Data-driven calibration of linear estimators with minimal penalties: Appendix

Sylvain Arlot *
CNRS ; Willow Project-Team
Laboratoire d'Informatique de
l'Ecole Normale Superieure
(CNRS/ENS/INRIA UMR 8548)
23, avenue d'Italie, F-75013 Paris, France
sylvain.arlot@ens.fr

Francis Bach †
INRIA ; Willow Project-Team
Laboratoire d'Informatique de
l'Ecole Normale Superieure
(CNRS/ENS/INRIA UMR 8548)
23, avenue d'Italie, F-75013 Paris, France
francis.bach@ens.fr

This appendix is mainly devoted to the proof of Theorem 1 in [1], which is splitted into two results. First, Proposition 1 shows that $n^{-1}\sigma^2(2\operatorname{tr}(A_\lambda)-\operatorname{tr}(A_\lambda^\top A_\lambda))$ is a minimal penalty, so that $\widehat{C}$ defined in the Algorithm of Section 4.1 in [1] consistently estimates σ^2 . Second, Proposition 2 shows that penalizing the empirical risk with $2\widehat{C}\operatorname{tr}(A_\lambda)n^{-1}$ and $\widehat{C}\approx\sigma^2$ leads to an oracle inequality. Proving Theorem 1 in [1] is straightforward by combining Propositions 1 and 2.

In Section 1, we introduce some notation and make some computations that will be used in the following. Proposition 1 is proved in Section 2. Proposition 2 is proved in Section 3. Concentration inequalities needed for proving Propositions 1 and 2 are stated and proved in Section 4. Computations specific to the kernel ridge regression example are made in Section 5.

1 Notation and first computations

Recall that

$$Y = F + \varepsilon$$

where $F=(f(x_i))_{1\leq i\leq n}\in\mathbb{R}^n$ is deterministic, $\varepsilon=(\varepsilon_i)_{1\leq i\leq n}\in\mathbb{R}^n$ is centered with covariance matrix $\sigma^2 I_n$ and I_n is the $n\times n$ identity matrix. For every $\lambda\in\Lambda$, $\widehat{F}_\lambda = A_\lambda Y$ for some $n\times n$ real-valued matrix A_λ , so that

$$\left\|\widehat{F}_\lambda - F\right\|_2^2 = \|(A_\lambda - I_n)F\|_2^2 + \|A_\lambda\varepsilon\|_2^2 + 2\left\langle A_\lambda\varepsilon,\,(A_\lambda-I_n)F\right\rangle \quad , \tag{1}$$

$$\left\|\widehat{F}_\lambda - Y\right\|_2^2 = \left\|\widehat{F}_\lambda - F\right\|_2^2 + \|\varepsilon\|_2^2 - 2\left\langle\varepsilon,\,A_\lambda\varepsilon\right\rangle + 2\left\langle\varepsilon,\,(I_n-A_\lambda)F\right\rangle \quad , \tag{2}$$

where $\forall t,u\in\mathbb{R}^n$, $\langle t,\,u\rangle=\sum_{i=1}^n t_i u_i$ and $\|t\|_2^2=\langle t,\,t\rangle$.

Now, define, for every $\lambda\in\Lambda$,

$$\begin{aligned}
b(\lambda) &= \|(A_\lambda - I_n)F\|_2^2\\
v_1(\lambda) &= \operatorname{tr}(A_\lambda)\sigma^2\\
\delta_1(\lambda) &= \langle\varepsilon,\,A_\lambda\varepsilon\rangle - \operatorname{tr}(A_\lambda)\sigma^2\\
v_2(\lambda) &= \operatorname{tr}(A_\lambda^\top A_\lambda)\sigma^2\\
\delta_2(\lambda) &= \|A_\lambda\varepsilon\|_2^2 - \operatorname{tr}(A_\lambda^\top A_\lambda)\sigma^2\\
\delta_3(\lambda) &= 2\left\langle A_\lambda\varepsilon,\,(A_\lambda-I_n)F\right\rangle\\
\delta_4(\lambda) &= 2\left\langle\varepsilon,\,(I_n-A_\lambda)F\right\rangle \quad ,
\end{aligned}$$

*http://www.di.ens.fr/∼arlot/
†http://www.di.ens.fr/∼fbach/

so that Eq. (1) and (2) can be rewritten

$$\left\|\widehat{F}_\lambda - F\right\|_2^2 = b(\lambda) + v_2(\lambda) + \delta_2(\lambda) + \delta_3(\lambda) \tag{3}$$

$$\left\|\widehat{F}_\lambda - Y\right\|_2^2 = \left\|\widehat{F}_\lambda - F\right\|_2^2 - 2v_1(\lambda) - 2\delta_1(\lambda) + \delta_4(\lambda) + \|\varepsilon\|_2^2 \ . \tag{4}$$

Note that $b(\lambda)$, $v_1(\lambda)$ and $v_2(\lambda)$ are deterministic, and for all $\lambda \in \Lambda$, all $\delta_i(\lambda)$ are random with zero mean. In particular, we deduce the following expressions of the risk and the empirical risk of $\widehat{F}_\lambda$:

$$\mathbb{E}\left[n^{-1}\left\|\widehat{F}_\lambda - F\right\|_2^2\right] = n^{-1}\left\|(A_\lambda - I_n)F\right\|_2^2 + \frac{\operatorname{tr}(A_\lambda^\top A_\lambda)\sigma^2}{n} \ , \tag{5}$$

$$\mathbb{E}\left[n^{-1}\left\|\widehat{F}_\lambda - Y\right\|_2^2\right] - \sigma^2 = n^{-1}\left\|(A_\lambda - I_n)F\right\|_2^2 - \frac{\left(2\operatorname{tr}(A_\lambda) - \operatorname{tr}(A_\lambda^\top A_\lambda)\right)\sigma^2}{n} \ . \tag{6}$$

Define

$$\|A_\lambda\| := \max \operatorname{Sp}(A_\lambda) = \sup_{t\in\mathbb{R}^n,\, t\neq 0}\left\{\frac{\|A_\lambda t\|_2}{\|t\|_2}\right\} \ .$$

