PAC-Bayesian aggregation of affine estimators

Lucie Montuelle, Erwan Le Pennec

PAC-Bayesian aggregation of affine estimators

Lucie Montuelle – LPMA, Université Paris Diderot
Erwan Le Pennec – CMAP, École Polytechnique.

Abstract

Aggregating estimators using exponential weights depending on their risk performs well in expectation, but sadly not in probability. Considering exponential weights of a penalized risk is a way to overcome this issue. We focus on the fixed design regression framework with sub-Gaussian noise and provide penalties allowing to obtain oracle inequalities in deviation for the aggregation of affine estimators. Sharp oracle inequalities are provided by a condition using the regression function's norm.

MSC 2010 subject classifications: Primary 62G08; secondary 62J02.
Keywords: Exponentially weighted aggregation, Regression, Oracle inequality, Deviation

Contents

1 Introduction

We consider here a classical fixed design regression model

$$\forall i \in \{1,\ldots,n\}, Y_i = f_0(x_i) + W_i$$

with f_0 an unknown function, x_i the fixed design points and $W = (W_i)_{i\leq n}$ a centered sub-Gaussian noise. Our aim is to estimate the function f_0 at the grid points.

Many regression estimators are available in the literature. For non parametric estimation, Nadaraya-Watson estimator [37, 50] and its fixed design counterpart [25] are widely used, just like projection estimators using trigonometric, wavelet [23] or spline [49] basis for example. In the parametric framework, least squares or maximum likelihood estimators are commonly employed, sometimes with minimization constraints, leading to LASSO [45], ridge [31], elastic net [58], AIC [1] or BIC [43] estimates.

Facing this variety, the statistician may wonder which procedure provides the best estimation. Unfortunately, the answer depends on the data. For instance, a rectangular function is well approximated by wavelets but not by trigonometric functions. Since the best estimator is not known in advance, our aim is to mimic its performances in term of risk. This is theoretically guaranteed by an oracle inequality:

$$R(f_0,\tilde{f}) \leq C_n \inf_{t\in\mathscr{T}} R(f_0,\hat{f}_t) + \varepsilon_n$$

comparing the risk of the constructed estimator $\tilde{f}$ to the risk of the best available procedure in the collection $\{\hat{f}_t, t \in \mathscr{T}\}$. Our strategy is based on convex combination of these preliminary estimators and relies on PAC-Bayesian aggregation to obtain a single adaptive estimator. We focus on a wide family, commonly used in practice : affine estimators $\{\hat{f}_t(Y) = A_tY + b_t | A_t \in \mathscr{S}_n^+(\mathbb{R}), b_t \in \mathbb{R}^n, t \in \mathscr{T}\}$.

Aggregation procedures have been introduced by Vovk [48], Littlestone and Warmuth [35], Cesa-Bianchi et al. [14], Cesa-Bianchi and Lugosi [13]. They are a central ingredient of bagging [9], boosting [24, 42] or random forest (Amit and Geman [3] or Breiman [10]; or more recently Biau et al. [8], Biau and Devroye [7], Biau [6], Genuer [26]).

The general aggregation framework is detailed in Nemirovski [38] and studied in Catoni [11, 12] through a PAC-Bayesian framework as well as in Yang [51, 52, 53, 54, 55, 56, 57]. See for instance Tsybakov [47] for a survey. Optimal rates of aggregation in regression and density estimation are studied by Tsybakov [46], Lounici [36], Rigollet and Tsybakov [40], Rigollet [39] and Lecué [33].

A way to translate the confidence of each preliminary estimate is to aggregate according to a measure exponentially decreasing when the estimate's risk rises. This widely used strategy is called exponentially weighted aggregation.

More precisely, the weight of each element $\hat{f}_t$ in the collection is proportional to $\exp\left(-\frac{\widetilde{r}_t}{\beta}\right)\pi(t)$ where $\widetilde{r}_t$ is a, possibly penalized, estimate of the risk of $\hat{f}_t$, β is a positive parameter, called the temperature, that has to be calibrated and π is a prior measure over $\mathscr{T}$. The main interest of exponential weights resides in Lemma 1 [12] since they explicitly minimize the aggregated risk penalized by the Kullback-Leibler divergence to the prior measure π. Our aim is to give sufficient conditions on the risk estimate $\widetilde{r}_t$ and the temperature β to obtain an oracle inequality for the risk of the aggregate. Note that when the family $\mathscr{T}$ is countable, the exponentially weighted aggregate is a weighted sum of the preliminary estimates.

This procedure has shown its efficiency, offering lower risk than model selection because we bet on several estimators. Aggregation of projections has already been addressed by Leung and Barron [34]. They have proved by the mean of an oracle inequality, that in expectation, the aggregate performs almost as well as the best projection in the collection. Those results have been extended to several settings and noise conditions [19, 20, 21, 28, 22, 5, 17, 29, 44, 41] under a *frozen* estimator assumption: they should not depend on the observed sample. This restriction, not present in the work by Leung and Barron [34], has been removed by Dalalyan and Salmon [18] within the context of affine estimator and exponentially weighted aggregation. Nevertheless, they make additional assumptions on the matrices A_t and the Gaussian noise to obtain an optimal oracle inequality in expectation for affine estimates.

However, Dai et al. [15] have shown the sub-optimality in deviation of exponential weighting, not allowing to obtain a sharp oracle inequality in probability. Nevertheless, with Gaussian white noise, penalizing the risk in the weights and taking a temperature at least 20 times greater than the noise variance allows to upper bound the risk of the aggregate, based on affine estimators, in probability [16]. Furthermore, the corresponding oracle inequality is not sharp. Another work, by Gerchinovitz [27] , provides weak oracle inequality with high probability for projection estimates on non linear models. Alquier and Lounici [2] consider *frozen* and bounded preliminary estimators and obtain a sharp oracle inequality in deviation for the excess risk under a sparsity assumption, if the regression function is bounded, with a modified version of exponential weights. Under strong assumptions and independent noise, Bellec [4] provides a sharp oracle inequality with optimal rate for an other aggregation procedure called Q-aggregation. It is similar to exponential weights but the criterion to minimize is modified and the weights no longer are explicit.

In this article, we obtain an oracle inequality under a general sub-Gaussian noise assumption that does not require a coordinate independent setting. We also conduct an analysis of the relationship between the choice of the penalty and the temperature. In particular, we show that there is a continuum between the sup norm type penalty and the noise based one.

2 Framework and estimate

Recall that we observe

$$\forall i \in \{1,\dots,n\}, Y_i = f_0(x_i) + W_i$$

with f_0 an unknown function and x_i the fixed grid points. Our only assumption will be on the noise. We do not assume any independence between the coordinates W_i but only that $W = (W_i)_{i\leq n} \in \mathbb{R}^n$ is a centered sub-Gaussian variable. More precisely, we assume that $\mathbb{E}(W) = 0$ and there exists $\sigma^2 \in \mathbb{R}^+$ such that

$$\forall \alpha \in \mathbb{R}^n, \mathbb{E}\left[\exp\left(\alpha^\top W\right)\right] \leq \exp\left(\frac{\sigma^2}{2}\|\alpha\|_2^2\right),$$

where $\|.\|_2$ is the usual euclidean norm in $\mathbb{R}^n$. If W is a centered Gaussian vector with covariance matrix Σ then σ^2 is nothing but the largest eigenvalue of Σ.

The quality of our estimate will be measured through its error at the design points. More precisely, we will consider the classical euclidean loss, related to the squared norm

$$\|g\|_2^2 = \sum_{i=1}^n g(x_i)^2.$$

Thus, our unknown is the vector $(f_0(x_i))_{i=1}^n$ rather than the function f_0.

As announced, we will consider affine estimators $\hat{f}_t(Y) = A_tY + b_t$ corresponding to affine smoothed projection. We will assume that

$$\hat{f}_t(Y) = A_tY + b_t = \sum_{i=1}^n \rho_{t,i}\langle Y, g_{t,i}\rangle g_{t,i} + b_t$$

where $(g_{t,i})_{i=1}^n$ is an orthonormal basis, $(\rho_{t,i})_{i=1}^n$ a sequence of non-negative real numbers and $b_t \in \mathbb{R}^n$. By construction, A_t is thus a symmetric positive semi-definite real matrix. We assume furthermore that the matrix collection $\{A_t\}_{t\in\mathscr{T}}$ is such that there exists a finite $V > 0$ for which $\sup_{t\in\mathscr{T}} \|A_t\|_2 \leq V$. For sake of simplicity, we only use the notation $\hat{f}_t(Y) = A_tY + b_t$ in the following.

To define our estimate from the collection $\{\hat{f}_t(Y) = A_tY + b_t | A_t \in \mathscr{S}_n^+(\mathbb{R}), b_t \in \mathbb{R}^n, t \in \mathscr{T}\}$, we specify the estimate $\widetilde{r}_t$ of the (penalized) risk of the estimator $\hat{f}_t(Y)$, choose a prior probability measure π over $\mathscr{T}$ and a temperature $\beta > 0$. We define the exponentially weighted measure ρ_{EWA}, a probability measure over $\mathcal{T}$, by

$$d\rho_{EWA}(t) = \frac{\exp\left(-\frac{1}{\beta}\widetilde{r_t}\right)}{\int \exp\left(-\frac{1}{\beta}\widetilde{r_{t'}}\right) d\pi(t')} d\pi(t)$$

and the exponentially weighted aggregate f_{EWA} by $f_{EWA} = \int \hat{f}_t \, d\rho_{EWA}(t)$. If $\mathcal{T}$ is countable then

$$f_{EWA} = \sum_{t\in\mathcal{T}} \frac{e^{-\widetilde{r_t}/\beta}\pi_t}{\sum_{t'\in\mathcal{T}} e^{-\widetilde{r_{t'}}/\beta}\pi_{t'}} \hat{f}_t.$$

This construction naturally favors low risk estimates. When the temperature goes to zero this estimator becomes very similar to the one minimizing the risk estimate while it becomes an indiscriminate average when β grows to infinity. The choice of the temperature appears thus to be crucial and a low temperature seems to be desirable.

Our choice for the risk estimate $\widetilde{r_t}$ is to use the classical Stein unbiased estimate, which is sufficient to obtain optimal oracle inequalities in expectation,

$$r_t = \|Y - \hat{f}_t(Y)\|_2^2 + 2\sigma^2 \operatorname{Tr}(A_t) - n\sigma^2$$

and add a penalty $\operatorname{pen}(t)$. We will consider simultaneously the case of a penalty independent of f_0 and the one where the penalty may depend on an upper bound of (kind of) sup norm.

More precisely, we allow the use, at least in the analysis, of an upper bound $\widetilde{\|f_0\|_\infty}$ which can be thought as the supremum of the sup norm of the coefficients of f_0 in any basis appearing in $\mathcal{T}$. Indeed, we define $\widetilde{\|f_0\|_\infty}$ as the smallest non-negative real number C such that for any $t \in \mathcal{T}$,

$$\|A_t f_0\|_2^2 \leq C^2 \operatorname{Tr}(A_t^2).$$

By construction, $\widetilde{\|f_0\|_\infty}$ is smaller than the sup norm of any coefficients of f_0 in any basis appearing in the collection of estimators. Note that $\widetilde{\|f_0\|_\infty}$ can also be upper bounded by $\|f_0\|_1$, $\|f_0\|_2$ or $\sqrt{n}\|f_0\|_\infty$ where the ℓ_1 and sup norm can be taken in any basis.

