned by David Donoho and Matan Gavish and by Lucien Birgé.

R.4.6. *Infinite estimator collections (Section 8.3.5)*

Pierre Bellec shows the noticeable fact that some infinite estimator collections —ordered linear smoothers— can be tuned optimally —with an oracle inequality with leading constant *equal to 1* and a remainder term of order σ^2/n — by an aggregation procedure (Bellec and Yang, 2019). This phenomenon can be related to the result by Arlot and Bach (2011) on kernel ridge regression, that is mentioned in Section 8.3.5. Let me nevertheless point out three important differences. First, the result by Bellec and Yang (2019) makes a milder assumption on the estimator collection: kernel ridge regressors form a family of ordered linear smoothers, but other examples exist (see for instance Kneip, 1994, Section 2). Second, the result by Bellec and Yang (2019) does not rely on some explicit discretization of the estimator collection —chaining arguments make such a discretization only implicitly—, hence it could be easier to generalize further, for instance to multivariate families of kernel ridge regressors. Third, the key deviation bound proved by Bellec and Yang (2019) includes a non-negative term $c\|(A_\lambda - A^*)\|^2$, while generalizing the results of Arlot and Bach (2011) would require a similar uniform deviation bound *without* this term.

The first two differences make the result of Bellec and Yang (2019) quite interesting for generalizing the analysis of minimal-penalty algorithms, but the third difference make me join Pierre Bellec’s interrogative statement at the end of his discussion. At least, this should motivate further work on the topic.

At a broader level, this underlines that analyzing penalization and aggregation procedures require similar concentration inequalities. Therefore, solving the open problems mentioned in Section 8.6 would have consequences far beyond minimal-penalty algorithms, including the theoretical analysis of aggregation procedures such as Q-aggregation.

R.4.7. *Phase transitions (Section 8.5)*

As mentioned in Section R.4.1, Lucien Birgé suggests to consider the problem of tuning ρ -estimators thanks to some kind of minimal penalty. This open problem might share some common points with the use of phase transitions for tuning Goldenshluger-Lepski’s and related procedures (Section 8.5.1), because both are based on pairwise comparisons. Note however that the theoretical results mentioned in Section 8.5.1 are for the least-squares risk, while the tuning of ρ -estimators should be done for the Hellinger distance, which increases the difficulty of the challenge.

The use of phase transitions for tuning thresholding procedures (Section 8.5.2) might also be related to the setting of David Donoho and Matan Gavish’s discussion (see also the comments made in Sections R.2.2 and R.4.5 above). In particular, I notice that the minimal and optimal thresholds depend on the noise distribution —through a multiplicative factor σ in the Gaussian case—, which is a problem if this distribution is not known. If the minimal threshold can be estimated from data —for instance, with an histogram representation, as suggested by David Donoho and Matan Gavish—, it will provide valuable information about the noise distribution, which can then be used for computing a (near-)optimal data-driven threshold. At least, in the Gaussian case, such a procedure could be used for estimating σ , even if other strategies exist (Gavish and Donoho, 2014, Section III.E).

R.5. Overpenalization (Section 8.4)

The overpenalization problem appears at least at two places in the discussion.

Sylvain Sardy shows that the need for overpenalization can occur for tuning the adaptive Lasso, because of the high variance of the unbiased estimate of the risk. He suggests quantile universal thresholding (QUT; [Giacobino et al., 2017](#)) for solving the issue, which adds a lead to the resolution of the open problem of data-driven overpenalization mentioned in Section 8.4.

Adrien Saumard provides arguments against the fact that minimal-penalty algorithms could sharply overpenalize in general. This is an important addition to Section 8.4, in which I may have been too optimistic about the power and versatility of minimal-penalty algorithms. I am pleased to see that several of us are considering overpenalization as an important issue, and I encourage other researchers to join us on this subject and explore other ideas to solve it, well beyond minimal penalties.

References

- Arlot, S. (2019). Minimal penalties and the slope heuristics: a survey. *Journal de la SFdS*. arXiv:1901.07277.
- Arlot, S. and Bach, F. (2011). Data-driven calibration of linear estimators with minimal penalties. arXiv:0909.1884v2.
- Baraud, Y. and Birgé, L. (2018). Rho-estimators revisited: general theory and applications. *Ann. Statist.*, 46(6B):3767–3804.
- Baraud, Y., Birgé, L., and Sart, M. (2017). A new method for estimation and model selection: ρ -estimation. *Invent. Math.*, 207(2):425–517.
- Bellec, P. C. and Yang, D. (2019). The cost-free nature of optimally tuning Tikhonov regularizers and other ordered smoothers. arXiv:1905.12517v1.
- Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statist. Sci.*, 16(3):199–231.
- Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *Ann. Statist.*, 43(1):177–214.
- Garivier, A. and Lerasle, M. (2011). Oracle approach and slope heuristic in context tree estimation. arXiv:1111.2191v1.
- Gavish, M. and Donoho, D. L. (2014). The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Trans. Inform. Theory*, 60(8):5040–5053.
- Giacobino, C., Sardy, S., Diaz-Rodriguez, J., and Hengartner, N. (2017). Quantile universal threshold. *Electron. J. Statist.*, 11(2):4701–4722.
- Kneip, A. (1994). Ordered linear smoothers. *Ann. Statist.*, 22(2):835–866.
- Lerasle, M. (2011). Optimal model selection for stationary data under various mixing conditions. *Ann. Statist.*, 39(4):1852–1877.
- Saumard, A. (2010). *Estimation par Minimum de Contraste Régulier et Heuristique de Pente en Sélection de Modèles*. PhD thesis, Université de Rennes 1. Available at <http://tel.archives-ouvertes.fr/tel-00569372v1>.