Since $\operatorname{tr}(A_\lambda) \leq \sqrt{n\operatorname{tr}(A_\lambda^\top A_\lambda)}$, we have

$$v_1(\lambda) \leq \sigma\sqrt{n v_2(\lambda)} \ . \tag{7}$$

In addition, if A_λ has a spectrum $\operatorname{Sp}(A_\lambda) \subset [0,1]$, then

$$0 \leq \operatorname{tr}(A_\lambda^\top A_\lambda) \leq \operatorname{tr}(A_\lambda) \leq 2\operatorname{tr}(A_\lambda) - \operatorname{tr}(A_\lambda^\top A_\lambda) \leq 2\operatorname{tr}(A_\lambda) \ ,$$

so that

$$0 \leq v_2(\lambda) \leq v_1(\lambda) \leq 2v_1(\lambda) - v_2(\lambda) \leq 2v_1(\lambda) \ . \tag{8}$$

2 Minimal penalty

Define

$$\forall C \geq 0, \qquad \widehat{\lambda}_0(C) \in \arg\min_{\lambda\in\Lambda}\left\{\left\|\widehat{F}_\lambda - Y\right\|_2^2 + C\left(2\operatorname{tr}(A_\lambda) - \operatorname{tr}(A_\lambda^\top A_\lambda)\right)\right\} \ . \tag{9}$$

We will prove the following proposition in this section.

Proposition 1 *Let $\widehat{\lambda}_0$ be defined by Eq. (9). Assume that $\forall\lambda \in \Lambda$, A_λ is symmetric with $\operatorname{Sp}(A_\lambda) \subset [0,1]$, that ε_i are i.i.d. Gaussian with zero mean and variance $\sigma^2 > 0$, and that*

$$\exists\lambda_1 \in \Lambda, \quad \operatorname{df}(\lambda_1) \geq \frac{n}{2} \quad \textit{and} \quad b(\lambda_1) \leq \sigma^2\sqrt{n\ln(n)} \tag{$\mathbf{A_1}$}$$

$$\exists\lambda_2 \in \Lambda, \quad \operatorname{df}(\lambda_2) \leq \sqrt{n} \quad \textit{and} \quad b(\lambda_2) \leq \sigma^2\sqrt{n\ln(n)} \ . \tag{$\mathbf{A_2}$}$$

Then, a numerical constant $C_1 > 0$ exists such that for every $n \geq C_1$, for every $\gamma \geq 1$,

$$\forall 0 \leq C < \left(1 - 91\gamma\sqrt{\frac{\ln(n)}{n}}\right)\sigma^2, \quad \operatorname{df}(\widehat{\lambda}_0(C)) \geq \frac{n}{10} \tag{10}$$

$$\textit{and} \quad \forall C > \left(1 + \frac{44\gamma\sqrt{\ln(n)}}{n^{1/4}}\right)\sigma^2, \quad \operatorname{df}(\widehat{\lambda}_0(C)) \leq n^{3/4} \tag{11}$$

hold with probability at least $1 - 8\operatorname{Card}(\Lambda)n^{-\gamma}$.

If $\operatorname{Card}(\Lambda) \leq Kn^\alpha$, Proposition 1 with $\gamma = \alpha + 2$ proves that with probability at least $1 - 8Kn^{-2}$, $\widehat{C}$ defined in the Algorithm of Section 4.1 in [1] exists and

$$\left(1 - 91(\alpha+2)\sqrt{\frac{\ln(n)}{n}}\right)\sigma^2 \leq \widehat{C} \leq \left(1 + \frac{44(\alpha+2)\sqrt{\ln(n)}}{n^{1/4}}\right)\sigma^2 \ .$$

Remark 1 *If* $(\mathbf{A_1})$ *is replaced by*

$$\exists \lambda_1 \in \Lambda, \quad \mathrm{df}(\lambda_1) \geq a_n \quad \textit{and} \quad b(\lambda_1) \leq \sigma^2 \beta_n \tag{$\mathbf{A_1'}$}$$

for some $a_n \geq \ln(n)$ *and* $\beta_n \geq 0$*, then Proposition 1 still holds with Eq.* (10) *replaced by*

$$\forall 0 \leq C < \left(1 - \frac{3\beta_n}{a_n} - 62\gamma\sqrt{\frac{\ln(n)}{a_n}}\right)\sigma^2, \quad \mathrm{df}(\widehat{\lambda}_0(C)) \geq \frac{a_n}{5} \ . \tag{12}$$

Remark 2 *If* $(\mathbf{A_2})$ *is replaced by*

$$\exists \lambda_2 \in \Lambda, \quad \mathrm{df}(\lambda_2) \leq n^a \quad \textit{and} \quad b(\lambda_2) \leq \sigma^2 \beta_n' \tag{$\mathbf{A_2'}$}$$

for some $a \in [1/2, 1)$ *and* $\beta_n' \geq \sqrt{n\ln(n)}$*, then for every* $\beta \in (a, 1)$ *Proposition 1 still holds for* $n \geq \max\left\{C_1, 4^{1/(\beta-a)}\right\}$ *with Eq.* (11) *replaced by*

$$\forall C > \left(1 + 44\gamma\beta_n' n^{-\beta}\right)\sigma^2, \quad \mathrm{df}(\widehat{\lambda}_0(C)) \leq n^\beta \ . \tag{13}$$

Remark 3 *On the event defined in Proposition 1, we can derive from Eq.* (3)*,* (39)*,* (48)*, and* $\|A_\lambda\| \leq 1$*, that*

$$\forall \lambda \in \Lambda \textit{ such that } \mathrm{df}(\lambda) \geq \frac{n}{\ln(n)}, \quad n^{-1}\left\|\widehat{F}_\lambda - F\right\|_2^2 \geq \left(\frac{1}{2\ln(n)} - \frac{8\gamma\ln(n)}{n}\right)\sigma^2 \ .$$

Hence, the blow up of $\mathrm{df}(\widehat{\lambda}_0(C))$ *holding when the penalty is below the minimal penalty also implies a blow up of the risk* $n^{-1}\left\|\widehat{F}_{\widehat{\lambda}_0(C)} - F\right\|_2^2$.

