Submitted to the Annals of Statistics

TECHNICAL APPENDIX TO "V-FOLD CROSS-VALIDATION IMPROVED: V-FOLD PENALIZATION"

By Sylvain Arlot

Université Paris-Sud

This is a technical appendix to "V-fold cross-validation improved: V-fold penalization". We present some additional simulation experiments, a few remarks about expectations of inverses, and the proofs which have been skipped or shortened in the main paper.

Throughout this appendix, we use the notations of the main paper [Arl08]. In order to distinguish references within the appendix from references to the main paper, we denote the former ones by (1) or 1, and the latter ones by **(1)** or **1**.

Following the ordering of [Arl08], we first present the additional simulation studies mentioned in Sect. **4**. Then, we add a few comments to Appendix **A.1**. Finally, we give some technical proofs.

1. Simulation study. We consider in this section eight experiments (called S1000, S$\sqrt{0.1}$, S0.1, Svar2, Sqrt, His6, DopReg and Dop2bin) in which we have compared the same procedures as in Sect. **4**, with the same benchmarks, but with only $N = 250$ samples for each experiment.

Data are generated according to

$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i$$

with X_i i.i.d. uniform on $\mathcal{X} = [0;1]$ and $\epsilon_i \sim \mathcal{N}(0,1)$ independent from X_i. The experiments differ from

- the regression function s:
  - S1000, S$\sqrt{0.1}$, S0.1 and Svar2 have the same smooth function as S1 and S2, see Fig. 1.
  - Sqrt has $s(x) = \sqrt{x}$, which is smooth except around 0, see Fig. 6.
  - His6 has a regular histogram with 5 jumps (hence it belongs to the regular histogram model of dimension 6), see Fig. 8.
  - DopReg and Dop2bin have the Doppler function, as defined by Donoho and Johnstone [DJ95], see Fig. 10.

AMS 2000 subject classifications: Primary 62G09; secondary 62G08, 62M20

Keywords and phrases: non-parametric statistics, statistical learning, resampling, non-asymptotic, V-fold cross-validation, model selection, penalization, non-parametric regression, adaptivity, heteroscedastic data

- the noise level σ:
  - $\sigma(x) = 1$ for S1000, Sqrt, His6, DopReg and Dop2bin.
  - $\sigma(x) = \sqrt{0.1}$ for S$\sqrt{0.1}$.
  - $\sigma(x) = 0.1$ for S0.1.
  - $\sigma(x) = \mathbb{1}_{x \geq 1/2}$ for Svar2.
- the sample size n:
  - $n = 200$ for S$\sqrt{0.1}$, S0.1, Svar2, Sqrt and His6.
  - $n = 1000$ for S1000.
  - $n = 2048$ for DopReg and Dop2bin.
- the family of models: with the notations introduced in Sect. **4**,
  - for S1000, S$\sqrt{0.1}$, S0.1, Sqrt and His6, we use the "regular" collection, as for S1:
$$\mathcal{M}_n = \left\{ 1, \dots, \left\lfloor \frac{n}{\ln(n)} \right\rfloor \right\} \ .$$
  - for Svar2, we use the "regular with two bin sizes" collection, as for S2:
$$\mathcal{M}_n = \{1\} \cup \left\{ 1, \dots, \left\lfloor \frac{n}{2\ln(n)} \right\rfloor \right\}^2 \ .$$
  - for DopReg, we use the "regular dyadic" collection, as for HSd1:
$$\mathcal{M}_n = \left\{ 2^k \text{ s.t. } 0 \leq k \leq \ln_2(n) - 1 \right\} \ .$$
  - for Dop2bin, we use the "regular dyadic with two bin sizes" collection, as for HSd2:
$$\mathcal{M}_n = \{1\} \cup \left\{ 2^k \text{ s.t. } 0 \leq k \leq \ln_2(n) - 2 \right\}^2 \ .$$

Notice that contrary to HSd2, Dop2bin is an homoscedastic problem. The interest of considering two bin sizes for it is that the smoothness of the Doppler function is quite different for small x and for $x \geq 1/2$.

Instances of data sets for each experiment are given in Fig. 2–5, 7, 9 and 11.

Compared to S1, S2, HSd1 and HSd2, these eight experiments consider larger signal-to-noise ratio data (S1000, S$\sqrt{0.1}$, S0.1), another kind of heteroscedasticity (Svar2) and other regression functions, with different kinds of unsmoothness (Sqrt, His6, DopReg and Dop2bin).

We consider for each of these experiments the same algorithms as in Sect. **4**, adding to them Mal*, which is Mallows' C_p penalty with the true value of the variance: $\mathrm{pen}(m) = 2\mathbb{E}\left[\sigma^2(X)\right] D_m n^{-1}$. Although it can not be used on real data sets, it is an interesting point