Our aim is to obtain sufficient conditions on the penalty $\operatorname{pen}(t)$ and the temperature β so that an oracle inequality of type

$$\begin{aligned}\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\mu\in\mathcal{M}_+^1(\mathcal{T})} & (1+\varepsilon)\int \|f_0 - \hat{f}_t\|_2^2 d\mu(t) \\ & + (1+\varepsilon')\left(\int \operatorname{price}(t) d\mu(t) + 2\beta \operatorname{KL}(\mu,\pi) + \beta \ln\frac{1}{\eta}\right)\end{aligned}$$

holds either in probability or in expectation. Here, ε and ε' are some small non-negative numbers possibly equal to 0 and price(t) a loss depending on the choice of pen(t) and β. When $\mathscr{T}$ is countable, such an oracle proves that the risk of our aggregate estimate is of the same order as the one of the best estimate in the collection as it implies

$$\|f_0 - f_{EWA}\|_2^2 \leq \inf_{t\in\mathscr{T}} \left\{ (1+\varepsilon)\|f_0 - \hat{f}_t\|_2^2 + (1+\varepsilon') \left(\text{price}(t) + \beta \ln \frac{1}{\pi(t)^2\eta} \right) \right\}.$$

Before stating our more general result, which is in Section 4, we provide a comparison with some similar results in the literature.

3 Penalization strategies and preliminary results

The most similar result in the literature is the one from Dai et al. [16] which holds under a Gaussian white noise assumption and uses a penalty proportional to the known variance σ^2:

Proposition 1 (Dai et al. [16])**.** *If* $\text{pen}(t) = 2\sigma^2 \operatorname{Tr}(A_t)$*, and* $\beta \geq 4\sigma^2 \max(16, 5V)$*, then for all* $\eta > 0$*, with probability at least* $1-\eta$*,*

$$\|f_0 - f_{EWA}\|^2 \leq \min_t \left\{ \left(1 + \frac{128\sigma^2}{3\beta}\right) \|f_0 - \hat{f}_t\|^2 + 8\sigma^2 \operatorname{Tr}(A_t) \right.$$
$$\left. + 3\beta \ln \frac{1}{\pi_t} + 3\beta \ln \frac{1}{\eta} \right\}.$$

Our result generalizes this result to the non necessarily independent sub-Gaussian noise. We obtain, under the mild assumption that $V > 1/2$,

Proposition 2. *If* $\beta \geq 20\sigma^2 V$*, there exists* $\gamma \geq 0$*, such that if* $\text{pen}(t) \geq \frac{4\sigma^2}{\beta - 4\sigma^2 V} \operatorname{Tr}(A_t^2)\sigma^2$*, for any* $\eta > 0$*, with probability at least* $1-\eta$*,*

$$\|f_0 - f_{EWA}\|^2 \leq \inf_t \left\{ (1+\varepsilon) \|f_0 - \hat{f}_t\|^2 \right.$$
$$\left. + (1+\varepsilon') \left(\text{pen}(t) + 2\sigma^2 \operatorname{Tr}(A_t) + 2\beta \ln \frac{1}{\pi_t} + \beta \ln \frac{1}{\eta} \right) \right\}.$$

with $\varepsilon = \dfrac{4V^2\gamma}{(2V-1)(1-2V\gamma)}$ *and* $\varepsilon' = \dfrac{2V\gamma}{1-2V\gamma}$.

The parameter γ is explicit and satisfies $\varepsilon = O(\frac{\sigma^2}{\beta})$. We recover thus a similar weak oracle inequality under a weaker assumption on the noise. It should be noted that [4] obtains a sharp oracle inequality for a slightly different aggregation procedure but only under the very strong assumption that $\mathrm{Tr}(A_t) \le \ln \frac{1}{\pi(t)}$.

Following Guedj and Alquier [30], a lower bound on the penalty, that involves the sup norm of f_0, can be given. In that case, the oracle inequality is sharp as $\varepsilon = \varepsilon' = 0$. Furthermore, the parameter γ is not necessary and the minimum temperature is lower.

Proposition 3. *If $\beta > 4\sigma^2 V$, and*

$$\mathrm{pen}(t) \ge \frac{4\sigma^2}{\beta - 4\sigma^2 V}\left(\sigma^2 \,\mathrm{Tr}(A_t^2) + 2\left[\widetilde{\|f_0\|}_\infty^2 \,\mathrm{Tr}(A_t^2) + \|b_t\|_2^2\right]\right),$$

then for any $\eta > 0$, with probability at least $1-\eta$,

$$\begin{aligned}\|f_0 - f_{EWA}\|^2 \le \inf_t \Big\{ &\|f_0 - \hat{f}_t\|^2 + 2\sigma^2 \,\mathrm{Tr}(A_t) \\ &+ \frac{8\sigma^2}{\beta - 4\sigma^2 V}\left[\widetilde{\|f_0\|}_\infty^2 \,\mathrm{Tr}(A_t^2) + \|b_t\|_2^2\right] \\ &\qquad + \mathrm{pen}(t) + 2\beta \ln \frac{1}{\pi_t} + \beta \ln \frac{1}{\eta} \Big\}.\end{aligned}$$

The two results can be combined in a single one. Indeed, to obtain the first oracle inequality, we rely in the proof on bounds of type

$$\|(A_t - A_u) f_0 + b_t - b_u\|_2^2 \le C_1 \|\hat{f}_t - f_0\|_2^2 + C_2 \|\hat{f}_u - f_0\|_2^2,$$

with some constants C_1 and C_2 depending on γ which allows to link $\|A_t f_0 + b_t - A_u f_0 + b_u\|_2^2$ to $\|A_t Y + b_t - f_0\|_2^2$ and $\|A_u Y + b_u - f_0\|_2^2$. Whereas, for the second inequality we rely on bounds of type

$$\begin{aligned}\|(A_t - A_u) f_0 + b_t - b_u\|_2^2 &\le 4(\|A_t f_0\|_2^2 + \|A_u f_0\|_2^2 + \|b_t\|_2^2 + \|b_u\|_2^2) \\ &\le 4\left[\widetilde{\|f_0\|}_\infty^2 (\mathrm{Tr}(A_t^2) + \mathrm{Tr}(A_u^2)) + \|b_t\|_2^2 + \|b_u\|_2^2\right].\end{aligned}$$

Combining these two upper bounds produce weak oracle inequalities for a wider range of temperatures than Proposition 2, drawing a continuum between Proposition 2 and Proposition 3. More precisely, one obtains

Proposition 4. *For any $\delta \in [0,1]$, if $\beta \ge 4\sigma^2 V(1+4\delta)$ and $\beta > 4\sigma^2 V$, there exists $\gamma \ge 0$, such that if*

$$\mathrm{pen}(t) \ge \frac{4\sigma^2}{\beta - 4\sigma^2 V}\left(\sigma^2 \,\mathrm{Tr}(A_t^2) + 2(1-\delta)(1+2\gamma V)^2\left[\widetilde{\|f_0\|}_\infty^2 \,\mathrm{Tr}(A_t^2) + \|b_t\|_2^2\right]\right),$$

then for any $\eta > 0$*, with probability at least* $1-\eta$*,*

$$\|f_0 - f_{EWA}\|^2 \leq \inf_t \big\{ (1+\varepsilon) \|f_0 - \hat{f}_t\|^2 \\ + (1+\varepsilon') \left(\operatorname{price}(t) + 2\beta \ln \frac{1}{\pi_t} + \beta \ln \frac{1}{\eta} \right) \bigg\}.$$

with $\varepsilon = \dfrac{4V^2\gamma}{(2V-1)(1-2V\gamma)}$, $\varepsilon' = \dfrac{2V\gamma}{1-2V\gamma}$ *and*

$$\operatorname{price}(t) = \operatorname{pen}(t) + 2\sigma^2 \operatorname{Tr}(A_t) + \frac{8\sigma^2(1-\delta)(1+2\gamma V)^2}{\beta - 4\sigma^2 V} \left[\widetilde{\|f_0\|}_\infty^2 \operatorname{Tr}(A_t^2) + \|b_t\|_2^2 \right].$$

The convex combination parameter δ measures the account for signal to noise ratio in the penalty. We are now ready to state the central result of this paper, which gives an explicit expression for γ and introduce an optimization parameter $\nu > 0$.

4 A general oracle inequality

We consider now the general case for which $\mathscr{T}$ is not necessarily countable. Recall that we have defined the exponentially weighted measure ρ_{EWA}, a probability measure over $\mathscr{T}$, by

$$d\rho_{EWA}(t) = \frac{\exp\left(-\frac{1}{\beta}\widetilde{r}_t\right)}{\int \exp\left(-\frac{1}{\beta}\widetilde{r}_{t'}\right) d\pi(t')} d\pi(t)$$

and the exponentially weighted aggregate f_{EWA} by $f_{EWA} = \int \hat{f}_t \, d\rho_{EWA}(t)$. We will directly consider a lower bound on the penalty of the same type than in Proposition 4 and propositions similar to Propositions 2 and 3 will be obtained as straightforward corollaries.

Our main result is the following:

Theorem 4.1. *For any* $\delta \in [0,1]$*, if* $\beta \geq 4\sigma^2 V(1+4\delta)$*, and* $\beta > 4\sigma^2 V$*, let*

$$\gamma = \frac{\beta - 4\sigma^2 V(1+2\delta) - \sqrt{\beta - 4\sigma^2 V}\sqrt{\beta - 4\sigma^2 V(1+4\delta)}}{16\sigma^2 \delta V^2} \mathbb{1}_{\delta > 0}.$$

If for any $t \in \mathscr{T}$,

$$\operatorname{pen}(t) \geq \frac{4\sigma^2}{\beta - 4\sigma^2 V} \left(\sigma^2 \operatorname{Tr}(A_t^2) + 2(1-\delta)(1+2\gamma V)^2 \left[\widetilde{\|f_0\|}_\infty^2 \operatorname{Tr}(A_t^2) + \|b_t\|_2^2 \right] \right),$$

then

- *for any* $\eta \in (0,1]$, *with probability at least* $1-\eta$,

$$\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\nu \in N} \inf_{\mu \in \mathscr{M}_+^1(\mathscr{T})} (1+\varepsilon(\nu)) \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t)$$
$$+ (1+\varepsilon'(\nu)) \int \mathrm{price}(t) d\mu(t) + \beta(1+\varepsilon'(\nu)) \left(2\mathrm{KL}(\mu,\pi) + \ln\frac{1}{\eta}\right).$$

- *Furthermore*

$$\mathbb{E}\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\nu \in N} \inf_{\mu \in \mathscr{M}_+^1(\mathscr{T})} (1+\varepsilon(\nu)) \int \mathbb{E}\|f_0 - \hat{f}_t\|_2^2 d\mu(t)$$
$$+ (1+\varepsilon'(\nu)) \int \mathrm{price}(t) d\mu(t) + 2\beta(1+\varepsilon'(\nu))\mathrm{KL}(\mu,\pi),$$

with $\varepsilon(\nu) = \dfrac{1+\nu}{\nu}\dfrac{(1+\nu)\gamma}{1-(1+\nu)\gamma}$, $\varepsilon'(\nu) = \dfrac{(1+\nu)\gamma}{1-(1+\nu)\gamma}$,

$$\mathrm{price}(t) = \mathrm{pen}(t) + 2\sigma^2 \mathrm{Tr}(A_t) + \frac{8\sigma^2(1-\delta)}{\beta - 4\sigma^2 V}(1+2\gamma V)^2 \left[\widetilde{\|f_0\|}_\infty^2 \mathrm{Tr}(A_t^2) + \|b_t\|_2^2\right]$$

and $N = \{\nu > 0 | (1+\nu)\gamma < 1\}$.