Let us now prove Proposition 1.

2.1 General starting point

Combining Eq. (9) with Eq. (3) and (4), for every $C \geq 0$, $\widehat{\lambda}_0(C)$ also minimizes over $\lambda \in \Lambda$

$$\begin{aligned}\mathrm{crit}_C(\lambda) &:= \left\|\widehat{F}_\lambda - Y\right\|_2^2 - \|\varepsilon\|_2^2 + C\left(2\,\mathrm{tr}(A_\lambda) - \mathrm{tr}(A_\lambda^\top A_\lambda)\right) \\ &= b(\lambda) + (\sigma^{-2}C - 1)\left(2v_1(\lambda) - v_2(\lambda)\right) - 2\delta_1(\lambda) + \delta_2(\lambda) + \delta_3(\lambda) + \delta_4(\lambda) \ .\end{aligned}$$

We now use the concentration inequalities of Eq. (39), (40), (47) and (48) proved in Section 4: For every $\lambda \in \Lambda$ and $x \geq 1$, an event of probability $1 - 8e^{-x}$ exists on which for every $C \geq 0$ and $\theta > 0$,

$$\mathrm{crit}_C(\lambda) \geq \frac{b(\lambda)}{3} + (\sigma^{-2}C - 1)(2v_1(\lambda) - v_2(\lambda)) - 3\theta v_1(\lambda) - 6(2 + \theta^{-1})x\sigma^2 \tag{14}$$

$$\mathrm{crit}_C(\lambda) \leq \frac{5b(\lambda)}{3} + (\sigma^{-2}C - 1)(2v_1(\lambda) - v_2(\lambda)) + 3\theta v_1(\lambda) + 6(2 + \theta^{-1})x\sigma^2 \ , \tag{15}$$

using also that $v_2 \leq v_1$ by Eq. (8) and that $\|A_\lambda\| \leq 1$.

For every $x \geq 1$, let Ω_x be the event on which the inequalities appearing in Eq. (39), (40), (47) and (48) hold for every $\theta > 0$ and $\lambda \in \Lambda$. The union bound shows that $\mathbb{P}(\Omega_x) \geq 1 - 8\,\mathrm{Card}(\Lambda)e^{-x}$.

2.2 Below the minimal penalty

We assume in this subsection that $C \in [0, \sigma^2)$. We will prove Eq. (12) using assumption $(\mathbf{A_1'})$, since when $a_n = n/2$ and $\beta_n = \sqrt{n\ln(n)}$, Eq. (12) is Eq. (10) and $(\mathbf{A_1'})$ is $(\mathbf{A_1})$.

Using Eq. (8) and taking $\theta = \sqrt{x/\,\mathrm{df}(\lambda)}$ in Eq. (14) and (15), we get that for every $x \geq 1$, on Ω_x, for every $\lambda \in \Lambda$,

$$\mathrm{crit}_C(\lambda) \geq \frac{b(\lambda)}{3} + 2(C - \sigma^2)\,\mathrm{df}(\lambda) - \left(9\sqrt{x\,\mathrm{df}(\lambda)} + 12x\right)\sigma^2 \tag{16}$$

$$\mathrm{crit}_C(\lambda) \leq \frac{5b(\lambda)}{3} + (C - \sigma^2)\,\mathrm{df}(\lambda) + \left(9\sqrt{x\,\mathrm{df}(\lambda)} + 12x\right)\sigma^2 \ . \tag{17}$$

Let $\lambda \in \Lambda$. Two cases can be distinguished:

1. If $\mathrm{df}(\lambda) < a_n/5$, then Eq. (16) implies

$$\mathrm{crit}_C(\lambda) \geq \frac{2(C-\sigma^2)a_n}{5} - \left(9\sqrt{\frac{xa_n}{5}} + 12x\right)\sigma^2 \ . \tag{18}$$

2. If $\mathrm{df}(\lambda) \geq a_n$, then Eq. (17) implies

$$\mathrm{crit}_C(\lambda) \leq \frac{5b(\lambda)}{3} + (C-\sigma^2)a_n + (9\sqrt{xa_n} + 12x)\,\sigma^2 \ . \tag{19}$$

We now take $x = \gamma \ln(n)$ so that $\mathbb{P}(\Omega_x) \geq 1 - 8\,\mathrm{Card}(\Lambda) n^{-\gamma}$.

On the one hand, Eq. (18) implies

$$\inf_{\lambda\in\Lambda,\,\mathrm{df}(\lambda)<a_n/5}\{\mathrm{crit}_C(\lambda)\} \geq \frac{2(C-\sigma^2)a_n}{5} - \left(9\sqrt{\frac{\gamma a_n \ln(n)}{5}} + 12\gamma\ln(n)\right)\sigma^2 \ . \tag{20}$$

On the other hand, for $\lambda = \lambda_1$ given by assumption $(\mathbf{A_1'})$, Eq. (19) implies

$$\mathrm{crit}_C(\lambda_1) \leq \frac{5\sigma^2\beta_n}{3} + (C-\sigma^2)a_n + \left(9\sqrt{\gamma a_n \ln(n)} + 12\gamma\ln(n)\right)\sigma^2 \ . \tag{21}$$

Comparing Eq. (20) and Eq. (21), we get that

$$\mathrm{crit}_C(\lambda_1) < \inf_{\lambda\in\Lambda,\,\mathrm{df}(\lambda)<a_n/5}\{\mathrm{crit}_C(\lambda)\}$$

hence $\mathrm{df}(\widehat{\lambda}_0(C)) \geq a_n/5$ if

$$1 - \sigma^{-2}C > \frac{3\beta_n}{a_n} + 62\gamma\sqrt{\frac{\ln(n)}{a_n}} \ .$$

2.3 Above the minimal penalty

We assume in this subsection that $C > \sigma^2$. We will prove Eq. (13) using assumption $(\mathbf{A_2'})$, since when $a = 1/2$, $\beta_n' = \sqrt{n\ln(n)}$ and $\beta = (1+a)/2 = 3/4$, Eq. (13) is Eq. (11) and $(\mathbf{A_2'})$ is $(\mathbf{A_2})$.