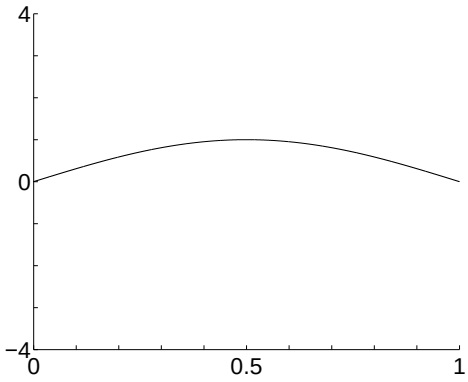

FIG 1. $s(x) = \sin(\pi x)$

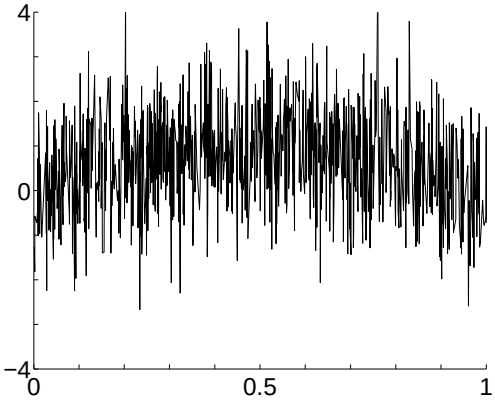

FIG 2. *Data sample for S1000*

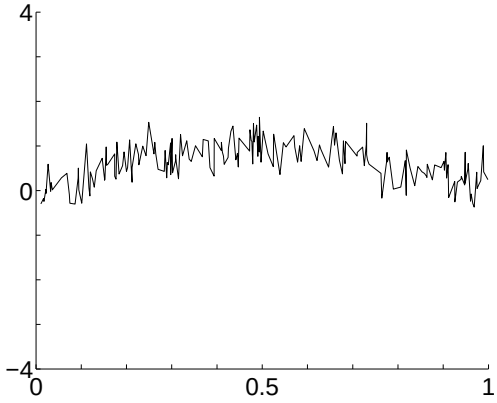

FIG 3. *Data sample for* $S\sqrt{0.1}$

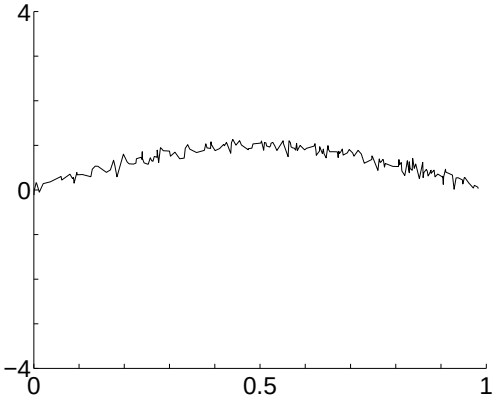

FIG 4. *Data sample for S0.1*

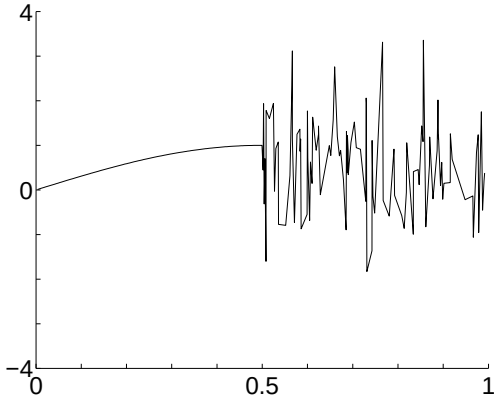

FIG 5. *Data sample for Svar2*

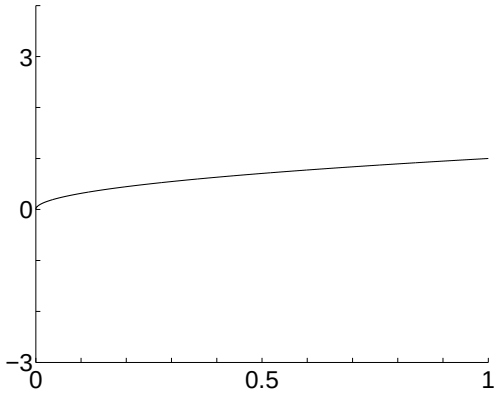

FIG 6. $s(x) = \sqrt{x}$

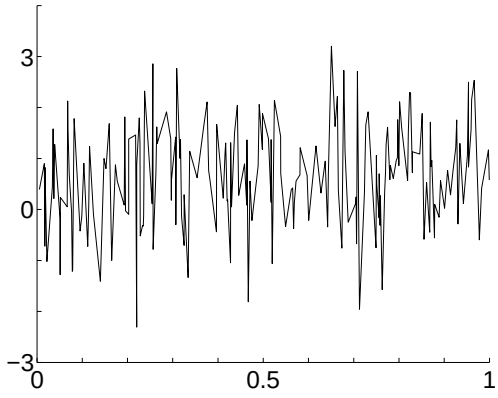

FIG 7. *Data sample for Sqrt*

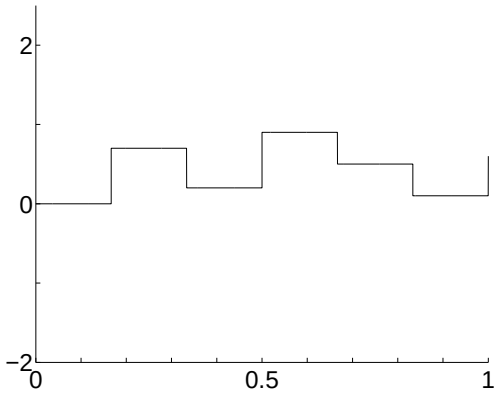

FIG 8. $s(x) = \mathrm{His}_6(x)$

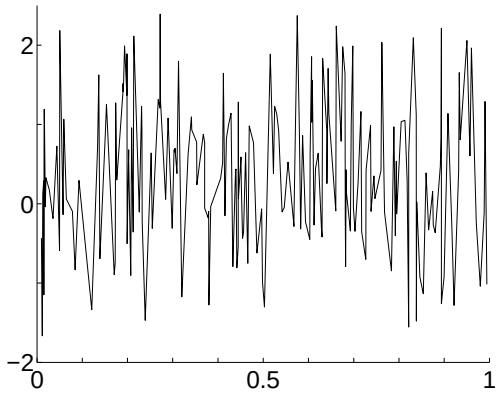

FIG 9. *Data sample for His6*

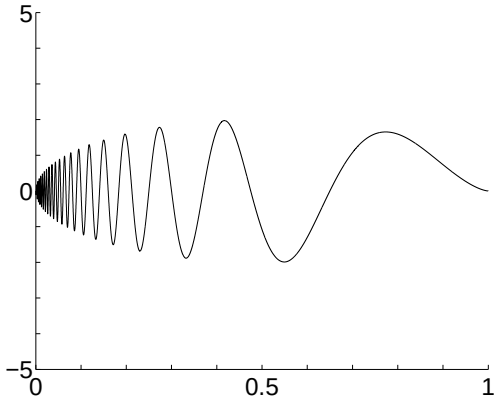

FIG 10. $s(x) = \mathrm{Doppler}(x)$ *(see [DJ95])*

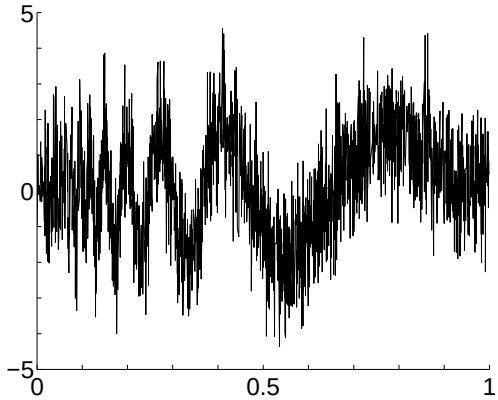

FIG 11. *Data sample for DopReg and Dop2bin*

of comparison, which does not have possible weaknesses coming from the variance estimator $\widehat{\sigma}^2$. Our estimates of C_{or} (and uncertainties for these estimates) for the procedures we consider are reported in Tab. 1 to 3 (we report here again the results for S1, S2, HSd1 and HSd2 to make comparisons easier). On the last line of these Tables, we also report

$$\frac{\mathbb{E}\left[\inf_{m\in\mathcal{M}_n} l(s,\widehat{s}_m)\right]}{\inf_{m\in\mathcal{M}_n}\left\{\mathbb{E}\left[l(s,\widehat{s}_m)\right]\right\}} = \frac{C'_{\text{or}}}{C_{\text{or}}} \qquad \text{where} \qquad C'_{\text{or}} := \frac{\mathbb{E}\left[l(s,\widehat{s}_{\widehat{m}})\right]}{\inf_{m\in\mathcal{M}_n}\left\{\mathbb{E}\left[l(s,\widehat{s}_m)\right]\right\}}$$

is the leading constant which appear in most of the classical oracle inequalities. Notice that C'_{or} is always smaller than C_{or}.

It appears that the choice of V is still difficult for VFCV: $V = 2$ is optimal in S1000 and Sqrt and $V = 20$ in the six other ones. On the contrary, $V = n$ is (almost) always better for penVF and penVF+, and overpenalization often improves the quality of the algorithm (but not always: see DopReg and S0.1). These eight experiments mainly show that the assumptions of Thm. **2** are not necessary for penVF to be efficient.