When $\mathscr{T}$ is discrete, one can replace the minimization over all the probability measure $\mathscr{M}_+^1(\mathscr{T})$ by the minimization overall Dirac measure δ_{f_t} with $t \in \mathscr{T}$. Propositions 2, 3 and 4 are then straightforward corollaries corresponding to the allowed choice $\nu = 2V - 1$ when $V > 1/2$, and δ equals respectively to 1, 0 and anything between 0 and 1. Note that the result in expectation requires the same penalty, which is known not to be necessary, at least in the Gaussian case, as shown by Dalalyan and Salmon [18].

The parameter ν is a technical parameter that can be optimized, provided N is non empty. If $\delta > 0$, then for any $\beta \geq 4\sigma^2 V(1+4\delta)$, $0 < 2\gamma V \leq 1$. Thus $(0, 2V-1) \subseteq N$ as soon as $V > 1/2$ with $2V - 1 \in N$ if we assume that $\beta > 4\sigma^2 V(1+4\delta)$. If we assume $V \in (0, 1/2)$, we have to impose $\beta > 4\sigma^2 V + 2\sigma^2\delta(1+2V)^2$ in order to have a non empty N. Finally, if $\delta = 0$ then $\gamma = 0$ and $\varepsilon'(\nu) = 0, \varepsilon(\nu) = 0$, and no optimization is required.

We can specialize the results for $\delta = 1$ and $\delta = 0$: we obtain a weak oracle inequality

Corollary 1. *Under the assumptions of Theorem 4.1, if* $\beta \geq 20\sigma^2 V$, *let*

$$\gamma = \frac{\beta - 12\sigma^2 V - \sqrt{\beta - 4\sigma^2 V}\sqrt{\beta - 20\sigma^2 V}}{16\sigma^2 V^2}.$$

If for any $t \in \mathscr{T}$,

$$\mathrm{pen}(t) \geq \frac{4\sigma^4}{\beta - 4\sigma^2 V} \mathrm{Tr}(A_t^2),$$

then the oracle inequalities of Theorem 4.1 hold with

$$\mathrm{price}(t) = \mathrm{pen}(t) + 2\sigma^2 \mathrm{Tr}(A_t).$$

and an oracle inequality with leading constant 1 if we assume we know an upper bound of $\widetilde{\|f_0\|}_\infty^2$

Corollary 2. *Under the assumptions of Theorem 4.1, if $\beta > 4\sigma^2 V$, if for any $t \in \mathscr{T}$,*

$$\mathrm{pen}(t) \geq \frac{4\sigma^2}{\beta - 4\sigma^2 V} \left(\sigma^2 \mathrm{Tr}(A_t^2) + 2 \left[\widetilde{\|f_0\|}_\infty^2 \mathrm{Tr}(A_t^2) + \|b_t\|_2^2 \right] \right),$$

then the oracle inequalities of Theorem 4.1 hold with $\varepsilon = \varepsilon' = 0$ and

$$\mathrm{price}(t) = \mathrm{pen}(t) + 2\sigma^2 \mathrm{Tr}(A_t) + \frac{8\sigma^2}{\beta - 4\sigma^2 V} \left[\widetilde{\|f_0\|}_\infty^2 \mathrm{Tr}(A_t^2) + \|b_t\|_2^2 \right].$$

If we assume the penalty is given

$$\mathrm{pen}(t) = \kappa \mathrm{Tr}(A_t^2) \sigma^2,$$

one can give rewrite the assumption in term of κ. The weak oracle inequality holds for any temperature greater than $20\sigma^2 V$ as soon as $\kappa \geq \frac{4\sigma^2}{\beta - 4\sigma^2 V}$. while an exact oracle inequality holds for any vector f_0 and any temperature β greater than $4\sigma^2 V$ as soon as

$$\frac{\beta - 4\sigma^2 V}{4\sigma^2} \kappa - 1 \geq (1 + 2\gamma V)^2 \frac{\widetilde{\|f_0\|}_\infty^2}{\sigma^2}.$$

For fixed κ and β, this corresponds to a low peak signal to noise ratio $\frac{\widetilde{\|f_0\|}_\infty^2}{\sigma^2}$. Theorem 4.1 shows that there is a continuum between those two cases as weak oracle inequalities, with smaller leading constant than the one of Corollary 1, hold as soon as there exists $\delta \in [0,1]$ such that $\beta \geq 4\sigma^2(1 + 4\delta)V$ and

$$\frac{\beta - 4\sigma^2 V}{4\sigma^2} \kappa - 1 \geq (1 - \delta)(1 + 2\gamma V)^2 \frac{\widetilde{\|f_0\|}_\infty^2}{\sigma^2},$$

where the signal to noise ratio guides the transition. The temperature required remains nevertheless always above $4\sigma^2 V$.

Finally, the minimal temperature of $4\sigma^2 V(1+4\delta)$ can be replaced by some smaller values if one further restrict the smoothed projections used. As it appears in the proof, the temperature can be replaced by $4\sigma^2(1+\delta)$ or even $2\sigma^2(2+\delta)$ when the smoothed projections are respectively classical projections (see Theorem 5.1) and projections in the same basis. The question of the minimality of such temperature is still open. Note that in this proof, there is no loss due to the sub-Gaussianity assumption, since the same upper bound on the exponential moment of the deviation as in the Gaussian case are found, providing the same penalty and bound on temperature.

The proof of this result is quite long and thus postponed in Appendix 6. We provide first the generic proof of the oracle inequalities, highlighting the role of Gibbs measure and of some control in deviation. Then, we focus on the aggregation of projection estimators in the Gaussian model. This example already conveys all the ideas used in the complete proof of the deviation lemma : exponential moments inequalities for Gaussian quadratic form and the control of the bias $\|f_0 - A_t f\|_2^2$ by $\widetilde{\|f_0\|_\infty}^2$ on the one hand, to obtain an exact oracle inequality, and by $\|f_0 - A_t Y\|_2^2$ on the other hand, giving a weak inequality.

The extension to the general case is obtained by showing that similar exponential moments inequalities can be obtained for quadratic form of sub-Gaussian random variables, working along the fact that the systematic bias $\|f_0 - A_t f\|_2^2$ is no longer always smaller than $\|f_0 - A_t Y\|_2^2$ and providing a fine tuning optimization allowing the equality in the constraint on β and an optimization on the parameters ε.

We provide in the next section the sketch of proof of Theorem 4.1, as well as a proof in the simple case of orthogonal projection with Gaussian noise, meant to be compared with the one of [16].

5 Sketch of proof of the oracle inequalities

5.1 General sketch of proof

Theorem 4.1 relies on the characterization of Gibbs measure (Lemma 1) and a control of deviation of the empirical risk of any aggregate around its true risk, allowed by Lemma 2 or Lemma 3.

ρ is a Gibbs measure. Therefore it maximizes the entropy for a given expected energy. That is the subject of Lemma 1.1.3 in Catoni [12]:

Lemma 1. *For any bounded measurable function $h : \mathscr{T} \to \mathbb{R}$, and any probability* *distribution* $\rho \in \mathscr{M}_+^1(\mathscr{T})$ *such that* $\mathrm{KL}(\rho,\pi) < \infty$,

$$\log\left(\int \exp(h)d\pi\right) = \int h d\rho - \mathrm{KL}(\rho,\pi) + \mathrm{KL}(\rho,\pi_{\exp(h)}),$$

where by definition $\frac{d\pi_{\exp(h)}}{d\pi} = \frac{\exp[h(t)]}{\int \exp(h)d\pi}$. *Consequently,*

$$\log\left(\int \exp(h)d\pi\right) = \sup_{\rho\in\mathscr{M}_+^1(\mathscr{T})} \int h d\rho - \mathrm{KL}(\rho,\pi).$$

With $h(t) = -\frac{1}{\beta}[r_t + \mathrm{pen}(t)]$, this lemma states that for any probability distribution $\mu \in \mathscr{M}_+^1(\mathscr{T})$ such that $\mathrm{KL}(\mu,\pi) < \infty$,

$$\int h d\rho - \mathrm{KL}(\rho,\pi) \geq \int h d\mu - \mathrm{KL}(\mu,\pi).$$

Equivalently,

$$\begin{aligned}
&\int \|f_0 - \hat{f}_t\|_2^2 d\rho(t) + \int \left(r_t - \|f_0 - \hat{f}_t\|_2^2 + \mathrm{pen}(t)\right) d\rho(t) + \beta \mathrm{KL}(\rho,\pi) \\
&\qquad \leq \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t) + \int \left(r_t - \|f_0 - \hat{f}_t\|_2^2 + \mathrm{pen}(t)\right) d\mu(t) + \beta \mathrm{KL}(\mu,\pi) \\
\Leftrightarrow &\int \|f_0 - \hat{f}_t\|_2^2 d\rho(t) - \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t) \leq \int \left(\|f_0 - \hat{f}_t\|_2^2 - r_t\right) d\rho(t) \\
&\quad - \beta \mathrm{KL}(\rho,\pi) - \int \left(\|f_0 - \hat{f}_t\|_2^2 - r_t\right) d\mu(t) - \int \mathrm{pen}(t) d\rho(t) \\
&\qquad + \int \mathrm{pen}(t) d\mu(t) + \beta \mathrm{KL}(\mu,\pi).
\end{aligned}$$

The key is to upper bound the right-hand side with terms that may depend on ρ, but only through $\int \|f_0 - \hat{f}_t\|_2^2 d\rho(t)$ and Kullback-Leibler distance. This is the purpose of Lemma 2 in the case of Gaussian noise with projections estimators and Lemma 3 in the sub-Gaussian case. Under mild assumptions, they provide upper bounds in probability (and in expectation) of type:

$$\begin{aligned}
&\int \left(\|f_0 - \hat{f}_t\|_2^2 - r_t\right) d\rho(t) - \int \left(\|f_0 - \hat{f}_u\|_2^2 - r_u\right) d\mu(u) \\
&\qquad \leq C_1 \int \|f_0 - \hat{f}_t\|_2^2 d\rho(t) + C_2 \int \|f_0 - \hat{f}_u\|_2^2 d\mu(u) \\
&\qquad\qquad + \int \left(C_3 \mathrm{Tr}(A_t^2) + C_4 \|b_t\|_2^2\right) d\rho(t) \\
&\qquad + C_5 \int \mathrm{Tr}(A_u) d\mu(u) + \int \left(C_6 \mathrm{Tr}(A_u^2) + C_7 \|b_u\|_2^2\right) d\mu(u) \\
&\qquad\qquad + \beta \left(\mathrm{KL}(\rho,\pi) + \mathrm{KL}(\mu,\pi) + \ln\frac{1}{\eta}\right)
\end{aligned}$$

where C_1 to C_7 are known functions. Combining with the previous inequality and taking $\mathrm{pen}(t) \geq C_3 \operatorname{Tr}(A_t^2) + C_4\|b_t\|_2^2$ gives

$$
\begin{aligned}
(1-C_1)\int \|f_0-\hat{f}_t\|_2^2 d\rho(t) - (1+C_2)\int \|f_0-\hat{f}_t\|_2^2 d\mu(t) \\
\leq C_5 \int \operatorname{Tr}(A_u) d\mu(u) + \int \left(C_6 \operatorname{Tr}(A_u^2) + C_7\|b_u\|_2^2\right) d\mu(u) + \int \mathrm{pen}(u) d\mu(u) \\
+ \beta\left(2\mathrm{KL}(\mu,\pi) + \ln\frac{1}{\eta}\right).
\end{aligned}
$$

The additional condition $C_1 < 1$ allows to conclude. It is now clear that the whole work lies in the obtention of the lemma.