Using Eq. (8) and taking $\theta = \sqrt{x/\,\mathrm{df}(\lambda)}$ in Eq. (14) and (15), we get that for every $x \geq 1$, on Ω_x, for every $\lambda \in \Lambda$,

$$\mathrm{crit}_C(\lambda) \geq \frac{b(\lambda)}{3} + (C-\sigma^2)\,\mathrm{df}(\lambda) - \left(9\sqrt{x\,\mathrm{df}(\lambda)} + 12x\right)\sigma^2 \tag{22}$$

$$\mathrm{crit}_C(\lambda) \leq \frac{5b(\lambda)}{3} + 2(C-\sigma^2)\,\mathrm{df}(\lambda) + \left(9\sqrt{x\,\mathrm{df}(\lambda)} + 12x\right)\sigma^2 \ . \tag{23}$$

Let $\lambda \in \Lambda$, and $\beta \in (a,1)$. As in Section 2.2, we consider two cases.

1. If $\mathrm{df}(\lambda) \leq n^a$, Eq. (23) implies

$$\mathrm{crit}_C(\lambda) \leq 2b(\lambda) + 2(C-\sigma^2)n^a + \left(9\sqrt{xn^a} + 12x\right)\sigma^2 \ . \tag{24}$$

2. If $\mathrm{df}(\lambda) \geq n^\beta$, Eq. (22) implies

$$\mathrm{crit}_C(\lambda) \geq (C-\sigma^2)n^\beta - \left(9\sqrt{xn^\beta} + 12x\right)\sigma^2 \ . \tag{25}$$

We now take $x = \gamma\ln(n)$ as in Section 2.2.

On the one hand, for $\lambda = \lambda_2$ given by assumption $(\mathbf{A_2'})$, Eq. (24) implies

$$\mathrm{crit}_C(\lambda_2) \leq 2\sigma^2\beta_n' + (C-\sigma^2)\frac{n^\beta}{2} + \left(9\sqrt{\gamma\ln(n)n^a} + 12\gamma\ln(n)\right)\sigma^2 \tag{26}$$

if $n^{\beta-a} \geq 4$.

On the other hand, Eq. (25) implies

$$\inf_{\lambda\in\Lambda,\,\mathrm{df}(\lambda)\geq n^\beta} \{\,\mathrm{crit}_C(\lambda)\,\} \geq (C-\sigma^2)n^\beta - \left(9\sqrt{\gamma\ln(n)n^\beta} + 12\gamma\ln(n)\right)\sigma^2 \quad . \tag{27}$$

Comparing Eq. (26) and Eq. (27), we get that

$$\mathrm{crit}_C(\lambda_2) < \inf_{\lambda\in\Lambda,\,\mathrm{df}(\lambda)\geq n^\beta} \{\,\mathrm{crit}_C(\lambda)\,\}$$

hence $\mathrm{df}(\widehat{\lambda}_0(C)) < n^\beta$ if

$$n \geq 4^{1/(\beta-a)}\,, \quad \sqrt{n/\ln(n)} \geq 12\,, \quad \text{and} \quad \sigma^{-2}C - 1 > 44\gamma\beta'_n n^{-\beta} \quad .$$

■

3 Oracle inequality

Define

$$\forall C \geq 0\,, \qquad \widehat{\lambda}_{\mathrm{opt}}(C) \in \arg\min_{\lambda\in\Lambda}\left\{ \left\|\widehat{F}_\lambda - Y\right\|_2^2 + 2C\,\mathrm{tr}(A_\lambda)\right\} \quad . \tag{28}$$

We will prove the following proposition in this section.

Proposition 2 *Let $\widehat{\lambda}_{\mathrm{opt}}$ be defined by Eq.* (28)*. Assume that $\forall\lambda\in\Lambda$, A_λ is symmetric with $\mathrm{Sp}(A_\lambda)\subset[0,1]$, that ε_i are i.i.d. Gaussian with zero mean and variance $\sigma^2>0$.*

Then, a numerical constant $C_2>0$ exists such that for every $n\geq C_2$, $\gamma\geq 1$, $\eta^+\geq(\ln(n))^{-1}$, and $C>0$ such that $\sigma^{-2}C\in\left[1+(\ln(n))^{-1}, 1+\eta^+\right]$,

$$n^{-1}\left\|\widehat{F}_{\widehat{\lambda}_{\mathrm{opt}}(C)} - F\right\|_2^2 \leq \left(1+\frac{3}{\ln(n)}\right)\inf_{\lambda\in\Lambda}\left\{ n^{-1}\left\|\widehat{F}_\lambda - F\right\|_2^2 + 4\eta^+\frac{\sigma^2\,\mathrm{tr}(A_\lambda)}{n}\right\} + \frac{14\gamma\,(\ln(n))^2\,\sigma^2}{n} \quad . \tag{29}$$

holds with probability at least $1-8\,\mathrm{Card}(\Lambda)n^{-\gamma}$.

If in addition

$$\exists\kappa\geq 1\,,\,\forall\lambda\in\Lambda\,, \quad v_1(\lambda) \leq \kappa\left(v_2(\lambda)+b(\lambda)+(\ln(n))^2\,\sigma^2\right) \quad , \tag{$\mathbf{A'_3}$}$$

then a constant $C_3>0$ depending only on κ exists such that for every $n\geq C_3$, $\gamma\geq 1$, and $C>0$ such that $\sigma^{-2}C\in\left[1-(\ln(n))^{-1}, 1+(\ln(n))^{-1}\right]$,

$$n^{-1}\left\|\widehat{F}_{\widehat{\lambda}_{\mathrm{opt}}(C)} - F\right\|_2^2 \leq \left(1+\frac{40\kappa}{\ln(n)}\right)\inf_{\lambda\in\Lambda}\left\{ n^{-1}\left\|\widehat{F}_\lambda - F\right\|_2^2\right\} + \frac{36(\kappa+\gamma)\ln(n)\sigma^2}{n} \tag{30}$$

holds with probability at least $1-8\,\mathrm{Card}(\Lambda)n^{-\gamma}$.