For the sake of completeness, we also reported the results for the twelve experiments in terms of the other benchmark

$$C_{\text{path}-\text{or}} := \mathbb{E}\left[\frac{l(s,\widehat{s}_{\widehat{m}})}{\inf_{m\in\mathcal{M}_n} l(s,\widehat{s}_m)}\right]$$

in Tab. 4 to Tab. 6. They are indeed quite similar to the previous ones.

2. Addendum to Appendix A.1. Whereas Lemma **3** is stated for the particular case of Binomial variables, it is worth noticing that ingredients of its proof can be successfully used in order to derive non-asymptotic bounds on $e^+_{\mathcal{L}(Z)}$ or $e^0_{\mathcal{L}(Z)}$ for several other distributions than the Binomial one. This has for instance be used in Sect. 6.7 of [Arl07] for the Hypergeometric and Poisson case.

First, the lower bound in **(15)** comes from Jensen's inequality:

$$e^+_Z \geq \mathbb{P}\left(Z > 0\right) \quad .$$

Second, taking $\theta = 0.16$ in the proof of Lemma **3** gives the absolute upper bound

$$e^0_Z \leq \kappa_4 = 7.8$$

instead of the smaller value given by Lemma 4.1 of [GKKW02]. Hence, the proof of Lemma **3** only uses that $\mathbb{P}(0 < Z < c_Z) = 0$ for some $c_Z > 0$ and that Z satisfies a concentration inequality similar to Bernstein's inequality. This covers a wide class of random variables.

Finally, notice that taking $\theta = 3\ln(A)/A$ at the end of the proof of Lemma **3**, instead of $\theta = A^{-1/2}$, leads to an upper bound

$$1 + \kappa_5\sqrt{\frac{\ln(A)}{A}} \geq \sup_{np\geq A}\left\{e^+_{\mathcal{B}(n,p)}\right\}$$

for some numerical constant κ_5, showing that the rate $A^{-1/4}$ is far from optimal.

TABLE 1
Accuracy indexes C_{or} for experiments S1, S2, HSd1 and HSd2 ($N = 1000$). Uncertainties reported are empirical standard deviations divided by $\sqrt{N}$.

Experiment	S1	S2	HSd1	HSd2
s	$\sin(\pi\cdot)$	$\sin(\pi\cdot)$	HeaviSine	HeaviSine
$\sigma(x)$	1	x	1	x
n (sample size)	200	200	2048	2048
$\mathcal{M}_n$	regular	2 bin sizes	dyadic, regular	dyadic, 2 bin sizes
Mal	1.928 ± 0.04	3.687 ± 0.07	1.015 ± 0.003	1.373 ± 0.010
Mal+	1.800 ± 0.03	3.173 ± 0.07	1.002 ± 0.003	1.411 ± 0.008
Mal*	2.028 ± 0.04	2.657 ± 0.06	1.044 ± 0.004	1.513 ± 0.005
Mal*+	1.827 ± 0.03	2.437 ± 0.05	1.004 ± 0.003	1.548 ± 0.003
$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}\right]$	1.919 ± 0.03	2.296 ± 0.05	1.028 ± 0.004	1.102 ± 0.004
$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}\right]$+	1.792 ± 0.03	2.028 ± 0.04	1.003 ± 0.003	1.089 ± 0.004
2-FCV	2.078 ± 0.04	2.542 ± 0.05	1.002 ± 0.003	1.184 ± 0.004
5-FCV	2.137 ± 0.04	2.582 ± 0.06	1.014 ± 0.003	1.115 ± 0.005
10-FCV	2.097 ± 0.04	2.603 ± 0.06	1.021 ± 0.003	1.109 ± 0.004
20-FCV	2.088 ± 0.04	2.578 ± 0.06	1.029 ± 0.004	1.105 ± 0.004
LOO	2.077 ± 0.04	2.593 ± 0.06	1.034 ± 0.004	1.105 ± 0.004
pen2-F	2.578 ± 0.06	3.061 ± 0.07	1.038 ± 0.004	1.103 ± 0.004
pen5-F	2.219 ± 0.05	2.750 ± 0.06	1.037 ± 0.004	1.104 ± 0.004
pen10-F	2.121 ± 0.04	2.653 ± 0.06	1.034 ± 0.004	1.104 ± 0.004
pen20-F	2.085 ± 0.04	2.639 ± 0.06	1.034 ± 0.004	1.105 ± 0.004
penLoo	2.080 ± 0.04	2.593 ± 0.06	1.034 ± 0.004	1.105 ± 0.004
pen2-F+	2.175 ± 0.05	2.748 ± 0.06	1.011 ± 0.003	1.106 ± 0.004
pen5-F+	1.913 ± 0.03	2.378 ± 0.05	1.006 ± 0.003	1.102 ± 0.004
pen10-F+	1.872 ± 0.03	2.285 ± 0.05	1.005 ± 0.003	1.098 ± 0.004
pen20-F+	1.898 ± 0.03	2.254 ± 0.05	1.004 ± 0.003	1.098 ± 0.004
penLoo+	1.844 ± 0.03	2.215 ± 0.05	1.004 ± 0.003	1.096 ± 0.004
$C'_{\mathrm{or}}/C_{\mathrm{or}}$	0.768	0.753	0.999	0.854

TABLE 2
Accuracy indexes C_{or} for experiments S1000, S$\sqrt{0.1}$, S0.1 and Svar2 ($N = 250$). Uncertainties reported are empirical standard deviations divided by $\sqrt{N}$.