5.2 The expository case of Gaussian noise and projection estimates

In this subsection, to provide a simplified proof, we assume that A_t are the matrices of orthogonal projections, $b_t = 0$, and the noise W is a centered Gaussian random variable with variance $\sigma^2 I$. The previous theorem becomes:

Theorem 5.1. *Let π be an arbitrary prior measure over $\mathscr{T}$. For any $\delta \in [0,1]$, any $\beta > 4\sigma^2(\delta+1)$, the aggregate estimator f_{EWA} defined with*

$$\mathrm{pen}(t) \geq \frac{2\sigma^4}{\beta - 4\sigma^2}\left(1 + 2(1-\delta)\frac{\widetilde{\|f_0\|}_\infty^2}{\sigma^2}\right)\operatorname{Tr}(A_t)$$

satisfies the oracle inequalities of Theorem 4.1 with $\varepsilon = 2\varepsilon' = \dfrac{8\sigma^2\delta}{\beta - 4\sigma^2(\delta+1)}$ and

$$\mathrm{price}(t) = \mathrm{pen}(t) + 2\left(1 + \frac{2(1-\delta)\sigma^2}{\beta - 4\sigma^2}\frac{\widetilde{\|f_0\|}_\infty^2}{\sigma^2}\right)\operatorname{Tr}(A_t)\sigma^2$$

Note that the result may be further simplified using $\mathrm{price}(t) \leq 2\left(\mathrm{pen}(t) + \sigma^2 \operatorname{Tr}(A_t)\right)$.

As announced in the scheme of proof of the oracle inequalities (section 5), the key is a control of the deviation of the empirical risk of any aggregate around its true risk. It is allowed by Lemma 2 in this case.

Lemma 2. *For any prior probability distribution π, any $\delta \in [0,1]$ and any $\beta > 4\sigma^2$, for any probability distributions ρ and μ,*

- *For any* $\eta > 0$, *with probability at least* $1-\eta$,

$$\begin{aligned}\int \left(\|f_0-\hat{f}_t\|_2^2 - r_t\right) d\rho(t) - \int \left(\|f_0-\hat{f}_u\|_2^2 - r_u\right) d\mu(u) \\ \leq \frac{4\delta\sigma^2}{\beta-4\sigma^2}\left(\int \|f_0-\hat{f}_t\|_2^2 d\rho(t) + \int \|f_0-\hat{f}_u\|_2^2 d\mu(u)\right) \\ + \frac{4\sigma^2}{\beta-4\sigma^2}\left(\sigma^2 + (1-\delta)\widetilde{\|f_0\|}_\infty^2\right)\int \mathrm{Tr}(A_t) d\rho(t) \\ + 2\sigma^2\left(1+\frac{2(1-\delta)\widetilde{\|f_0\|}_\infty^2}{\beta-4\sigma^2}\right)\int \mathrm{Tr}(A_u) d\mu(u) \\ + \beta\left(\mathrm{KL}(\rho,\pi) + \mathrm{KL}(\mu,\pi) + \ln\frac{1}{\eta}\right)\end{aligned}$$

- *Moreover,*

$$\begin{aligned}\mathbb{E}\left[\int \left(\|f_0-\hat{f}_t\|_2^2 - r_t\right) d\rho(t) - \int \left(\|f_0-\hat{f}_u\|_2^2 - r_u\right) d\mu(u)\right] \\ \leq \mathbb{E}\left[\frac{4\delta\sigma^2}{\beta-4\sigma^2}\left(\int \|f_0-\hat{f}_t\|_2^2 d\rho(t) + \int \|f_0-\hat{f}_u\|_2^2 d\mu(u)\right)\right. \\ + \frac{4\sigma^2}{\beta-4\sigma^2}\left(\sigma^2 + (1-\delta)\widetilde{\|f_0\|}_\infty^2\right)\int \mathrm{Tr}(A_t) d\rho(t) \\ + 2\sigma^2\left(1+\frac{2(1-\delta)\widetilde{\|f_0\|}_\infty^2}{\beta-4\sigma^2}\right)\int \mathrm{Tr}(A_u) d\mu(u) \\ \left. + \beta\left(\mathrm{KL}(\rho,\pi) + \mathrm{KL}(\mu,\pi)\right)\right].\end{aligned}$$

We follow the scheme of proof given in section 5 and use Lemma 2, leading to the following result: for any $\eta > 0$, any prior probability distribution π, any $\delta \in [0,1]$ and any $\beta > 4\sigma^2(1+\delta)$, with probability at least $1-\eta$, for any probability distribution μ,

$$\begin{aligned}
\int \|f_0-\hat{f}_t\|_2^2 d\rho(t) &- \int \|f_0-\hat{f}_t\|_2^2 d\mu(t) \\
&\le \frac{4\delta\sigma^2}{\beta-4\sigma^2}\left(\int \|f_0-\hat{f}_t\|_2^2 d\rho(t) + \int \|f_0-\hat{f}_u\|_2^2 d\mu(u)\right) \\
&+ \frac{4\sigma^2}{\beta-4\sigma^2}\left(\sigma^2+(1-\delta)\widetilde{\|f_0\|}_\infty^2\right)\int \mathrm{Tr}(A_t)d\rho(t) - \int \mathrm{pen}(t)d\rho(t) \\
&+ 2\sigma^2\left(1+\frac{2(1-\delta)\widetilde{\|f_0\|}_\infty^2}{\beta-4\sigma^2}\right)\int \mathrm{Tr}(A_t)d\mu(t) + \int \mathrm{pen}(t)d\mu(t) \\
&+ \beta\left(2\mathrm{KL}(\mu,\pi)+\ln\frac{1}{\eta}\right).
\end{aligned}$$

With $\mathrm{pen}(t) \ge \frac{4\sigma^2}{\beta-4\sigma^2}\left(\sigma^2+(1-\delta)\widetilde{\|f_0\|}_\infty^2\right)\mathrm{Tr}(A_t)$, the previous inequality becomes

$$\begin{aligned}
\left(1-\frac{4\delta\sigma^2}{\beta-4\sigma^2}\right)&\int \|f_0-\hat{f}_t\|_2^2 d\rho(t) - \left(1+\frac{4\delta\sigma^2}{\beta-4\sigma^2}\right)\int \|f_0-\hat{f}_t\|_2^2 d\mu(t) \\
&\le 2\sigma^2\left(1+\frac{2(1-\delta)\widetilde{\|f_0\|}_\infty^2}{\beta-4\sigma^2}\right)\int \mathrm{Tr}(A_t)d\mu(t) + \int \mathrm{pen}(t)d\mu(t) \\
&+ \beta\left(2\mathrm{KL}(\mu,\pi)+\ln\frac{1}{\eta}\right).
\end{aligned}$$

Furthermore, using

$$\|f_0-f_{EWA}\|_2^2 \le \int \|f_0-\hat{f}_t\|_2^2 d\rho(t),$$

if $\beta > 4\sigma^2(\delta+1)$, we obtain

$$\begin{aligned}
\|f_0-f_{EWA}\|_2^2 \le \inf_{\mu\in\mathscr{M}_+^1(\mathscr{T})} &\left(1+\frac{8\sigma^2\delta}{\beta-4\sigma^2(\delta+1)}\right)\int \|f_0-\hat{f}_t\|_2^2 d\mu(t) \\
&+ \left(1+\frac{4\sigma^2\delta}{\beta-4\sigma^2(\delta+1)}\right)2\sigma^2\left(1+\frac{2(1-\delta)\widetilde{\|f_0\|}_\infty^2}{\beta-4\sigma^2}\right)\int \mathrm{Tr}(A_t)d\mu(t) \\
&+ \left(1+\frac{4\sigma^2\delta}{\beta-4\sigma^2(\delta+1)}\right)\int \mathrm{pen}(t)d\mu(t) \\
&+ \beta\left(1+\frac{4\sigma^2\delta}{\beta-4\sigma^2(\delta+1)}\right)\left(2\mathrm{KL}(\mu,\pi)+\ln\frac{1}{\eta}\right).
\end{aligned}$$

In addition, taking $\varepsilon = \frac{4\sigma^2\delta}{\beta - 4\sigma^2(\delta+1)}$, gives

$$\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\mu \in \mathscr{M}_+^1(\mathscr{T})} (1+2\varepsilon) \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t)$$
$$+ 2\sigma^2 (1+\varepsilon) \left(1 + \frac{2(1-\delta)\widetilde{\|f_0\|}_\infty^2}{\beta - 4\sigma^2}\right) \int \operatorname{Tr}(A_t) d\mu(t)$$
$$+ (1+\varepsilon) \left(\int \operatorname{pen}(t) d\mu(t) + 2\beta \operatorname{KL}(\mu, \pi) + \beta \ln \frac{1}{\eta}\right).$$

We focus now on the proof of Lemma 2 mixing control of exponential moments of a quadratic form of a Gaussian random variable with basic inequalities like Jensen, Fubini, and the important link between $\|f_0 - A_t f_0\|_2^2$ and $\|f_0 - A_t Y\|_2^2$. Note that this link is obvious in the case of orthogonal projections and need to be established differently in the general case, leading to technicalities (the introduction of γ).

Proof. For the sake of clarity, for any $t, u \in \mathscr{T}$, let

$$\Delta_{t,u} = \|f_0 - \hat{f}_t\|_2^2 - r_t - \|f_0 - \hat{f}_u\|_2^2 + r_u.$$

A simple calculation yields

$$\Delta_{t,u} = 2\left(W^\top (A_t - A_u) W + W^\top (A_t - A_u) f_0 - \sigma^2 \operatorname{Tr}(A_t - A_u)\right).$$

Since $(A_t)_{t \in \mathscr{T}}$ are positive semi-definite matrices, $W^\top (A_t - A_u) W \leq W^\top A_t W$, and there exist an orthogonal matrix U and a diagonal matrix D such that $A_t = U^\top D U$.