If $\mathrm{Card}(\Lambda)\leq Kn^\alpha$, Proposition 2 with $\gamma=\alpha+2$ proves that with probability at least $1-8Kn^{-2}$, $\widehat{\lambda}$ defined in the Algorithm of Section 4.1 in [1] satisfies an oracle inequality if assumption ($\mathbf{A'_3}$) holds.

Remark 4 *Assumption ($\mathbf{A'_3}$) holds as soon as ($\mathbf{A_3}$) holds, i.e.,*

$$\mathbb{E}\left[n^{-1}\left\|\widehat{F}_\lambda - F\right\|_2^2\right] = n^{-1}\left(v_2(\lambda)+b(\lambda)\right) \geq \kappa^{-1}\frac{\sigma^2\,\mathrm{tr}(A_\lambda)}{n} \quad ,$$

which is the parametric rate of estimation in a model of dimension $\mathrm{tr}(A_\lambda)$.

In the ordinary least-squares regression example, where all A_λ are projection matrices, assumption ($\mathbf{A_3}$) always holds with $\kappa=1$ because $v_1(\lambda)=v_2(\lambda)$.

In the kernel ridge regression example, a sufficient condition for $(\mathbf{A_3})$ *is that the eigenvalues* $(\mu_j)_{1\leq j\leq n}$ *of the kernel matrix* K *satisfy*

$$\exists \alpha, L_1, L_2 > 0, \forall 1 \leq j \leq n, \quad L_1 j^{-\alpha} \leq \mu_j \leq L_2 j^{-\alpha} \ ,$$

which is a classical assumption in kernel ridge regression with a random design; see Section 5.2 for details.

Remark 5 *When* $\operatorname{tr}(A_\lambda^\top A_\lambda) \ll \operatorname{tr}(A_\lambda)$ *for most* $\lambda \in \Lambda$*, taking* C *slightly larger than* σ^2 *as in the first part of Proposition 2 is necessary to obtain an oracle inequality. Indeed, Proposition 1 then shows that*

$$\left(2\operatorname{tr}(A_\lambda) - \operatorname{tr}(A_\lambda^\top A_\lambda)\right)\sigma^2 n^{-1} \approx 2\operatorname{tr}(A_\lambda)\sigma^2 n^{-1}$$

is a minimal penalty. So, any underestimation of the constant C *in the penalty* $2C\operatorname{tr}(A_\lambda)n^{-1}$ *may lead to selecting* $\widehat{\lambda} = \widehat{\lambda}_{\mathrm{opt}}(C)$ *with* $\mathrm{df}(\widehat{\lambda}) \geq n/(\ln(n))$*.*

Such a phenomenon holds for instance when $A_\lambda = \lambda I_n$ *and* $\Lambda \subset [0,1]$*, since* $\operatorname{tr}(A_\lambda^\top A_\lambda) = \operatorname{tr}(A_\lambda)^2 n^{-1} \ll \operatorname{tr}(A_\lambda)$ *unless* $\operatorname{tr}(A_\lambda) \propto n$*.*

Remark 6 *The remainder terms in Eq.* (29) *and* (30)*,* $14\gamma(\ln(n))^2\sigma^2 n^{-1}$ *and* $36(\kappa + \gamma)\ln(n)\sigma^2 n^{-1}$*, are negligible in front of the risk of the oracle provided that* $v_2(\lambda^\star)$ *tends grows with* n *faster than* $(\ln(n))^2$*, since the risk of* $\widehat{F}_{\lambda^\star}$ *is at least of order* $v_2(\lambda^\star)n^{-1}$*. This usually holds when the bias is not exactly zero for some* $\lambda \in \Lambda$ *with* $\operatorname{tr}(A_\lambda^\top A_\lambda)$ *too small, as often assumed in the model selection literature for proving asymptotic optimality results.*

Let us now prove Proposition 2.

3.1 General starting point

Combining Eq. (4) and (28), we obtain that for every $C > 0$ such that $\sigma^{-2}C \in [1-\eta^-, 1+\eta^+]$ and every $\lambda \in \Lambda$,

$$\begin{aligned} &\left\|\widehat{F}_{\widehat{\lambda}_{\mathrm{opt}}(C)} - F\right\|_2^2 - 2\eta^- v_1(\widehat{\lambda}_{\mathrm{opt}}(C)) + \widehat{\Delta}(\widehat{\lambda}_{\mathrm{opt}}(C)) \\ &\qquad \leq \inf_{\lambda\in\Lambda}\left\{\left\|\widehat{F}_\lambda - F\right\|_2^2 + 2\eta^+ v_1(\lambda) + \widehat{\Delta}(\lambda)\right\} \ . \end{aligned} \tag{31}$$

where

$$\widehat{\Delta}(\lambda) := -2\delta_1(\lambda) + \delta_4(\lambda) \ .$$

Inequality (31) implies an oracle inequality as soon as $\widehat{\Delta}(\lambda)$ is small compared to $\left\|\widehat{F}_\lambda - F\right\|_2^2$ and η^-, η^+ are small enough.