Experiment	S1000	S$\sqrt{0.1}$	S0.1	Svar2
s	$\sin(\pi\cdot)$	$\sin(\pi\cdot)$	$\sin(\pi\cdot)$	$\sin(\pi\cdot)$
$\sigma(x)$	1	$\sqrt{0.1}$	0.1	$\mathbb{1}_{x\geq 1/2}$
n (sample size)	1000	200	200	200
$\mathcal{M}_n$	regular	regular	regular	2 bin sizes
Mal	1.667 ± 0.04	1.611 ± 0.03	1.400 ± 0.02	5.643 ± 0.22
Mal+	1.619 ± 0.03	1.593 ± 0.03	1.426 ± 0.02	4.647 ± 0.22
Mal*	1.745 ± 0.05	1.925 ± 0.03	3.204 ± 0.05	4.481 ± 0.21
Mal*+	1.617 ± 0.03	2.073 ± 0.04	3.641 ± 0.07	3.544 ± 0.17
$\mathbb{E}\left[\text{pen}_{\text{id}}\right]$	1.745 ± 0.05	1.571 ± 0.03	1.373 ± 0.02	2.409 ± 0.13
$\mathbb{E}\left[\text{pen}_{\text{id}}\right]$+	1.617 ± 0.03	1.554 ± 0.03	1.392 ± 0.02	2.005 ± 0.10
2-FCV	1.668 ± 0.04	1.663 ± 0.04	1.394 ± 0.02	2.960 ± 0.15
5-FCV	1.756 ± 0.07	1.693 ± 0.04	1.393 ± 0.02	2.950 ± 0.16
10-FCV	1.746 ± 0.04	1.684 ± 0.04	1.385 ± 0.02	2.681 ± 0.14
20-FCV	1.774 ± 0.05	1.645 ± 0.03	1.382 ± 0.02	2.742 ± 0.16
LOO	1.768 ± 0.05	1.639 ± 0.04	1.379 ± 0.02	2.641 ± 0.15
pen2-F	2.066 ± 0.08	1.809 ± 0.05	1.390 ± 0.02	3.209 ± 0.18
pen5-F	1.816 ± 0.05	1.638 ± 0.04	1.400 ± 0.02	2.749 ± 0.15
pen10-F	1.783 ± 0.05	1.706 ± 0.04	1.374 ± 0.02	2.598 ± 0.15
pen20-F	1.801 ± 0.05	1.657 ± 0.03	1.385 ± 0.02	2.684 ± 0.15
penLoo	1.776 ± 0.05	1.641 ± 0.04	1.379 ± 0.02	2.656 ± 0.15
pen2-F+	1.809 ± 0.05	1.714 ± 0.04	1.416 ± 0.02	2.808 ± 0.16
pen5-F+	1.683 ± 0.04	1.616 ± 0.03	1.399 ± 0.02	2.460 ± 0.14
pen10-F+	1.627 ± 0.04	1.613 ± 0.03	1.385 ± 0.02	2.398 ± 0.14
pen20-F+	1.644 ± 0.04	1.583 ± 0.03	1.390 ± 0.02	2.316 ± 0.13
penLoo+	1.626 ± 0.03	1.587 ± 0.03	1.401 ± 0.02	2.349 ± 0.13
$C'_{\text{or}}/C_{\text{or}}$	0.8	0.801	0.816	0.779

TABLE 3
Accuracy indexes C_{or} for experiments Sqrt, His6, DopReg and Dop2bin ($N = 250$). Uncertainties reported are empirical standard deviations divided by $\sqrt{N}$.

Experiment	Sqrt	His6	DopReg	Dop2bin
s	$\sqrt{\cdot}$	His_6	Doppler	Doppler
$\sigma(x)$	1	1	1	1
n (sample size)	200	200	2048	2048
$\mathcal{M}_n$	regular	regular	dyadic, regular	dyadic, 2 bin sizes
Mal	2.295 ± 0.11	1.969 ± 0.11	1.039 ± 0.01	1.052 ± 0.01
Mal+	1.989 ± 0.08	1.799 ± 0.09	1.090 ± 0.00	1.047 ± 0.01
Mal*	2.483 ± 0.12	2.021 ± 0.11	1.013 ± 0.01	1.061 ± 0.01
Mal*+	2.075 ± 0.09	1.836 ± 0.10	1.070 ± 0.00	1.041 ± 0.01
$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}\right]$	2.365 ± 0.11	1.805 ± 0.10	1.025 ± 0.01	1.056 ± 0.01
$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}\right]$+	2.012 ± 0.09	1.632 ± 0.08	1.083 ± 0.00	1.040 ± 0.01
2-FCV	2.489 ± 0.12	2.788 ± 0.13	1.097 ± 0.00	1.165 ± 0.01
5-FCV	2.777 ± 0.16	2.316 ± 0.12	1.064 ± 0.01	1.049 ± 0.01
10-FCV	2.571 ± 0.13	2.074 ± 0.11	1.043 ± 0.01	1.051 ± 0.01
20-FCV	2.561 ± 0.12	2.071 ± 0.11	1.034 ± 0.01	1.053 ± 0.01
LOO	2.695 ± 0.14	2.059 ± 0.11	1.026 ± 0.01	1.058 ± 0.01
pen2-F	4.088 ± 0.23	3.210 ± 0.14	1.048 ± 0.01	1.062 ± 0.01
pen5-F	3.024 ± 0.18	2.485 ± 0.13	1.033 ± 0.01	1.055 ± 0.01
pen10-F	3.009 ± 0.18	2.192 ± 0.12	1.029 ± 0.01	1.056 ± 0.01
pen20-F	2.723 ± 0.14	2.150 ± 0.12	1.031 ± 0.01	1.056 ± 0.01
penLoo	2.695 ± 0.14	2.063 ± 0.12	1.026 ± 0.01	1.058 ± 0.01
pen2-F+	3.015 ± 0.17	2.728 ± 0.12	1.084 ± 0.00	1.084 ± 0.01
pen5-F+	2.409 ± 0.13	2.080 ± 0.09	1.080 ± 0.00	1.063 ± 0.01
pen10-F+	2.305 ± 0.11	1.869 ± 0.09	1.082 ± 0.00	1.050 ± 0.01
pen20-F+	2.180 ± 0.10	1.832 ± 0.09	1.079 ± 0.00	1.052 ± 0.01
penLoo+	2.152 ± 0.10	1.858 ± 0.10	1.082 ± 0.00	1.048 ± 0.01
$C'_{\mathrm{or}}/C_{\mathrm{or}}$	0.795	0.996	0.998	0.977

TABLE 4
Accuracy indexes $C_{\text{path}-\text{or}}$ for experiments S1, S2, HSd1 and HSd2 ($N = 1000$). Uncertainties reported are empirical standard deviations divided by $\sqrt{N}$.

Experiment	S1	S2	HSd1	HSd2
s	$\sin(\pi\cdot)$	$\sin(\pi\cdot)$	HeaviSine	HeaviSine
$\sigma(x)$	1	x	1	x
n (sample size)	200	200	2048	2048
$\mathcal{M}_n$	regular	2 bin sizes	dyadic, regular	dyadic, 2 bin sizes
Mal	2.064 ± 0.04	4.129 ± 0.10	1.015 ± 0.002	1.316 ± 0.010
Mal+	1.921 ± 0.03	3.500 ± 0.09	1.002 ± 0.001	1.354 ± 0.008
Mal*	2.168 ± 0.04	2.907 ± 0.07	1.045 ± 0.003	1.453 ± 0.006
Mal*+	1.941 ± 0.03	2.645 ± 0.06	1.004 ± 0.001	1.487 ± 0.005
$\mathbb{E}\left[\text{pen}_{\text{id}}\right]$	2.053 ± 0.04	2.458 ± 0.06	1.029 ± 0.003	1.050 ± 0.002
$\mathbb{E}\left[\text{pen}_{\text{id}}\right]$+	1.903 ± 0.03	2.142 ± 0.04	1.003 ± 0.001	1.038 ± 0.002
2-FCV	2.230 ± 0.05	2.755 ± 0.06	1.002 ± 0.001	1.134 ± 0.004
5-FCV	2.290 ± 0.05	2.827 ± 0.08	1.014 ± 0.002	1.064 ± 0.003
10-FCV	2.237 ± 0.05	2.832 ± 0.08	1.021 ± 0.002	1.057 ± 0.002
20-FCV	2.225 ± 0.05	2.794 ± 0.07	1.029 ± 0.003	1.054 ± 0.002
LOO	2.212 ± 0.05	2.832 ± 0.08	1.034 ± 0.003	1.053 ± 0.002
pen2-F	2.770 ± 0.07	3.340 ± 0.08	1.039 ± 0.003	1.052 ± 0.003
pen5-F	2.383 ± 0.06	2.982 ± 0.08	1.038 ± 0.003	1.053 ± 0.002
pen10-F	2.256 ± 0.05	2.867 ± 0.07	1.035 ± 0.003	1.053 ± 0.002
pen20-F	2.219 ± 0.05	2.869 ± 0.08	1.035 ± 0.003	1.053 ± 0.002
penLoo	2.215 ± 0.05	2.832 ± 0.08	1.034 ± 0.003	1.053 ± 0.002
pen2-F+	2.328 ± 0.05	2.979 ± 0.07	1.011 ± 0.002	1.056 ± 0.003
pen5-F+	2.050 ± 0.04	2.540 ± 0.06	1.006 ± 0.001	1.052 ± 0.002
pen10-F+	1.997 ± 0.03	2.436 ± 0.05	1.005 ± 0.001	1.048 ± 0.002
pen20-F+	2.018 ± 0.04	2.416 ± 0.06	1.004 ± 0.001	1.047 ± 0.002
penLoo+	1.959 ± 0.03	2.397 ± 0.06	1.004 ± 0.001	1.045 ± 0.002