For any $\beta > 0$,

$$\mathbb{E}\left[\exp \frac{\Delta_{t,u}}{\beta}\right] \leq \mathbb{E}\left[\exp \frac{2}{\beta}\left((UW)^\top D (UW) + (UW)^\top U (A_t - A_u) f\right.\right.$$
$$\left.\left. - \sigma^2 \operatorname{Tr}(A_t - A_u)\right)\right].$$

Following lemma 2.4 of Hsu et al. [32], if $\beta > 4\sigma^2$,

$$\mathbb{E}\left[\exp \frac{\Delta_{t,u}}{\beta}\right] \leq \exp \frac{2\sigma^2}{\beta}\left(\operatorname{Tr}(A_u) + \frac{2\sigma^2 \operatorname{Tr}(A_t) + \|(A_t - A_u) f_0\|_2^2}{\beta - 4\sigma^2}\right). \tag{5.1}$$

Note that

$$\|(A_t - A_u) f_0\|_2^2 \leq 2\left(\|f_0 - A_t f_0\|_2^2 + \|f_0 - A_u f_0\|_2^2\right) \leq 2\left(\|f_0 - A_t Y\|_2^2 + \|f_0 - A_u Y\|_2^2\right)$$

and

$$\|(A_t - A_u)f_0\|_2^2 \leq 2\left(\|A_t f_0\|_2^2 + \|A_u f_0\|_2^2\right) \leq 2\widetilde{\|f_0\|}_\infty^2 \left(\mathrm{Tr}(A_t) + \mathrm{Tr}(A_u)\right).$$

Thus, for any $\beta > 4\sigma^2$, for any $\delta \in [0,1]$,

$$\begin{aligned}\mathbb{E}\exp\Bigg[\frac{\Delta_{t,u}}{\beta} - \frac{2\sigma^2}{\beta}\left(\mathrm{Tr}(A_u) + \frac{2\sigma^2\,\mathrm{Tr}(A_t)}{\beta - 4\sigma^2}\right) \\ - \frac{4\sigma^2\delta}{\beta(\beta - 4\sigma^2)}\left(\|f_0 - \hat{f}_t\|_2^2 + \|f_0 - \hat{f}_u\|_2^2\right) \\ - \frac{4\sigma^2}{\beta(\beta - 4\sigma^2)}(1-\delta)\widetilde{\|f_0\|}_\infty^2 \left(\mathrm{Tr}(A_t) + \mathrm{Tr}(A_u)\right)\Bigg] \leq 1.\end{aligned}$$

Along the same lines as Alquier and Lounici [2], we first integrate according to the prior π and use Fubini's theorem,

$$\begin{aligned}\mathbb{E}\int\int \exp\frac{1}{\beta}\Bigg[\Delta_{t,u} - 2\sigma^2\left(\mathrm{Tr}(A_u) + \frac{2\sigma^2\,\mathrm{Tr}(A_t)}{\beta - 4\sigma^2}\right) \\ - \frac{4\sigma^2\delta}{\beta - 4\sigma^2}\left(\|f_0 - \hat{f}_t\|_2^2 + \|f_0 - \hat{f}_u\|_2^2\right) \\ - \frac{4\sigma^2}{\beta - 4\sigma^2}(1-\delta)\widetilde{\|f_0\|}_\infty^2 \left(\mathrm{Tr}(A_t) + \mathrm{Tr}(A_u)\right)\Bigg] d\pi(t)d\pi(u) \leq 1,\end{aligned}$$

then introduce the probability distributions ρ and μ, and $\eta > 0$

$$\begin{aligned}\mathbb{E}\int\int \exp\frac{1}{\beta}\Bigg[\Delta_{t,u} - 2\sigma^2\left(\mathrm{Tr}(A_u) + \frac{2\sigma^2\,\mathrm{Tr}(A_t)}{\beta - 4\sigma^2}\right) \\ - \frac{4\sigma^2\delta}{\beta - 4\sigma^2}\left(\|f_0 - \hat{f}_t\|_2^2 + \|f_0 - \hat{f}_u\|_2^2\right) \\ - \frac{4\sigma^2(1-\delta)}{\beta - 4\sigma^2}\widetilde{\|f_0\|}_\infty^2 \left(\mathrm{Tr}(A_t) + \mathrm{Tr}(A_u)\right) \\ - \beta\left(\ln\frac{d\rho}{d\pi}(t) + \ln\frac{d\mu}{d\pi}(u) + \ln\frac{1}{\eta}\right)\Bigg] d\rho(t)d\mu(u) \leq \eta,\end{aligned}$$

before applying Jensen's inequality

$$\begin{aligned}\mathbb{E}\exp\frac{1}{\beta}\Bigg[\int\int\Delta_{t,u}d\rho(t)d\mu(u)\\
-\frac{4\delta\sigma^2}{\beta-4\sigma^2}\left(\int\|f_0-\hat{f}_t\|_2^2d\rho(t)+\int\|f_0-\hat{f}_u\|_2^2d\mu(u)\right)\\
-\frac{4\sigma^2}{\beta-4\sigma^2}\left(\sigma^2+(1-\delta)\widetilde{\|f_0\|}_\infty^2\right)\int\mathrm{Tr}(A_t)d\rho(t)\\
-\beta\left(\mathrm{KL}(\rho,\pi)+\mathrm{KL}(\mu,\pi)+\ln\frac{1}{\eta}\right)\\
-2\sigma^2\left(1+\frac{2(1-\delta)\widetilde{\|f_0\|}_\infty^2}{\beta-4\sigma^2}\right)\int\mathrm{Tr}(A_u)d\mu(u)\Bigg]\leq\eta.\end{aligned}\tag{5.2}$$

Finally, using the basic inequality $\exp(x)\geq\mathbb{1}_{\mathbb{R}_+}(x)$,

$$\begin{aligned}\mathbb{P}\Bigg[\int\int\Delta_{t,u}d\rho(t)d\mu(u)\leq\frac{4\delta\sigma^2}{\beta-4\sigma^2}\left(\int\|f_0-\hat{f}_t\|_2^2d\rho(t)+\int\|f_0-\hat{f}_u\|_2^2d\mu(u)\right)\\
+\frac{4\sigma^2}{\beta-4\sigma^2}\left(\sigma^2+(1-\delta)\widetilde{\|f_0\|}_\infty^2\right)\int\mathrm{Tr}(A_t)d\rho(t)\\
+\beta\left(\mathrm{KL}(\rho,\pi)+\mathrm{KL}(\mu,\pi)+\ln\frac{1}{\eta}\right)\\
+\frac{2\sigma^2}{n}\left(1+\frac{2(1-\delta)n\widetilde{\|f_0\|}_\infty^2}{\beta n-4\sigma^2}\right)\int\mathrm{Tr}(A_u)d\mu(u)\Bigg]\geq1-\eta.\end{aligned}$$

The result in expectation is obtained by Equation (5.2) with $\eta=1$:

$$\begin{aligned}\mathbb{E}\exp\frac{1}{\beta}\Bigg[\int\int\Delta_{t,u}d\rho(t)d\mu(u)\\
-\frac{4\delta\sigma^2}{\beta-4\sigma^2}\left(\int\|f_0-\hat{f}_t\|_2^2d\rho(t)+\int\|f_0-\hat{f}_u\|_2^2d\mu(u)\right)\\
-\frac{4\sigma^2}{\beta-4\sigma^2}\left(\sigma^2+(1-\delta)\widetilde{\|f_0\|}_\infty^2\right)\int\mathrm{Tr}(A_t)d\rho(t)-\beta\left(\mathrm{KL}(\rho,\pi)+\mathrm{KL}(\mu,\pi)\right)\\
-2\sigma^2\left(1+\frac{2(1-\delta)\widetilde{\|f_0\|}_\infty^2}{\beta-4\sigma^2}\right)\int\mathrm{Tr}(A_u)d\mu(u)\Bigg]\leq1,\end{aligned}$$

combined with the inequality $t\leq\exp(t)-1$. □

6 Appendix : Proofs in the sub-Gaussian case

6.1 Proof of Theorem 4.1

The proof follows from the scheme described in section 5. The main point is still to control

$$\int \left(\|f_0 - \hat{f}_t\|_2^2 - r_t\right) d\rho(t) - \int \left(\|f_0 - \hat{f}_t\|_2^2 - r_t\right) d\mu(t).$$

We recall that A_t is a symmetric positive semi-definite matrix, there exists $V > 0$ such that $\sup_{t\in\mathscr{T}} \|A_t\|_2 \leq V$ and W is a centered sub-Gaussian noise. For any $t, u \in \mathscr{T}$, we still denote $\Delta_{t,u} = \|f_0 - \hat{f}_t\|_2^2 - r_t - \|f_0 - \hat{f}_u\|_2^2 + r_u$.

Lemma 3. *Let π be an arbitrary prior probability. For any $\delta \in [0,1]$, any $\beta > 4\sigma^2 V$ and $\beta \geq 4\sigma^2 V(1+4\delta)$, let*

$$\gamma = \frac{1}{16\sigma^2\delta V^2}\left(\beta - 4\sigma^2 V(1+2\delta) - \sqrt{\beta - 4\sigma^2 V}\sqrt{\beta - 4\sigma^2 V(1+4\delta)}\right)\mathbb{1}_{\delta>0}.$$

Then, for any probability distributions ρ and μ, for any $\nu > 0$,

- *for any $\eta \in (0,1]$, with probability at least $1-\eta$,*

$$\begin{aligned}
\int\int \Delta_{t,u} d\rho(t) d\mu(u) &\leq (1+\nu)\gamma \int \|\hat{f}_t - f_0\|_2^2 d\rho(t) \\
&+ \frac{8\sigma^2}{\beta - 4\sigma^2 V}(1-\delta)(1+2\gamma V)^2 \int \left[\widetilde{\|f_0\|}_\infty^2 \operatorname{Tr}(A_t^2) + \|b_t\|_2^2\right] d\rho(t) \\
&+ \frac{4\sigma^4}{\beta - 4\sigma^2 V}\int \operatorname{Tr}(A_t^2) d\rho(t) + 2\sigma^2 \int \operatorname{Tr}(A_u) d\mu(u) \\
&+ \frac{8\sigma^2}{\beta - 4\sigma^2 V}(1-\delta)(1+2\gamma V)^2 \int \left[\widetilde{\|f_0\|}_\infty^2 \operatorname{Tr}(A_u^2) + \|b_u\|_2^2\right] d\mu(u) \\
&+ \left(1+\frac{1}{\nu}\right)\gamma \int \|\hat{f}_u - f_0\|_2^2 d\mu(u) + \beta\left(\mathrm{KL}(\rho,\pi) + \mathrm{KL}(\mu,\pi) + \ln\frac{1}{\eta}\right).
\end{aligned}$$

- *Moreover,*

$$\mathbb{E}\left[\int\int \Delta_{t,u} d\rho(t) d\mu(u)\right] \leq \mathbb{E}\Big[(1+\nu)\gamma\int \|\hat{f}_t - f_0\|_2^2 d\rho(t)$$
$$+ \frac{8\sigma^2}{\beta - 4\sigma^2 V}(1-\delta)(1+2\gamma V)^2 \int \left[\widetilde{\|f_0\|}_\infty^2 \operatorname{Tr}(A_t^2) + \|b_t\|_2^2\right] d\rho(t)$$
$$+ \frac{4\sigma^4}{\beta - 4\sigma^2 V}\int \operatorname{Tr}(A_t^2) d\rho(t) + 2\sigma^2 \int \operatorname{Tr}(A_u) d\mu(u)$$
$$+ \frac{8\sigma^2}{\beta - 4\sigma^2 V}(1-\delta)(1+2\gamma V)^2 \int \left[\widetilde{\|f_0\|}_\infty^2 \operatorname{Tr}(A_u^2) + \|b_u\|_2^2\right] d\mu(u)$$
$$+ \left(1 + \frac{1}{\nu}\right)\gamma \int \|\hat{f}_u - f_0\|_2^2 d\mu(u) + \beta\left(\mathrm{KL}(\rho,\pi) + \mathrm{KL}(\mu,\pi)\right)\Big].$$