3.2 Make use of concentration inequalities

Let Ω_x denote the same event as in Section 2. From Eq. (40) and (47), since $\|A_\lambda\| \leq 1$, we deduce that on Ω_x

$$\forall \lambda \in \Lambda, \forall \theta > 0, \quad \left|\widehat{\Delta}(\lambda)\right| \leq \theta b(\lambda) + 2\theta v_1(\lambda) + (4 + 5\theta^{-1})x\sigma^2 \ . \tag{32}$$

In addition, combining Eq. (3), (39) and (48) with $\theta = 1/2$, and $\|A_\lambda\| \leq 1$, we have on Ω_x,

$$\forall \lambda \in \Lambda, \quad b(\lambda) + v_2(\lambda) \leq 2\left\|\widehat{F}_\lambda - F\right\|_2^2 + 16x\sigma^2 \ . \tag{33}$$

3.3 First result: with a slightly enlarged penalty

Assume in this subsection that $\sigma^{-2}C \in \left[1 + (\ln(n))^{-1}; 1+\eta^+\right]$ with $\eta^+ \geq (\ln(n))^{-1}$. Then, Eq. (31) and (32) with $\theta = (\ln(n))^{-1}$ imply

$$\left\|\widehat{F}_{\widehat{\lambda}_{\mathrm{opt}}(C)} - F\right\|_2^2 \leq \frac{1 + (\ln(n))^{-1}}{1 - (\ln(n))^{-1}} \inf_{\lambda\in\Lambda}\left\{\left\|\widehat{F}_\lambda - F\right\|_2^2 + 4\eta^+ v_1(\lambda)\right\} + (9 + 12\ln(n))x\sigma^2 \ , \tag{34}$$

if $\ln(n) \geq 5$.

Taking $x = \gamma \ln(n)$ with $\gamma \geq 1$, then $\mathbb{P}(\Omega_x) \geq 1 - 8\,\mathrm{Card}(\Lambda) n^{-\gamma}$ and Eq. (34) implies Eq. (29) for every $n \geq C_2 = e^5$.

3.4 Second result: with assumption $(\mathbf{A}'_\mathbf{3})$

Assume in this subsection that $\sigma^{-2} C \in [1-\eta^-; 1+\eta^+]$ with $0 \leq \eta^-, \eta^+ \leq (\ln(n))^{-1}$, and that $(\mathbf{A}'_\mathbf{3})$ holds.

Then, Eq. (32) with $\theta = (\ln(n))^{-1}$ and Eq. (33) imply

$$\begin{aligned} &\left\| \widehat{F}_{\widehat{\lambda}_{\mathrm{opt}}(C)} - F \right\|_2^2 - 2\eta^- v_1(\widehat{\lambda}_{\mathrm{opt}}(C)) + \widehat{\Delta}(\widehat{\lambda}_{\mathrm{opt}}(C)) \\ &\geq \left(1 - \frac{10\kappa}{\ln(n)}\right) \left\| \widehat{F}_{\widehat{\lambda}_{\mathrm{opt}}(C)} - F \right\|_2^2 - \left[\left(4 + \frac{80\kappa}{\ln(n)}\right) x + 9\ln(n)\kappa \right] \sigma^2 \ , \end{aligned} \tag{35}$$

and for every $\lambda \in \Lambda$,

$$\begin{aligned} &\left\| \widehat{F}_{\lambda} - F \right\|_2^2 + 2\eta^+ v_1(\lambda) + \widehat{\Delta}(\lambda) \\ &\leq \left(1 + \frac{10\kappa}{\ln(n)}\right) \left\| \widehat{F}_{\lambda} - F \right\|_2^2 + \left[\left(4 + \frac{80\kappa}{\ln(n)}\right) x + 9\ln(n)\kappa \right] \sigma^2 \ . \end{aligned} \tag{36}$$

Combining Eq. (31), (35) and (36), we get that on Ω_x,

$$\left\| \widehat{F}_{\widehat{\lambda}_{\mathrm{opt}}(C)} - F \right\|_2^2 \leq \left(1 + \frac{40\kappa}{\ln(n)}\right) \left\| \widehat{F}_{\lambda} - F \right\|_2^2 + 4 \left[\left(4 + \frac{80\kappa}{\ln(n)}\right) x + 9\ln(n)\kappa \right] \sigma^2 \tag{37}$$

if $\ln(n) \geq 20\kappa$.

Now, taking $x = \gamma \ln(n)$ with $\gamma \geq 1$ in Eq. (37) implies Eq. (30) for every $n \geq C_3(\kappa)$. ■

4 Concentration inequalities

The concentration inequalities used for proving Propositions 1 and 2 are proved in this section.

4.1 Linear functions of ε

We here prove concentration inequalities for $\delta_3(\lambda)$ and $\delta_4(\lambda)$. Let us first prove a classical result.

Proposition 3 *Let ξ be a standard Gaussian vector in $\mathbb{R}^n$, $\alpha \in \mathbb{R}^n$ and*

$$Z = \langle \xi,\, \alpha \rangle = \sum_{j=1}^{n} \alpha_j \xi_j \ .$$

Then, for every $x \geq 0$,

$$\mathbb{P}\left(|Z| \leq \sqrt{2x}\, \|\alpha\|_2 \right) \geq 1 - 2e^{-x} \ . \tag{38}$$

Proof Z is a Lipschitz function of ξ, with Lipschitz constant $\|\alpha\|_2$. Therefore, the Gaussian concentration theorem implies (see for instance Theorem 3.4 in [2]):

$$\forall t \geq 0, \qquad \mathbb{P}\left(|Z| \geq t \right) \leq 2\exp\left(-\frac{t^2}{2\,\|\alpha\|_2^2} \right) \ .$$

The result follows by taking $t = \sqrt{2x}\,\|\alpha\|_2$. ■

Now, remark that

$$\delta_3(\lambda) = \left\langle \sigma^{-1}\varepsilon,\, 2\sigma A_\lambda^\top (I_n - A_\lambda) F \right\rangle \quad \text{and} \quad \delta_4(\lambda) = \left\langle \sigma^{-1}\varepsilon,\, 2\sigma (I_n - A_\lambda) F \right\rangle \ ,$$

where $\sigma^{-1}\varepsilon$ is a standard Gaussian vector. Hence, Proposition 3 shows that for every $x \geq 0$ and $\lambda \in \Lambda$,

$$\mathbb{P}\left(|\delta_3(\lambda)| \leq 2\sigma\sqrt{x}\left\|A_\lambda^\top(I_n - A_\lambda)F\right\|_2\right) \geq 1 - 2e^{-x}$$
$$\mathbb{P}\left(|\delta_4(\lambda)| \leq 2\sigma\sqrt{x}\left\|(I_n - A_\lambda)F\right\|_2\right) \geq 1 - 2e^{-x} \ ,$$

which implies that

$$\mathbb{P}\left(\forall\theta > 0,\ |\delta_3(\lambda)| \leq \theta^{-1}x\left\|A_\lambda\right\|^2\sigma^2 + \theta\left\|(I_n - A_\lambda)F\right\|_2^2\right) \geq 1 - 2e^{-x} \tag{39}$$

$$\mathbb{P}\left(\forall\theta > 0,\ |\delta_4(\lambda)| \leq \theta^{-1}x\sigma^2 + \theta\left\|(I_n - A_\lambda)F\right\|_2^2\right) \geq 1 - 2e^{-x} \ , \tag{40}$$

since $\forall a, b, \theta > 0$, $2ab \leq \theta a^2 + \theta^{-1}b^2$.