TABLE 5
Accuracy indexes $C_{\mathrm{path-or}}$ for experiments S1000, S$\sqrt{0.1}$, S0.1 and Svar2 ($N = 250$). Uncertainties reported are empirical standard deviations divided by $\sqrt{N}$.

Experiment	S1000	S$\sqrt{0.1}$	S0.1	Svar2
s	$\sin(\pi\cdot)$	$\sin(\pi\cdot)$	$\sin(\pi\cdot)$	$\sin(\pi\cdot)$
$\sigma(x)$	1	$\sqrt{0.1}$	0.1	$\mathbb{1}_{x\geq 1/2}$
n (sample size)	1000	200	200	200
$\mathcal{M}_n$	regular	regular	regular	2 bin sizes
Mal	1.704 ± 0.04	1.654 ± 0.03	1.407 ± 0.02	7.212 ± 0.40
Mal+	1.670 ± 0.03	1.636 ± 0.03	1.436 ± 0.02	5.740 ± 0.34
Mal*	1.793 ± 0.04	2.018 ± 0.04	3.273 ± 0.06	5.597 ± 0.33
Mal*+	1.664 ± 0.03	2.175 ± 0.05	3.719 ± 0.08	4.284 ± 0.25
$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}\right]$	1.793 ± 0.04	1.611 ± 0.03	1.378 ± 0.01	2.785 ± 0.19
$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}\right]$+	1.194 ± 0.02	1.177 ± 0.02	1.128 ± 0.01	1.337 ± 0.07
2-FCV	1.721 ± 0.04	1.723 ± 0.04	1.400 ± 0.02	3.507 ± 0.19
5-FCV	1.801 ± 0.06	1.740 ± 0.04	1.399 ± 0.02	3.486 ± 0.24
10-FCV	1.802 ± 0.05	1.735 ± 0.04	1.388 ± 0.02	3.149 ± 0.20
20-FCV	1.832 ± 0.05	1.687 ± 0.03	1.388 ± 0.02	3.257 ± 0.23
LOO	1.815 ± 0.05	1.685 ± 0.04	1.385 ± 0.01	3.127 ± 0.24
pen2-F	2.108 ± 0.07	1.864 ± 0.05	1.394 ± 0.02	3.839 ± 0.27
pen5-F	1.852 ± 0.05	1.675 ± 0.04	1.404 ± 0.02	3.237 ± 0.23
pen10-F	1.812 ± 0.05	1.767 ± 0.04	1.381 ± 0.01	3.093 ± 0.23
pen20-F	1.839 ± 0.05	1.706 ± 0.03	1.391 ± 0.01	3.123 ± 0.23
penLoo	1.825 ± 0.05	1.687 ± 0.04	1.385 ± 0.01	3.152 ± 0.24
pen2-F+	1.852 ± 0.05	1.765 ± 0.05	1.420 ± 0.02	3.336 ± 0.23
pen5-F+	1.732 ± 0.04	1.664 ± 0.03	1.408 ± 0.02	2.890 ± 0.22
pen10-F+	1.663 ± 0.04	1.657 ± 0.03	1.394 ± 0.02	2.810 ± 0.21
pen20-F+	1.680 ± 0.04	1.623 ± 0.03	1.397 ± 0.01	2.657 ± 0.19
penLoo+	1.673 ± 0.03	1.624 ± 0.03	1.409 ± 0.02	2.659 ± 0.18

TABLE 6
Accuracy indexes $C_{\text{path}-\text{or}}$ for experiments Sqrt, His6, DopReg and Dop2bin ($N = 250$). Uncertainties reported are empirical standard deviations divided by $\sqrt{N}$.

Experiment	Sqrt	His6	DopReg	Dop2bin
s	$\sqrt{\cdot}$	His_6	Doppler	Doppler
$\sigma(x)$	1	1	1	1
n (sample size)	200	200	2048	2048
$\mathcal{M}_n$	regular	regular	dyadic, regular	dyadic, 2 bin sizes
Mal	2.557 ± 0.12	2.356 ± 0.18	1.040 ± 0.00	1.049 ± 0.00
Mal+	2.232 ± 0.10	2.041 ± 0.12	1.094 ± 0.00	1.045 ± 0.01
Mal*	2.838 ± 0.15	2.533 ± 0.21	1.013 ± 0.00	1.057 ± 0.00
Mal*+	2.349 ± 0.11	2.168 ± 0.16	1.073 ± 0.00	1.038 ± 0.00
$\mathbb{E}\left[\text{pen}_{\text{id}}\right]$	2.678 ± 0.14	2.182 ± 0.17	1.026 ± 0.00	1.053 ± 0.00
$\mathbb{E}\left[\text{pen}_{\text{id}}\right]$+	1.348 ± 0.07	1.230 ± 0.06	1.050 ± 0.00	1.038 ± 0.00
2-FCV	2.974 ± 0.17	3.713 ± 0.25	1.100 ± 0.00	1.164 ± 0.01
5-FCV	3.209 ± 0.21	2.977 ± 0.24	1.066 ± 0.00	1.046 ± 0.00
10-FCV	2.912 ± 0.16	2.639 ± 0.21	1.045 ± 0.00	1.047 ± 0.00
20-FCV	2.889 ± 0.15	2.584 ± 0.20	1.035 ± 0.00	1.050 ± 0.00
LOO	3.061 ± 0.17	2.568 ± 0.21	1.027 ± 0.00	1.055 ± 0.00
pen2-F	5.062 ± 0.37	4.462 ± 0.30	1.050 ± 0.00	1.059 ± 0.01
pen5-F	3.595 ± 0.25	3.458 ± 0.28	1.034 ± 0.00	1.052 ± 0.00
pen10-F	3.445 ± 0.22	2.744 ± 0.21	1.031 ± 0.00	1.053 ± 0.00
pen20-F	3.120 ± 0.17	2.670 ± 0.21	1.032 ± 0.00	1.053 ± 0.00
penLoo	3.063 ± 0.17	2.571 ± 0.21	1.027 ± 0.00	1.055 ± 0.00
pen2-F+	3.723 ± 0.29	3.777 ± 0.26	1.087 ± 0.00	1.082 ± 0.01
pen5-F+	2.790 ± 0.18	2.698 ± 0.19	1.083 ± 0.00	1.061 ± 0.01
pen10-F+	2.653 ± 0.14	2.364 ± 0.20	1.085 ± 0.00	1.047 ± 0.01
pen20-F+	2.497 ± 0.13	2.318 ± 0.20	1.082 ± 0.00	1.049 ± 0.01
penLoo+	2.437 ± 0.12	2.218 ± 0.18	1.085 ± 0.00	1.045 ± 0.00