Under the assumptions of the previous lemma, with probability at least $1-\eta$,

$$\int \|f_0 - \hat{f}_t\|_2^2 d\rho(t) - \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t) \leq (1+\nu)\gamma \int \|\hat{f}_t - f_0\|_2^2 d\rho(t)$$
$$+ \frac{4\sigma^2}{\beta - 4\sigma^2 V}\int \left(\sigma^2 \operatorname{Tr}(A_t^2) + 2(1-\delta)(1+2\gamma V)^2 \left[\widetilde{\|f_0\|}_\infty^2 \operatorname{Tr}(A_t^2) + \|b_t\|_2^2\right]\right) d\rho(t)$$
$$+ 2\sigma^2 \left(\int \operatorname{Tr}(A_t) d\mu(t) + \frac{4(1-\delta)(1+2\gamma V)^2}{\beta - 4\sigma^2 V}\int \left[\widetilde{\|f_0\|}_\infty^2 \operatorname{Tr}(A_t^2) + \|b_t\|_2^2\right] d\mu(t)\right)$$
$$- \int \operatorname{pen}(t) d\rho(t) + \int \operatorname{pen}(t) d\mu(t) + \left(1 + \frac{1}{\nu}\right)\gamma \int \|\hat{f}_t - f_0\|_2^2 d\mu(t)$$
$$+ \beta\left(2\mathrm{KL}(\mu,\pi) + \ln\frac{1}{\eta}\right).$$

Taking

$$\operatorname{pen}(t) \geq \frac{4\sigma^2}{\beta - 4\sigma^2 V}\left(\sigma^2 \operatorname{Tr}(A_t^2) + 2(1-\delta)(1+2\gamma V)^2 \left[\widetilde{\|f_0\|}_\infty^2 \operatorname{Tr}(A_t^2) + \|b_t\|_2^2\right]\right),$$

and $\nu \in N = \{\nu > 0 | (1+\nu)\gamma < 1\}$, such that the inequality stays informative,

$$(1 - (1+\nu)\gamma)\int \|f_0 - \hat{f}_t\|_2^2 d\rho(t) \leq \left(1 + \left(1 + \frac{1}{\nu}\right)\gamma\right)\int \|f_0 - \hat{f}_t\|_2^2 d\mu(t)$$
$$+ 2\sigma^2 \left(\int \operatorname{Tr}(A_t) d\mu(t) + \frac{4(1-\delta)(1+2\gamma V)^2}{\beta - 4\sigma^2 V}\int \left[\widetilde{\|f_0\|}_\infty^2 \operatorname{Tr}(A_t^2) + \|b_t\|_2^2\right] d\mu(t)\right)$$
$$+ \int \operatorname{pen}(t) d\mu(t) + \beta\left(2\mathrm{KL}(\mu,\pi) + \ln\frac{1}{\eta}\right).$$

Finally, since $\|f_0 - f_{EWA}\|_2^2 \leq \int \|f_0 - \hat{f}_t\|_2^2 d\rho(t)$,

$$
\begin{aligned}
\|f_0 - f_{EWA}\|_2^2 \leq \left(1 + \frac{(1+\nu)^2\gamma}{\nu(1-(1+\nu)\gamma)}\right) \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t) \\
+ \frac{2\sigma^2}{1-(1+\nu)\gamma} \left(\int \mathrm{Tr}(A_t) d\mu(t) \right. \\
+ \frac{4(1-\delta)(1+2\gamma V)^2}{\beta - 4\sigma^2 V} \int \left[\widetilde{\|f_0\|}_\infty^2 \mathrm{Tr}(A_t^2) + \|b_t\|_2^2 \right] d\mu(t) \Bigg) \\
+ \frac{1}{1-(1+\nu)\gamma} \left(\int \mathrm{pen}(t) d\mu(t) + \beta \left(2\mathrm{KL}(\mu, \pi) + \ln \frac{1}{\eta} \right) \right).
\end{aligned}
$$

The result in expectation is obtained in the same fashion.

6.2 Proof of Lemma 3

The exponential moment of $\Delta_{t,u}$ is easily controlled by a term involving $\|A_t f_0 + b_t - f_0\|_2^2$ (see Equation (5.1)). Since A_t are not projections, $\|A_t f_0 + b_t - f_0\|_2^2 \leq \|A_t Y - f_0\|_2^2$ does not hold any more. The presence of $\|A_t Y - f_0\|_2^2$ allows us to obtain a weak oracle inequality. To overcome this difficulty, $\|(A_t - A_u)Y\|_2^2$ is introduced and for an arbitrary $\gamma \geq 0$, we try to control $\Delta_{t,u} - \gamma\|(A_t - A_u)Y\|_2^2$.

Proof. A simple calculation yields

$$
\begin{aligned}
\Delta_{t,u} - \gamma\|\hat{f}_t - \hat{f}_u\|_2^2 = W^\top (2I - \gamma(A_t - A_u)^\top)(A_t - A_u)W \\
+ 2W^\top (I - \gamma(A_t - A_u)^\top) \left[(A_t - A_u) f_0 + b_t - b_u \right] \\
- 2\sigma^2 \mathrm{Tr}(A_t - A_u) - \gamma\|(A_t - A_u) f_0 + b_t - b_u\|_2^2.
\end{aligned}
$$

Noting that $W^\top(2I - \gamma(A_t - A_u)^\top)(A_t - A_u)W \leq 2W^\top(A_t - A_u)W$ and since $(A_t)_{t\in\mathscr{T}}$ are positive semi-definite matrices, $2W^\top(A_t - A_u)W \leq 2W^\top A_t W$. Thus, for any $\beta > 0$, any $\gamma \geq 0$,

$$
\begin{aligned}
&\mathbb{E} \exp\left(\frac{\Delta_{t,u}}{\beta} - \frac{\gamma}{\beta}\|\hat{f}_t - \hat{f}_u\|_2^2 \right) \\
&\leq \mathbb{E}\left[\exp \frac{2}{\beta} \left(W^\top A_t W + W^\top (I - \gamma(A_t - A_u)) \left[(A_t - A_u) f_0 + b_t - b_u \right] \right) \right] \\
&\qquad \times \exp \frac{-1}{\beta} \left(2\sigma^2 \mathrm{Tr}(A_t - A_u) + \gamma\|(A_t - A_u) f_0 + b_t - b_u\|_2^2 \right).
\end{aligned}
$$

The first step is to bring us back to the Gaussian case, using W's sub-Gaussianity and an idea of Hsu et al. [32]. Let Z be a standard Gaussian random variable, independent of W. Then,

$$
\begin{aligned}
&\mathbb{E}\exp\left(\frac{2}{\sqrt{\beta}}W^\top\sqrt{A_t}Z+\frac{2}{\beta}W^\top(I-\gamma(A_t-A_u))\left[(A_t-A_u)f_0+b_t-b_u\right]\right)\\
&\quad=\mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{2}{\sqrt{\beta}}W^\top\sqrt{A_t}Z+\frac{2}{\beta}W^\top(I-\gamma(A_t-A_u))\left[(A_t-A_u)f_0+b_t-b_u\right]\right)\big|W\right]\right]\\
&\quad=\mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{2}{\sqrt{\beta}}W^\top\sqrt{A_t}Z\right)\big|W\right]\right.\\
&\qquad\qquad\qquad\qquad\left.\times\exp\left(\frac{2}{\beta}W^\top(I-\gamma(A_t-A_u))\left[(A_t-A_u)f_0+b_t-b_u\right]\right)\right]\\
&\quad=\mathbb{E}\exp\frac{2}{\beta}\left(W^\top A_tW+W^\top(I-\gamma(A_t-A_u))\left[(A_t-A_u)f_0+b_t-b_u\right]\right).
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
&\mathbb{E}\left[\exp\frac{2}{\beta}\left(W^\top A_tW+W^\top(I-\gamma(A_t-A_u))\left[(A_t-A_u)f_0+b_t-b_u\right]\right)\right]\\
&=\mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{2}{\sqrt{\beta}}W^\top\sqrt{A_t}Z+\frac{2}{\beta}W^\top(I-\gamma(A_t-A_u))\left[(A_t-A_u)f_0+b_t-b_u\right]\right)\big|Z\right]\right].
\end{aligned}
$$

Since W is sub-Gaussian with parameter σ,

$$
\begin{aligned}
&\mathbb{E}\left[\exp\frac{2}{\beta}\left(W^\top A_tW+W^\top(I-\gamma(A_t-A_u))\left[(A_t-A_u)f_0+b_t-b_u\right]\right)\right]\\
&\le\mathbb{E}\exp\left(\frac{\sigma^2}{2}\left\|\frac{2}{\sqrt{\beta}}\left(\sqrt{A_t}Z+\frac{1}{\sqrt{\beta}}(I-\gamma(A_t-A_u))\left[(A_t-A_u)f_0+b_t-b_u\right]\right)\right\|_2^2\right)
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\mathbb{E}\exp\left(\frac{\Delta_{t,u}}{\beta}-\frac{\gamma}{\beta}\|\hat{f}_t-\hat{f}_u\|_2^2\right)\\
&\leq \mathbb{E}\left[\exp\frac{2\sigma^2}{\beta}\left(Z^\top A_t Z+\frac{2}{\sqrt{\beta}}Z^\top\sqrt{A_t}(I-\gamma(A_t-A_u))\left[(A_t-A_u)f_0+b_t-b_u\right]\right)\right]\\
&\quad\times\exp\left(\frac{2\sigma^2}{\beta^2}\left\|(I-\gamma(A_t-A_u))\left[(A_t-A_u)f_0+b_t-b_u\right]\right\|_2^2-\frac{2\sigma^2}{\beta}\operatorname{Tr}(A_t-A_u)\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\times\exp\left(-\frac{\gamma}{\beta}\|(A_t-A_u)f_0+b_t-b_u\|_2^2\right).
\end{aligned}
$$