4.2 Quadratic functions of ε

We here prove concentration inequalities for $\delta_2(\lambda)$ and $\delta_1(\lambda)$. Let us first prove (recall) a general result.

Proposition 4 *Let ξ be a standard Gaussian vector in $\mathbb{R}^n$, M a real-valued $n \times n$ matrix and*

$$Z = \left\|M\xi\right\|_2^2 - \mathrm{tr}(M^\top M) \ .$$

Then, for every $x \geq 0$,

$$\mathbb{P}\left(\forall\theta > 0,\ Z \leq \theta\,\mathrm{tr}(M^\top M) + 2(1+\theta^{-1})\left\|M\right\|^2 x\right) \geq 1 - e^{-x} \tag{41}$$

$$\mathbb{P}\left(\forall\theta > 0,\ Z \geq -\theta\,\mathrm{tr}(M^\top M) - \left[2x\left(\theta^{-1} - 1\right) + 1 - \theta\right]\left\|M\right\|^2\right) \geq 1 - e^{-x} \ . \tag{42}$$

Proof Define $W = \left\|M\xi\right\|_2$, and note that $\mathbb{E}\left[W^2\right] = \mathrm{tr}(M^\top M)$. Since W is a Lipschitz function of ξ with Lipschitz constant $\left\|M\right\|$, the Gaussian concentration theorem (see for instance Theorem 3.4 in [2]) shows that for every $x \geq 0$, an event Ω_x^+ of probability at least $1 - \exp(-x)$ exists on which

$$W \leq \mathbb{E}\left[W\right] + \sqrt{2x}\left\|M\right\| \ , \tag{43}$$

and an event Ω_x^- of probability at least $1 - \exp(-x)$ exists on which

$$W \geq \mathbb{E}\left[W\right] - \sqrt{2x}\left\|M\right\| \ . \tag{44}$$

In addition, Proposition 3.5 in [2] shows that $\mathrm{var}(W) \leq \left\|M\right\|^2$. Therefore,

$$0 \leq \mathbb{E}\left[W^2\right] - \left(\mathbb{E}\left[W\right]\right)^2 = \mathrm{var}(W) \leq \left\|M\right\|^2 \ . \tag{45}$$

We now combine Eq. (45) with the two concentration inequalities above for proving the result.

On the one hand, on Ω_x^+,

$$\begin{aligned} W^2 &\leq \left(\mathbb{E}\left[W\right]\right)^2 + 2\mathbb{E}\left[W\right]\sqrt{2x}\left\|M\right\| + 2x\left\|M\right\|^2 \\ &\leq \mathbb{E}\left[W^2\right] + 2\sqrt{2x\mathbb{E}\left[W^2\right]}\left\|M\right\| + 2x\left\|M\right\|^2 \\ &\leq (1+\theta)\mathbb{E}\left[W^2\right] + 2(1+\theta^{-1})x\left\|M\right\|^2 \end{aligned}$$

for every $\theta > 0$, using successively Eq. (45) and that $\forall a, b, \theta > 0$, $2\sqrt{ab} \leq a\theta + b\theta^{-1}$. This proves Eq. (41).

On the other hand, for every $x \geq 0$ such that $x \leq (\mathbb{E}\left[W^2\right] - \left\|M\right\|^2)/(2\left\|M\right\|^2)$, on Ω_x^-

$$\begin{aligned} W^2 &\geq \left(\sqrt{\mathbb{E}\left[W^2\right] - \left\|M\right\|^2} - \sqrt{2x}\left\|M\right\|\right)^2 \\ &\geq (1-\theta)\mathbb{E}\left[W^2\right] - \left[2x\left(\theta^{-1} - 1\right) + 1 - \theta\right]\left\|M\right\|^2 \ . \end{aligned} \tag{46}$$

This proves Eq. (42), since the lower bound in Eq. (46) is non-positive if $x > (\mathbb{E}\left[W^2\right] - \left\|M\right\|^2)/(2\left\|M\right\|^2)$. ■

Now, remark that if B_λ exists such that $A_\lambda = B_\lambda^\top B_\lambda$ —as in the kernel ridge regression example for instance, and more generally when A_λ is symmetric real-valued with $\mathrm{Sp}(A_\lambda) \subset [0,1]$ —, then

$$\sigma^{-2}\delta_1(\lambda) = \left\|B_\lambda(\sigma^{-1}\varepsilon)\right\|_2^2 - \mathrm{tr}(B_\lambda^\top B_\lambda) \quad \text{and} \quad \sigma^{-2}\delta_2(\lambda) = \left\|A_\lambda(\sigma^{-1}\varepsilon)\right\|_2^2 - \mathrm{tr}(A_\lambda^\top A_\lambda) \ .$$

Hence, Proposition 4 shows that for every $x \geq 0$ and $\lambda \in \Lambda$,

$$\mathbb{P}\left(\forall\theta > 0,\ |\delta_1(\lambda)| \leq \theta\sigma^2\,\mathrm{tr}(A_\lambda) + 2(1+\theta^{-1})x\,\|A_\lambda\|\,\sigma^2\right) \geq 1 - 2e^{-x} \tag{47}$$

$$\mathbb{P}\left(\forall\theta > 0,\ |\delta_2(\lambda)| \leq \theta\sigma^2\,\mathrm{tr}(A_\lambda^\top A_\lambda) + 2(1+\theta^{-1})x\,\|A_\lambda\|^2\,\sigma^2\right) \geq 1 - 2e^{-x} \ , \tag{48}$$

where we used in particular that $\|B_\lambda\|^2 = \|A_\lambda\|$.