3. Additional proofs.

3.1. *Proof of Lemma **6**.* In this proof, we denote by L any constant that may depend on a, b, $(c_i)_{1\leq i\leq 4}$, $(\kappa_i)_{1\leq i\leq 4}$, c_{rich} and C, possibly different from one place to another.

First of all, there is a model $m_1 \in \mathcal{M}_n$ such that

$$\ln(n)^{\kappa_1} \leq \left(2anb^{-1}\right)^{1/3} \leq D_{m_1} \leq \left(2anb^{-1}\right)^{1/3} + c_{\mathrm{rich}} \leq c_1 n(\ln(n))^{-1}$$

(at least for $n \geq L$). As a consequence, **(27)** implies that

$$\mathrm{crit}_1(m_1) \leq a^{1/3}b^{2/3}n^{-2/3}\left(3\times 2^{-2/3} + c_{\mathrm{rich}}\left(\frac{b}{an}\right)^{1/3}\right)\left(1+c_2\ln(n)^{-\kappa_2}\right) \quad . \tag{1}$$

With a similar argument, for $n \geq L$, there exists a model $m_2 \in \mathcal{M}_n$ such that

$$\mathrm{crit}_2(m_2) \leq a^{1/3}\,(bC)^{2/3}\,n^{-2/3}\left(3\times 2^{-2/3} + c_{\mathrm{rich}}\left(\frac{bC}{an}\right)^{1/3}\right)\left(1+c_2\ln(n)^{-\kappa_2}\right) \quad . \tag{2}$$

We will now derive from (2) some tight bounds on $D_{\widehat{m}}$. First, the upper bound in (2) is smaller than the lower bounds in both **(29)** and **(30)** for $n \geq L$. This proves that

$$\ln(n)^{\kappa_1} \leq D_{\widehat{m}} \leq \frac{c_1 n}{\ln(n)} \quad .$$

Then, according to **(49)**, we have for every $m \in \mathcal{M}_n$ of dimension $D_m = \left(\frac{2an}{bC}\right)^{1/3}(1+\delta)$ (which is between $\ln(n)^{\kappa_1}$ and $\frac{c_1 n}{\ln(n)}$ for $n \geq L$, as long as $1 \leq \delta > -1$):

$$\begin{aligned}\mathrm{crit}_2(m) &\geq a^{1/3}\,(bC)^{2/3}\,n^{-2/3}\left(2^{-2/3}(1+\delta)^{-2} + 2^{1/3}(1+\delta)\right)\left(1-c_2\ln(n)^{-\kappa_2}\right)\\ &\geq \mathrm{crit}_2(m_2)\times\frac{1-c_2\ln(n)^{-\kappa_2}}{1+c_2\ln(n)^{-\kappa_2}}\times\frac{f(\delta)}{3\times 2^{-2/3}+c_{\mathrm{rich}}\left(\frac{bC}{an}\right)^{1/3}}\end{aligned}$$

with f defined by $f(\delta) = 2^{-2/3}(1+\delta)^{-2} + 2^{1/3}(1+\delta)$. Using Lemma 1 below, we then have

$$\frac{\mathrm{crit}_2(m)}{\mathrm{crit}_2(m_2)} \geq \frac{1-c_2\ln(n)^{-\kappa_2}}{1+c_2\ln(n)^{-\kappa_2}}\times\frac{3\times 2^{-2/3}+3\times 2^{-14/3}\left(\delta^2\wedge 1\right)}{3\times 2^{-2/3}+c_{\mathrm{rich}}\left(\frac{bC}{an}\right)^{1/3}} \quad .$$

This lower bound is strictly larger than 1 as soon as $\delta^2 \geq \ln(n)^{-\kappa_2/2}$ and $n \geq L$, so that

$$\left(\frac{2an}{bC}\right)^{1/3}\left(1-\ln(n)^{-\kappa_2/4}\right) \leq D_{\widehat{m}} \leq \left(\frac{2an}{bC}\right)^{1/3}\left(1+\ln(n)^{-\kappa_2/4}\right) \quad . \tag{3}$$

We can now use **(27)** in order to bound $\mathrm{crit}_1(\widehat{m})$. For $n \geq L$, using again Lemma 1,

$$\begin{aligned}
\mathrm{crit}_1(\widehat{m}) &\geq a^{1/3}b^{2/3}n^{-2/3}\left(\left(\frac{C}{2}\right)^{2/3}+\left(\frac{C}{2}\right)^{-1/3}\right)\left(1-L\ln(n)^{-\kappa_2/4}\right)\\
&= a^{1/3}b^{2/3}n^{-2/3}f\left(C^{-1/3}-1\right)\left(1-L\ln(n)^{-\kappa_2/4}\right)\\
&\geq a^{1/3}b^{2/3}n^{-2/3}\left(3\times 2^{-2/3}+\left(C^{-1/3}-1\right)^2\right)\left(1-L\ln(n)^{-\kappa_2/4}\right)\\
&\geq \mathrm{crit}_1(m_1)\left(1+2^{2/3}\times 3^{-1}\left(C^{-1/3}-1\right)^2-\ln(n)^{-\kappa_2/5}\right) \ ,
\end{aligned}$$

which proves **(31)**. □

Remark 1. A similar argument proves that for $n \geq L$,

$$\mathrm{crit}_1(\widehat{m}) \leq \mathrm{crit}_1(m_1)\left(1+2^{2/3}\times 3^{-1}\left(C^{-1/3}-1\right)^2+L\ln(n)^{-\kappa_2/4}\right) \ .$$

Moreover, if crit_1 satisfies (ii) and (iii), we prove in a similar way that if $n \geq n_0$, for every $\widehat{m} \in \arg\min_{m\in\mathcal{M}_n}\mathrm{crit}_2(m)$,

$$\mathrm{crit}_1(\widehat{m}) \leq \left(1+K(C)+\ln(n)^{-\kappa_2/5}\right)\inf_{m\in\mathcal{M}_n}\{\mathrm{crit}_1(m)\} \ . \tag{4}$$

This justifies our first comment behind Thm. **1**.