The expectation is similar to the one obtained in the Gaussian case: the exponential of some quadratic form. The same recipe is applied. Since A_t is positive semi-definite, there exist an orthogonal matrix U and a diagonal matrix D such that $A_t=U^\top DU$. Note that UZ is a standard Gaussian variable. This diagonalization step and the non-negativity of the eigenvalues allow to apply Lemma 2.4 of Hsu et al. [32]. Then, for any $\beta>4\sigma^2V$, any $\gamma\geq 0$,

$$
\begin{aligned}
&\mathbb{E}\exp\left(\frac{\Delta_{t,u}}{\beta}-\frac{\gamma}{\beta}\|\hat{f}_t-\hat{f}_u\|_2^2\right)\leq\exp\left(\frac{2\sigma^2}{\beta}\operatorname{Tr}(A_t)+\frac{4\sigma^4}{\beta(\beta-4\sigma^2V)}\operatorname{Tr}(A_t^2)\right)\\
&\quad\times\exp\left(\frac{8\sigma^4}{\beta^2(\beta-4\sigma^2V)}\left\|\sqrt{A_t}(I-\gamma(A_t-A_u))\left[(A_t-A_u)f_0+b_t-b_u\right]\right\|_2^2\right)\\
&\quad\times\exp\left(\frac{2\sigma^2}{\beta^2}\left\|(I-\gamma(A_t-A_u))\left[(A_t-A_u)f_0+b_t-b_u\right]\right\|_2^2-\frac{2\sigma^2}{\beta}\operatorname{Tr}(A_t-A_u)\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\times\exp\left(-\frac{\gamma}{\beta}\|(A_t-A_u)f_0+b_t-b_u\|_2^2\right).
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
&\mathbb{E}\exp\left(\frac{\Delta_{t,u}}{\beta}+\frac{\gamma}{\beta}\left(\|(A_t-A_u)f_0+b_t-b_u\|_2^2-\|\hat{f}_t-\hat{f}_u\|_2^2\right)\right)\\
&\leq\exp\frac{2\sigma^2}{\beta}\left(\operatorname{Tr}(A_u)+\frac{2\sigma^2}{\beta-4\sigma^2V}\operatorname{Tr}(A_t^2)\right)\\
&\quad\times\exp\left(\frac{2\sigma^2}{\beta^2}\left(\frac{4\sigma^2V}{\beta-4\sigma^2V}(1+2\gamma V)^2+(1+2\gamma V)^2\right)\|(A_t-A_u)f_0+b_t-b_u\|_2^2\right)\\
&\leq\exp\frac{2\sigma^2}{\beta}\left(\operatorname{Tr}(A_u)+\frac{2\sigma^2}{\beta-4\sigma^2V}\operatorname{Tr}(A_t^2)+\frac{(1+2\gamma V)^2}{\beta-4\sigma^2V}\|(A_t-A_u)f_0+b_t-b_u\|_2^2\right).
\end{aligned}
$$

If an exact oracle inequality is wished, $\|(A_t-A_u)f_0+b_t-b_u\|_2^2$ should be upper bounded by some constant and γ should be set to zero. Else, γ is used to *replace* the terms in $\|(A_t - A_u)f_0 + b_t - b_u\|_2^2$ by $\|(A_t - A_u)Y + b_t - b_u\|_2^2$. Thus, the terms depending on f_0 will be upper bounded in two ways:

- on the one hand, using $\widetilde{\|f_0\|}_\infty^2$

$$\|(A_t - A_u)f_0 + b_t - b_u\|_2^2 \le 4\left(\|A_t f_0\|_2^2 + \|A_u f_0\|_2^2 + \|b_t\|_2^2 + \|b_u\|_2^2\right)$$
$$\le 4\left(\left(\mathrm{Tr}(A_t^2) + \mathrm{Tr}(A_u^2)\right)\widetilde{\|f_0\|}_\infty^2 + \|b_t\|_2^2 + \|b_u\|_2^2\right)$$

  For any $\delta \in [0,1]$,

$$\begin{aligned}
&\mathbb{E}\exp\left(\frac{\Delta_{t,u}}{\beta} + \frac{\gamma}{\beta}\left(\|(A_t - A_u)f_0 + b_t - b_u\|_2^2 - \|\hat{f}_t - \hat{f}_u\|_2^2\right)\right)\\
&\le \exp\frac{2\sigma^2}{\beta}\left(\mathrm{Tr}(A_u) + \frac{2\sigma^2}{\beta - 4\sigma^2 V}\mathrm{Tr}(A_t^2)\right)\\
&\qquad \times \exp\left(\frac{2\sigma^2}{\beta}\frac{(1+2\gamma V)^2(1-\delta)}{\beta - 4\sigma^2 V}\|(A_t - A_u)f_0 + b_t - b_u\|_2^2\right)\\
&\qquad\qquad \times \exp\left(\frac{2\sigma^2(1+2\gamma V)^2\delta}{\beta(\beta - 4\sigma^2 V)}\|(A_t - A_u)f_0 + b_t - b_u\|_2^2\right)\\
&\le \exp\frac{2\sigma^2}{\beta}\left(\mathrm{Tr}(A_u) + \frac{2\sigma^2}{\beta - 4\sigma^2 V}\mathrm{Tr}(A_t^2) + \frac{(1+2\gamma V)^2\delta}{\beta - 4\sigma^2 V}\|(A_t - A_u)f_0 + b_t - b_u\|_2^2\right)\\
&\times \exp\left(\frac{8\sigma^2(1+2\gamma V)^2(1-\delta)}{\beta(\beta - 4\sigma^2 V)}\left[\left(\mathrm{Tr}(A_t^2) + \mathrm{Tr}(A_u^2)\right)\widetilde{\|f_0\|}_\infty^2 + \|b_t\|_2^2 + \|b_u\|_2^2\right]\right).
\end{aligned}$$

- on the other hand, introducing $\|\hat{f}_t - f_0\|_2^2$ to obtain a weak oracle inequality: conditions should be found on γ such that

$$\begin{aligned}
&\frac{2\sigma^2(1+2\gamma V)^2\delta}{\beta - 4\sigma^2 V}\|(A_t - A_u)f_0 + b_t - b_u\|_2^2\\
&\qquad - \gamma\left(\|(A_t - A_u)f_0 + b_t - b_u\|_2^2 - \|\hat{f}_t - \hat{f}_u\|_2^2\right)\\
&\qquad\qquad \le C_1\|\hat{f}_t - f_0\|_2^2 + C_2\|\hat{f}_u - f_0\|_2^2
\end{aligned}$$

  for some non-negative constants C_1 and C_2 and with $\delta > 0$. Since for any $\nu > 0$, $\|\hat{f}_t - \hat{f}_u\|_2^2 \le (1+\nu)\|\hat{f}_t - f_0\|_2^2 + \left(1 + \frac{1}{\nu}\right)\|\hat{f}_u - f_0\|_2^2$, it suffices that

$$\frac{2\sigma^2(1+2\gamma V)^2\delta}{\beta - 4\sigma^2 V}\|(A_t - A_u)f_0 + b_t - b_u\|_2^2 - \gamma\|(A_t - A_u)f_0 + b_t - b_u\|_2^2 \le 0.$$

This condition may be fulfilled if $\beta \geq 4\sigma^2 V(1+4\delta)$. The smallest $\gamma \geq 0$ among all the possible ones is chosen :

$$\gamma = \frac{1}{16\sigma^2\delta V^2}\left(\beta - 4\sigma^2 V(1+2\delta) - \sqrt{\beta - 4\sigma^2 V}\sqrt{\beta - 4\sigma^2 V(1+4\delta)}\right)\mathbb{1}_{\delta>0}.$$

This leads to the following inequality : for any $\delta \in [0,1]$, for any $\beta > 4\sigma^2 V$ and $\beta \geq 4\sigma^2 V(1+4\delta)$, with γ previously defined, for any $\nu > 0$,

$$\begin{aligned}
&\mathbb{E}\exp\left(\frac{\Delta_{t,u}}{\beta} - \frac{\gamma}{\beta}\left((1+\nu)\|\hat{f}_t - f_0\|_2^2 + \left(1+\frac{1}{\nu}\right)\|\hat{f}_u - f_0\|_2^2\right)\right) \\
&\leq \exp\left(\frac{8\sigma^2(1+2\gamma V)^2(1-\delta)}{\beta(\beta-4\sigma^2 V)}\left[\left(\mathrm{Tr}(A_t^2)+\mathrm{Tr}(A_u^2)\right)\widetilde{\|f_0\|}_\infty^2 + \|b_t\|_2^2 + \|b_u\|_2^2\right]\right) \\
&\qquad\qquad \times \exp\frac{2\sigma^2}{\beta}\left(\mathrm{Tr}(A_u) + \frac{2\sigma^2}{\beta-4\sigma^2 V}\mathrm{Tr}(A_t^2)\right).
\end{aligned}$$

The rest of the proof follows the same steps as in the Gaussian case: we first integrate according to the prior π, use Fubini's theorem, introduce the probability measures ρ and μ and apply Jensen's inequality to obtain that for any $\eta \in (0,1]$,

$$\begin{aligned}
&\mathbb{E}\exp\frac{1}{\beta}\Big[\int\int \Delta_{t,u}d\rho(t)d\mu(u) - (1+\nu)\gamma\int\|\hat{f}_t - f_0\|_2^2 d\rho(t) \\
&-\frac{4\sigma^2}{\beta-4\sigma^2 V}\int\left(\sigma^2\mathrm{Tr}(A_t^2) + 2(1-\delta)(1+2\gamma V)^2\left[\widetilde{\|f_0\|}_\infty^2\mathrm{Tr}(A_t^2) + \|b_t\|_2^2\right]\right)d\rho(t) \\
&-2\sigma^2\left(\int\mathrm{Tr}(A_u)d\mu(u) + \frac{4(1-\delta)(1+2\gamma V)^2}{\beta-4\sigma^2 V}\int\left[\widetilde{\|f_0\|}_\infty^2\mathrm{Tr}(A_u^2) + \|b_u\|_2^2\right]d\mu(u)\right) \\
&-\left(1+\frac{1}{\nu}\right)\gamma\int\|\hat{f}_u - f_0\|_2^2 d\mu(u) - \beta\left(\mathrm{KL}(\rho,\pi) + \mathrm{KL}(\mu,\pi) + \ln\frac{1}{\eta}\right)\Big] \leq \eta.
\end{aligned} \tag{6.1}$$

Finally, using $\exp(x) \geq \mathbb{1}_{\mathbb{R}_+}(x)$, for any $\delta \in [0,1]$, any $\beta > 4\sigma^2 V$ and $\beta \geq 4\sigma^2 V(1+4\delta)$, with γ previously defined, for any $\eta \in (0,1]$, for any $\nu > 0$,

$$\begin{aligned}
&\mathbb{P}\Big[\int\int \Delta_{t,u}d\rho(t)d\mu(u) \leq (1+\nu)\gamma\int\|\hat{f}_t - f_0\|_2^2 d\rho(t) \\
&+\frac{4\sigma^2}{\beta-4\sigma^2 V}\int\left(\sigma^2\mathrm{Tr}(A_t^2) + 2(1-\delta)(1+2\gamma V)^2\left[\widetilde{\|f_0\|}_\infty^2\mathrm{Tr}(A_t^2) + \|b_t\|_2^2\right]\right)d\rho(t) \\
&+2\sigma^2\left(\int\mathrm{Tr}(A_u)d\mu(u) + \frac{4(1-\delta)(1+2\gamma V)^2}{\beta-4\sigma^2 V}\int\left[\widetilde{\|f_0\|}_\infty^2\mathrm{Tr}(A_u^2) + \|b_u\|_2^2\right]d\mu(u)\right) \\
&+\left(1+\frac{1}{\nu}\right)\gamma\int\|\hat{f}_u - f_0\|_2^2 d\mu(u) + \beta\left(\mathrm{KL}(\rho,\pi) + \mathrm{KL}(\mu,\pi) + \ln\frac{1}{\eta}\right)\Big] \geq 1-\eta.
\end{aligned}$$

The result in expectation comes from Equation (6.1) with $\eta = 1$, combined with the inequality $t \leq \exp(t) - 1$. $\square$

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[2] P. Alquier and K. Lounici. PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electron. J. Stat.*, 5:127–145, 2011. ISSN 1935-7524. doi: 10.1214/11-EJS601. URL `http://dx.doi.org/10.1214/11-EJS601`.