5 Kernel ridge regression example

Finally, let us make some computations that are specific to the kernel ridge regression example.

5.1 Explicit formulas for the deterministic terms

Let K be the $n \times n$ matrix such that $K_{i,j} = k(x_i, x_j)$. Then, the kernel regression estimator with regularization parameter λ is defined by

$$\widehat{F}_\lambda = A_\lambda Y \qquad \text{with} \quad A_\lambda = K(K + n\lambda I_n)^{-1} \ .$$

Then, A_λ is symmetric, real-valued (hence diagonalizable by orthogonal matrices) and $\mathrm{Sp}(A_\lambda) \subset [0,1]$.

Let $(e_j)_{1\leq j\leq n}$ be the (orthonormal) eigenvectors of K, with eigenvalues $(\mu_j)_{1\leq j\leq n}$, assuming that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n \geq 0$. We also assume that $\mu_1 > 0$, that is, K is not the null matrix. We can then decompose F in this basis: $F = \sum_j f_j e_j$.

Therefore, in the orthonormal basis $(e_j)_{1\leq j\leq n}$, A_λ is diagonal with coefficients

$$\left(\frac{\mu_j}{\mu_j + n\lambda}\right)_{1\leq j\leq n} \ .$$

Hence,

$$\begin{aligned}
\mathrm{tr}(A_\lambda) = \mathrm{df}(\lambda) &= \sum_{j=1}^n \left(\frac{\mu_j}{\mu_j + n\lambda}\right) \\
\mathrm{tr}(A_\lambda^\top A_\lambda) &= \sum_{j=1}^n \left(\frac{\mu_j}{\mu_j + n\lambda}\right)^2 \\
2\,\mathrm{tr}(A_\lambda) - \mathrm{tr}(A_\lambda^\top A_\lambda) &= \sum_{j=1}^n \left[\frac{2\mu_j}{\mu_j + n\lambda} - \left(\frac{\mu_j}{\mu_j + n\lambda}\right)^2\right] = \sum_{j=1}^n \left[\frac{\mu_j(\mu_j + 2n\lambda)}{(\mu_j + n\lambda)^2}\right] \\
b(\lambda) = \|(A_\lambda - I_n)F\|_2^2 &= \sum_{j=1}^n \left(1 - \frac{\mu_j}{\mu_j + n\lambda}\right)^2 f_j^2 \ .
\end{aligned}$$

Note that $\mathrm{df}(\lambda)$ and $\mathrm{tr}(A_\lambda^\top A_\lambda)$ are decreasing functions of λ, as well as $2\,\mathrm{tr}(A_\lambda) - \mathrm{tr}(A_\lambda^\top A_\lambda)$ since each term of the sum is nonincreasing, and at least one is decreasing. On the contrary, $b(\lambda)$ is an nondecreasing function of λ. Hence, $\mathrm{tr}(A_\lambda^\top A_\lambda)$ and $2\,\mathrm{tr}(A_\lambda) - \mathrm{tr}(A_\lambda^\top A_\lambda)$ are increasing functions of $\mathrm{df}(\lambda)$, and $b(\lambda)$ is a nonincreasing function of $\mathrm{df}(\lambda)$.

5.2 Sufficient condition for assumption $(\mathbf{A_3})$

Assumption $(\mathbf{A_3})$ holds in particular when

$$\exists\kappa \geq 1,\ \forall\lambda \in \Lambda, \quad \mathrm{tr}(A_\lambda) \leq \kappa\,\mathrm{tr}(A_\lambda^\top A_\lambda) \ .$$

If the eigenvalues of K satisfy

$$\exists \alpha, L_1, L_2 > 0, \forall 1 \leq j \leq n, \quad L_1 j^{-\alpha} \leq \mu_j \leq L_2 j^{-\alpha} \ ,$$

then, following [3],

$$\begin{aligned}
\operatorname{tr}(A_\lambda) &\leq \sum_{j=1}^{n} \frac{L_2 j^{-\alpha}}{L_2 j^{-\alpha} + n\lambda} = \sum_{j=1}^{n} \frac{1}{1 + n\lambda L_2^{-1} j^{\alpha}} \\
&\leq \int_0^\infty \frac{dt}{1 + n\lambda L_2^{-1} t^{\alpha}} = \left(\frac{L_2}{n\lambda}\right)^{1/\alpha} \int_0^\infty \frac{du}{1+u^\alpha}
\end{aligned}$$

and

$$\begin{aligned}
\operatorname{tr}(A_\lambda^\top A_\lambda) &\geq \sum_{j=1}^{n} \left(\frac{L_1 j^{-\alpha}}{L_1 j^{-\alpha} + n\lambda}\right)^2 = \sum_{j=1}^{n} \frac{1}{\left(1 + n\lambda L_1^{-1} j^{\alpha}\right)^2} \\
&\geq \int_1^\infty \frac{dt}{\left(1 + n\lambda L_1^{-1} t^{\alpha}\right)^2} = \left(\frac{L_1}{n\lambda}\right)^{1/\alpha} \int_1^\infty \frac{du}{(1+u^\alpha)^2} \ .
\end{aligned}$$

Therefore, $(\mathbf{A_3})$ holds with

$$\kappa = \left(\frac{L_2}{L_1}\right)^{1/\alpha} \int_0^\infty \frac{du}{1+u^\alpha} \left(\int_1^\infty \frac{du}{(1+u^\alpha)^2}\right)^{-1} \ .$$

References

[1] Sylvain Arlot and Francis Bach. Data-driven calibration of linear estimators with minimal penalties. In *Advances in Neural Information Processing Systems (NIPS)*, December 2009.

[2] P. 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.

[3] Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel fisher discriminant analysis, April 2008. oai:hal.archives-ouvertes.fr:hal-00270806_v1.