Lemma 1. *Let $f:(-1,+\infty)\mapsto\mathbb{R}$ be defined by $f(x)=2^{-2/3}(1+x)^{-2}+2^{1/3}(1+x)$. Then, for every $x>-1$,*

$$f(x) \geq 3\times 2^{-2/3}+3\times 2^{-14/3}\left(x^2\wedge 1\right) \ .$$

proof of Lemma 1. We apply the Taylor-Lagrange theorem to f (which is infinitely differentiable) at order two, between 0 and x. The result follows since $f(0)=3\times 2^{-2/3}$, $f'(0)=0$ and $f''(t)=6\times 2^{-2/3}\times(1+t)^{-4}\geq 3\times 2^{1/3-4}$ if $t\leq 1$. If $t>1$, the result follows from the fact that $f'\geq 0$ on $[0,+\infty)$. □

3.2. *End of the proof of Prop. 2.* We here compute $R_{1,\widetilde{W}}(n,\widehat{p}_\lambda)$ and $R_{2,\widetilde{W}}(n,\widehat{p}_\lambda)$ when V does not divide $n\widehat{p}_\lambda$, that we have skipped in Appendix **B.4.2**.

Since $(\widetilde{W}_i)_{X_i\in I_\lambda}$ is exchangeable and $\widetilde{W}_i$ takes only two values,

$$W_\lambda = \mathbb{E}_W\left[W_i \mid W_\lambda\right] = \frac{V}{V-1}\mathbb{P}\left(W_i=\frac{V}{V-1}\,\middle|\, W_\lambda\right) \ .$$

Thus,

$$\mathcal{L}\left(W_i \mid W_\lambda\right) = \frac{V}{V-1}\mathcal{B}(\kappa^{-1}W_\lambda)$$

so that

$$R_{2,W}(n,\widehat{p}_\lambda) = \frac{1}{V-1} \qquad \text{and} \qquad R_{1,W}(n,\widehat{p}_\lambda) = \frac{V}{V-1}\mathbb{E}\left(\widetilde{W}_\lambda^{-1}\right) - 1 \ .$$

There exists $a, b \in \mathbb{N}$ such that $0 \leq b \leq V-1$ and $n\widehat{p}_\lambda = aV + b$. Then,

$$\mathbb{P}\left(\widetilde{W}_\lambda = \frac{V(a(V-1)+b)}{(V-1)(aV+b)}\right) = \frac{V-b}{V} \quad \text{and} \quad \mathbb{P}\left(\widetilde{W}_\lambda = \frac{V(a(V-1)+b-1)}{(V-1)(aV+b)}\right) = \frac{b}{V}$$

so that

$$\begin{aligned}\mathbb{E}\left[\widetilde{W}_\lambda^{-1}\right] &= \frac{V-b}{V}\frac{(V-1)(aV+b)}{V(a(V-1)+b)} + \frac{b}{V}\frac{(V-1)(aV+b)}{V(a(V-1)+b-1)} \\ &= 1 - \frac{b}{V(a(V-1)+b)} + \frac{(V-1)(aV+b)b}{V^2(a(V-1)+b-1)(a(V-1)+b)} \ .\end{aligned}$$

We deduce

$$R_{1,\widetilde{W}}(n,\widehat{p}_\lambda) = \frac{1}{V-1} - \frac{b}{(V-1)(a(V-1)+b)} + \frac{(aV+b)b}{V(a(V-1)+b-1)(a(V-1)+b)} \ .$$

The result follows with

$$\delta^{(\mathrm{penV})}_{n,\widehat{p}_\lambda} = \frac{b}{n\widehat{p}_\lambda - a}\left(\frac{V-1}{V} \times \frac{n\widehat{p}_\lambda}{n\widehat{p}_\lambda - a - 1} - 1\right) \in \left[0; \frac{2}{n\widehat{p}_\lambda - 2}\right] \ . \quad \square$$

3.3. *Proof of Lemma **8**.* Although this lemma can be found in [Arl07] (where it is called Lemma 5.7), we recall here its proof for the sake of completeness.

First, split the penalty (without the constant C) into these two terms:

$$\widehat{p}_1(m) = \sum_{\lambda\in\Lambda_m} \mathbb{E}_W\left[\widehat{p}_\lambda\left(\widehat{\beta}^W_\lambda - \widehat{\beta}_\lambda\right)^2 \,\middle|\, W_\lambda > 0\right] \tag{5}$$

$$\widehat{p}_2(m) = \sum_{\lambda\in\Lambda_m} \mathbb{E}_W\left[\widehat{p}^W_\lambda\left(\widehat{\beta}^W_\lambda - \widehat{\beta}_\lambda\right)^2\right] \ . \tag{6}$$

This split into two terms is the equivalent of the split of $\mathrm{pen}_{\mathrm{id}}$ into p_1 and p_2 (plus a centered term).

We first compute this quantity, which appears in both $\widehat{p}_1$ and $\widehat{p}_2$: let $\lambda \in \Lambda_m$ and $W_\lambda > 0$,

$$\begin{aligned}\mathbb{E}_W\left[\widehat{p}_\lambda\left(\widehat{\beta}^W_\lambda - \widehat{\beta}_\lambda\right)^2 \,\middle|\, W_\lambda\right] &= \mathbb{E}_W\left[\widehat{p}_\lambda\left(\frac{1}{n\widehat{p}_\lambda}\sum_{X_i\in I_\lambda}(Y_i-\beta_\lambda)\left(1-\frac{W_i}{W_\lambda}\right)\right)^2 \,\middle|\, W_\lambda\right] \\ &= \frac{1}{n^2\widehat{p}_\lambda}\left[\sum_{X_i\in I_\lambda}(Y_i-\beta_\lambda)^2\,\mathbb{E}_W\left[\left(1-\frac{W_i}{W_\lambda}\right)^2 \,\middle|\, W_\lambda\right]\right. \\ &\quad \left. + \frac{1}{n^2\widehat{p}_\lambda}\sum_{i\neq j, X_i\in I_\lambda, X_j\in I_\lambda}(Y_i-\beta_\lambda)(Y_j-\beta_\lambda)\mathbb{E}_W\left[\left(1-\frac{W_i}{W_\lambda}\right)\left(1-\frac{W_j}{W_\lambda}\right) \,\middle|\, W_\lambda\right]\right] \ . \end{aligned} \tag{7}$$