[3] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Comput.*, 9(7):1545–1588, October 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.7.1545. URL `http://dx.doi.org/10.1162/neco.1997.9.7.1545`.

[4] P. C. Bellec. Optimal bounds for aggregation of affine estimators. *ArXiv e-prints*, October 2014.

[5] A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011. ISSN 0006-3444. doi: 10.1093/biomet/asr043. URL `http://dx.doi.org/10.1093/biomet/asr043`.

[6] G. Biau. Analysis of a random forests model. *J. Mach. Learn. Res.*, 13:1063–1095, 2012. ISSN 1532-4435.

[7] G. Biau and L. Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *J. Multivariate Anal.*, 101(10):2499–2518, 2010. ISSN 0047-259X. doi: 10.1016/j.jmva.2010.06.019. URL `http://dx.doi.org/10.1016/j.jmva.2010.06.019`.

[8] G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.*, 9:2015–2033, 2008. ISSN 1532-4435.

[9] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. ISSN 0885-6125. doi: 10.1023/A:1018054314350.

[10] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324.

[11] O. Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004. ISBN 3-540-22572-2. doi: 10.1007/b99352. URL `http://dx.doi.org/10.1007/b99352`. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001.

[12] O. Catoni. *Pac-Bayesian supervised classification: the thermodynamics of statistical learning*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 56. Institute of Mathematical Statistics, Beachwood, OH, 2007. ISBN 978-0-940600-72-0; 0-940600-72-2.

[13] N. Cesa-Bianchi and G. Lugosi. On prediction of individual sequences. *Ann. Statist.*, 27(6):1865–1895, 1999. ISSN 0090-5364. doi: 10.1214/aos/1017939242. URL `http://dx.doi.org/10.1214/aos/1017939242`.

[14] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *J. ACM*, 44(3):427–485, May 1997. ISSN 0004-5411. doi: 10.1145/258128.258179. URL `http://doi.acm.org/10.1145/258128.258179`.

[15] D. Dai, P. Rigollet, and T. Zhang. Deviation optimal learning using greedy Q-aggregation. *Ann. Statist.*, 40(3):1878–1905, 2012. ISSN 0090-5364. doi: 10.1214/12-AOS1025. URL `http://dx.doi.org/10.1214/12-AOS1025`.

[16] D. Dai, P. Rigollet, L. Xia, and T. Zhang. Aggregation of affine estimators. *Electron. J. Stat.*, 8:302–327, 2014. ISSN 1935-7524. doi: 10.1214/14-EJS886. URL `http://dx.doi.org/10.1214/14-EJS886`.

[17] A. S. Dalalyan. SOCP based variance free Dantzig selector with application to robust estimation. *C. R. Math. Acad. Sci. Paris*, 350(15-16):785–788, 2012. ISSN 1631-073X. doi: 10.1016/j.crma.2012.09.016. URL `http://dx.doi.org/10.1016/j.crma.2012.09.016`.

[18] A. S. Dalalyan and J. Salmon. Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.*, 40(4):2327–2355, 2012. ISSN 0090-5364. doi: 10.1214/12-AOS1038. URL `http://dx.doi.org/10.1214/12-AOS1038`.

[19] A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In NaderH. Bshouty and Claudio Gentile, ed-

itors, *Learning Theory*, volume 4539 of *Lecture Notes in Computer Science*, pages 97–111. Springer Berlin Heidelberg, 2007.

[20] A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61, 2008. doi: 10.1007/s10994-008-5051-0.

[21] A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.*, 78(5):1423–1443, 2012. ISSN 0022-0000. doi: 10.1016/j.jcss.2011.12.023. URL `http://dx.doi.org/10.1016/j.jcss.2011.12.023`.

[22] A. S. Dalalyan, M. Hebiri, K. Meziani, and J. Salmon. Learning heteroscedastic models by convex programming under group sparsity. In *ICML*, 2013.

[23] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B*, 57(2):301–369, 1995. ISSN 0035-9246. With discussion and a reply by the authors.

[24] Y. Freund. Boosting a weak learning algorithm by majority. *Inform. and Comput.*, 121(2):256–285, 1995. ISSN 0890-5401. doi: 10.1006/inco.1995.1136. URL `http://dx.doi.org/10.1006/inco.1995.1136`.

[25] Theo Gasser and Hans-Georg Müller. Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.*, 11(3):171–185, 1984. ISSN 0303-6898.

[26] R. Genuer. *Forêts aléatoires : aspects théoriques, sélection de variables et applications*. PhD thesis, Université Paris-Sud, 2011.

[27] S. Gerchinovitz. *Prediction of individual sequences and prediction in the statistical framework : some links around sparse regression and aggregation techniques*. These, Université Paris Sud, December 2011. URL `https://tel.archives-ouvertes.fr/tel-00653550`.

[28] C. Giraud. Mixing least-squares estimators when the variance is unknown. *Bernoulli*, 14(4):1089–1107, 2008. ISSN 1350-7265. doi: 10.3150/08-BEJ135. URL `http://dx.doi.org/10.3150/08-BEJ135`.

[29] C. Giraud, S. Huet, and N. Verzelen. High-dimensional regression with unknown variance. *Statist. Sci.*, 27(4):500–518, 2012. ISSN 0883-4237. doi: 10.1214/12-STS398. URL `http://dx.doi.org/10.1214/12-STS398`.

[30] B. Guedj and P. Alquier. PAC-Bayesian estimation and prediction in sparse additive models. *Electron. J. Stat.*, 7:264–291, 2013. ISSN 1935-7524. doi: 10.1214/13-EJS771. URL `http://dx.doi.org/10.1214/13-EJS771`.

[31] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. ISBN 978-0-387-84857-0. doi: 10.1007/978-0-387-84858-7. URL `http://dx.doi.org.revues.math.u-psud.fr:2048/10.1007/978-0-387-84858-7`. Data mining, inference, and prediction.

[32] D. Hsu, S. M. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.*, 17:no. 52, 6, 2012. ISSN 1083-589X. doi: 10.1214/ECP.v17-2079. URL `http://dx.doi.org/10.1214/ECP.v17-2079`.

[33] G. Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13(4):1000–1022, 2007. ISSN 1350-7265. doi: 10.3150/07-BEJ6044. URL `http://dx.doi.org/10.3150/07-BEJ6044`.

[34] G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006. ISSN 0018-9448. doi: 10.1109/TIT.2006.878172. URL `http://dx.doi.org/10.1109/TIT.2006.878172`.

[35] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212 – 261, 1994. ISSN 0890-5401. doi: http://dx.doi.org/10.1006/inco.1994.1009. URL `http://www.sciencedirect.com/science/article/pii/S0890540184710091`.

[36] K. Lounici. Generalized mirror averaging and D-convex aggregation. *Math. Methods Statist.*, 16(3):246–259, 2007. ISSN 1066-5307. doi: 10.3103/S1066530707030040. URL `http://dx.doi.org/10.3103/S1066530707030040`.

[37] É. Nadaraya. On non-parametric estimates of density functions and regression curves. *Theory of Probability & Its Applications*, 10(1):186–190, 1965. doi: 10.1137/1110024. URL `http://dx.doi.org/10.1137/1110024`.

[38] A. Nemirovski. Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*, volume 1738 of *Lecture Notes in Math.*, pages 85–277. Springer, Berlin, 2000.

[39] P. Rigollet. *Inégalités d'oracle, agrégration et adaptation*. PhD thesis, Université Pierre et Marie Curie- Paris VI, 2006.

[40] P. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Math. Methods Statist.*, 16(3):260–280, 2007. ISSN 1066-5307. doi: 10.3103/S1066530707030052. URL `http://dx.doi.org/10.3103/S1066530707030052`.

[41] P. Rigollet and A. B. Tsybakov. Sparse estimation by exponential weighting. *Statist. Sci.*, 27(4):558–575, 2012. ISSN 0883-4237. doi: 10.1214/12-STS393. URL `http://dx.doi.org/10.1214/12-STS393`.

[42] R. E. Schapire. The strength of weak learnability. *Mach. Learn.*, 5(2):197–227, July 1990. ISSN 0885-6125. doi: 10.1023/A:1022648800760. URL `http://dx.doi.org/10.1023/A:1022648800760`.

[43] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978. ISSN 0090-5364.

[44] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012. ISSN 0006-3444. doi: 10.1093/biomet/ass043. URL `http://dx.doi.org/10.1093/biomet/ass043`.

[45] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[46] A. B. Tsybakov. Optimal rates of aggregation. In B. Schölkopf and M. K. Warmuth, editors, *Learning Theory and Kernel Machines*, volume 2777 of *Lecture Notes in Computer Science*, pages 303–313. Springer Berlin Heidelberg, 2003.

[47] A. B. Tsybakov. Agrégation d'estimateurs et optimisation stochastique. *J. Soc. Fr. Stat. & Rev. Stat. Appl.*, 149(1):3–26, 2008. ISSN 1962-5197.

[48] V. G. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, COLT '90, pages 371–386, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc. ISBN 1-55860-146-5. URL `http://dl.acm.org/citation.cfm?id=92571.92672`.

[49] G. Wahba. *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial

and Applied Mathematics (SIAM), Philadelphia, PA, 1990. ISBN 0-89871-244-0. doi: 10.1137/1.9781611970128. URL http://dx.doi.org/10.1137/1.9781611970128.

[50] G. S. Watson. Smooth regression analysis. *Sankhya Ser. A*, 26:359–372, 1964. ISSN 0581-572X.

[51] Y. Yang. Mixing strategies for density estimation. *Ann. Statist.*, 28(1):75–87, 2000. ISSN 0090-5364. doi: 10.1214/aos/1016120365. URL http://dx.doi.org/10.1214/aos/1016120365.

[52] Y. Yang. Combining different procedures for adaptive regression. *J. Multivariate Anal.*, 74(1):135–161, 2000. ISSN 0047-259X. doi: 10.1006/jmva.1999.1884. URL http://dx.doi.org/10.1006/jmva.1999.1884.

[53] Y. Yang. Adaptive estimation in pattern recognition by combining different procedures. *Statist. Sinica*, 10(4):1069–1089, 2000. ISSN 1017-0405.

[54] Y. Yang. Adaptive regression by mixing. *J. Amer. Statist. Assoc.*, 96(454):574–588, 2001. ISSN 0162-1459. doi: 10.1198/016214501753168262. URL http://dx.doi.org/10.1198/016214501753168262.

[55] Y. Yang. Regression with multiple candidate models: selecting or mixing? *Statist. Sinica*, 13(3):783–809, 2003. ISSN 1017-0405.

[56] Y. Yang. Combining forecasting procedures: some theoretical results. *Econometric Theory*, 20(1):176–222, 2004. ISSN 0266-4666. doi: 10.1017/S0266466604201086. URL http://dx.doi.org/10.1017/S0266466604201086.

[57] Y. Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10(1):25–47, 2004. ISSN 1350-7265. doi: 10.3150/bj/1077544602. URL http://dx.doi.org/10.3150/bj/1077544602.

[58] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00503.x. URL http://dx.doi.org.revues.math.u-psud.fr:2048/10.1111/j.1467-9868.2005.00503.x.