Since the weights are exchangeable, $(W_i)_{X_i\in I_\lambda}$ is also exchangeable conditionally to W_λ and $(X_i)_{1\le i\le n}$. Thus, the "variance" term

$$R_V(n,n\widehat{p}_\lambda,W_\lambda,\mathcal{L}(W)) := \mathbb{E}_W\left[(W_i-W_\lambda)^2\,\middle|\,W_\lambda\right]$$

does not depend from i (provided that $X_i\in I_\lambda$), and the "covariance" term

$$R_C(n,n\widehat{p}_\lambda,W_\lambda,\mathcal{L}(W)) := \mathbb{E}_W\left[(W_i-W_\lambda)(W_j-W_\lambda)\,|\,W_\lambda\right]$$

does not depend from (i,j) (provided that $i\neq j$ and $X_i,X_j\in I_\lambda$). Moreover,

$$\begin{aligned}0&=\mathbb{E}_W\left[\left(\sum_{X_i\in I_\lambda}(W_i-W_\lambda)\right)^2\,\middle|\,W_\lambda\right]\\&=n\widehat{p}_\lambda R_V(n,n\widehat{p}_\lambda,W_\lambda,\mathcal{L}(W))+n\widehat{p}_\lambda(n\widehat{p}_\lambda-1)R_C(n,n\widehat{p}_\lambda,W_\lambda,\mathcal{L}(W))\end{aligned}$$

so that, if $n\widehat{p}_\lambda\ge 2$,

(8)
$$R_C(n,n\widehat{p}_\lambda,W_\lambda,W)=\frac{-1}{n\widehat{p}_\lambda-1}R_V(n,n\widehat{p}_\lambda,W_\lambda,\mathcal{L}(W))\qquad\text{and}\qquad R_V(n,1,W_\lambda,\mathcal{L}(W))=0\ .$$

Combining (7) and (8), we obtain

$$\begin{aligned}\mathbb{E}_W\left[\widehat{p}_\lambda\left(\widehat{\beta}_\lambda^W-\widehat{\beta}_\lambda\right)^2\,\middle|\,W_\lambda\right]&=\frac{R_V(n,n\widehat{p}_\lambda,W_\lambda,\mathcal{L}(W))}{W_\lambda n^2\widehat{p}_\lambda}\mathbb{1}_{n\widehat{p}_\lambda\ge 2}\\&\quad\times\left[\frac{n\widehat{p}_\lambda}{n\widehat{p}_\lambda-1}S_{\lambda,2}-\frac{1}{n\widehat{p}_\lambda-1}S_{\lambda,1}^2\right]\end{aligned}\tag{9}$$

Combining (9) and (5) (resp. (9) and (6)), we have the following expressions for $\widehat{p}_1$ and $\widehat{p}_2$:

$$\widehat{p}_1(m)=\sum_{\lambda\in\Lambda_m}\frac{R_{1,W}(n,\widehat{p}_\lambda)\mathbb{1}_{n\widehat{p}_\lambda\ge 2}}{n^2\widehat{p}_\lambda}\left[\frac{n\widehat{p}_\lambda}{n\widehat{p}_\lambda-1}S_{\lambda,2}-\frac{1}{n\widehat{p}_\lambda-1}S_{\lambda,1}^2\right]\tag{10}$$

$$\widehat{p}_2(m)=\sum_{\lambda\in\Lambda_m}\frac{R_{2,W}(n,\widehat{p}_\lambda)\mathbb{1}_{n\widehat{p}_\lambda\ge 2}}{n^2\widehat{p}_\lambda}\left[\frac{n\widehat{p}_\lambda}{n\widehat{p}_\lambda-1}S_{\lambda,2}-\frac{1}{n\widehat{p}_\lambda-1}S_{\lambda,1}^2\right]\ .\tag{11}$$

Remark that the terms of the sum for which $n\widehat{p}_\lambda=1$ are all equal to zero, which can be ensured with the convention $0\times\infty=0$ since $R_{1,W}(n,n^{-1})=R_{2,W}(n,n^{-1})=0$. The result follows. □

3.4. *Concentration of $\widetilde{p_1}$: detailed proof.* Within the proof of Prop. **9**, we used Lemma **4** in order to control the deviations of $\mathbb{E}^{\Lambda_m}[\widetilde{p_1}(m)]$ around its expectation. Implicitly, we used the following lemma (which is indeed a straightforward consequence of Lemma **4**).

Lemma 2. *We assume that* $\min_{\lambda\in\Lambda_m}\{np_\lambda\}\ge B_n\ge 1$.

1. *Lower deviations: let* $c_1 = 0.184$. *For all* $x \geq 0$, *with probability at least* $1 - e^{-x}$,

$$\mathbb{E}^{\Lambda_m}\left[\widetilde{p_1}(m)\right] \geq \mathbb{E}\left[\widetilde{p_1}(m)\right] - \theta^-(x, B_n, D_m, A, \sigma_{\min}) \times \mathbb{E}\left[p_2(m)\right] \tag{12}$$

$$\text{with} \qquad \theta^- := L\left[\varphi_1(c_1 B_n) + \frac{A^2}{\sigma_{\min}^2}\sqrt{e^{-c_1 B_n} + \frac{x}{D_m}}\right]$$

2. *Upper deviations: let* $c_2 = 0.28$ *and* $c_4 = 0.09$. *For every* $x \geq 0$, *with probability at least* $1 - e^{-x}$,

$$\mathbb{E}^{\Lambda_m}\left[\widetilde{p_1}(m)\right] \leq \mathbb{E}\left[\widetilde{p_1}(m)\right] + \theta^+(x, B_n, D_m, A, \sigma_{\min})\mathbb{E}\left[p_2(m)\right] \tag{13}$$

$$\text{with} \qquad \theta^+ := L\left[\varphi_1\left(c_2 B_n\right) + \frac{A^2}{\sigma_{\min}^2}\sqrt{x D_m^{-1} + e^{-c_4 B_n}}\left(1 \vee \sqrt{x + D_m e^{-c_4 B_n}}\right)\right] \ .$$

Proof. From **(19)** and **(37)**, we have an explicit expression for $\widetilde{p_1}$. We then apply Lemma **4**, with $X_\lambda = n\widehat{p}_\lambda$ and $a_\lambda = p_\lambda\left(\sigma_\lambda\right)^2 \geq 0$. For θ^+, we used the general upper bound

$$\max_{\lambda \in \Lambda_m}\left(\sigma_\lambda\right)^4 \left(\sum_{\lambda \in \Lambda_m} \sigma_\lambda^4\right)^{-1} \leq 1 \ .$$

□

Remark 2. If $B_n \geq \left(c_1^{-1} \vee c_4^{-1}\right)\ln(n)$, for every $\gamma > 0$,

$$\theta^- \vee \theta^+\left(\gamma \ln(n), B_n, D_m, A, \sigma_{\min}\right) \leq L_\gamma A^2 \sigma_{\min}^{-2} D_m^{-1/2} \ln(n)$$

since $D_m \leq n$.

REFERENCES

[Arl07] Sylvain Arlot. *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11, December 2007. Available online at `http://tel.archives-ouvertes.fr/tel-00198803/en/`.

[Arl08] Sylvain Arlot. V-fold cross-validation improved: V-fold penalization, February 2008. Preprint. arXiv:0802.0566.

[DJ95] David L. Donoho and Iain M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432):1200–1224, 1995.

[GKKW02] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

Sylvain Arlot
Univ Paris-Sud, UMR 8628,
Laboratoire de Mathématiques,
Orsay, F-91405 ; CNRS, Orsay, F-91405 ;
INRIA-Futurs, Projet Select
E-mail: sylvain.arlot@math.u-psud.fr