PAC-Bayesian aggregation of linear estimators

Lucie Montuelle, Erwan Le Pennec

PAC-Bayesian aggregation of linear estimators

L. Montuelle (Université Paris Sud) and E. Le Pennec (École polytechnique)

October 2, 2014

Abstract : We consider the aggregation of linear estimator in regresison with a sub-Gaussian noise assumption. Aggregating estimators using exponential weights depending on their risk performs well in expectation, but sadly not in probability. A way to overcome this issue is considering exponential weights of a penalized risk. In this case, an oracle inequality can be obtained in probability, but is not sharp. Taking into account the estimated function's norm in the penalty offers a sharp inequality.

Keywords : Exponentially weighted aggregation, Regression, Oracle inequality

1 Introduction

We consider here a classical fixed design regression model

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

with f_0 an unknown function, x_i the fixed design points and W a centered sub-Gaussian noise. Our aim is to estimate the function f_0 at the grid points. We study a strategy in which a collection of smoothed projection $\{\hat{f}_t(Y) = P_t Y | P_t \in \mathcal{S}_n^+(\mathbb{R}), t \in \mathcal{T}\}$ is aggregated into a single adaptive estimator using a PAC-Bayesian aggregation.

Aggregation procedures have been introduced by Vovk [1990], Littlestone and Warmuth [1994], Cesa-Bianchi et al. [1997], Cesa-Bianchi and Lugosi [1999]. They are a core ingredient of bagging [Breiman, 1996], boosting [Freund, 1995, Schapire, 1990] or random forest (Amit and Geman [1997] or Breiman [2001]; or more recently Biau et al. [2008], Biau and Devroye [2010], Biau [2012], Genuer [2011]).

The general aggregation framework is detailed in Nemirovski [2000] and studied in Catoni [2004, 2007] through a PAC-Baysian framework as well as in Yang [2000c,b,a, 2001, 2003, 2004a,b]. See for instance Tsybakov [2008] for a survey. Optimal rates of aggregation in regression and density estimation are studied by Tsybakov [2003], Lounici [2007], Rigollet and Tsybakov [2007], Rigollet [2006] and Lecué [2007]

We follow the exponentially weighted aggregation strategy, in which the weight of each element 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 $\mathcal{T}$. Our aim is to give sufficient conditions on the penalized risk estimate and the temperature to obtain an oracle inequality for the risk of our estimate.

This scheme here has been used first by Leung and Barron [2006] that have obtained the first exact oracle inequality in this setting by aggregation projection with exponential weights in a Gaussian regression framework. Those results have been extended to several setting Dalalyan and Tsybakov [2007, 2008, 2012], Giraud [2008], Dalalyan et al. [2013], Belloni et al. [2011], Dalalyan [2012], Giraud et al. [2012], Dalalyan and Salmon [2012], Sun and Zhang [2012], Rigollet and Tsybakov [2012] under a *frozen* estimator assumption: they should not depend on the observed sample. This restriction, not present in the work by Leung and Barron [2006], has been removed by Dalalyan and Salmon [2012] within the context of affine estimator and exponentially weighted aggregation. However, Dai et al. [2012] have shown the sub-optimality in deviation of exponential weighting, not allowing to obtain a sharp oracle inequality in probability. Nevertheless, 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 in probability [Dai et al., 2014]. Furthermore, the corresponding oracle inequality is not sharp.

Our contribution is twofold. First, we propose the first extension to general sub-Gaussian noise. Second, we conduct a fine analysis of the relationship between the choice of the penalty and the temperature. In particular, we are able to take into account the signal to noise ratio to provide sharp oracle inequalities for bounded functions.

Not that our results are similar to the one obtained for a slightly different aggregation scheme by Bellec [2014] in a preprint written while the authors were working independently on this 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 main assumption on the noise is that $W \in \mathbb{R}^n$ is a centered sub-Gaussian variable, i.e. $\mathbb{E}(W) = 0$ 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 point 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.

Assume that we have at hand a collection of data dependent smoothed projection estimates

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

where $(b_{t,i})_{i=1}^n$ is an orthonormal basis and $(\rho_{t,i})_{i=1}^n$ a sequence of non-negative real numbers. For such an estimate, it exists a symmetric positive semi-definite real matrix of size n, P_t such that $\hat{f}_t(Y) = P_t Y$. For the sake of simplicity, we use this representation of our estimators. Note that we depart from the affine estimator studied by Dalalyan and Salmon [2012] and Dai et al. [2014] because we consider only linear estimate. This choice was made to simplify our exposition but similar results as the one we obtain for linear estimates hold for affine ones.

To define our estimate from the collection $\{\hat{f}_t(Y) = P_t Y | P_t \in \mathcal{S}_n^+(\mathbb{R}), t \in \mathcal{T}\}$, we specify the estimate $\widetilde{r_t}$ of the risk of the estimator $\hat{f}_t(Y)$, choose a prior probability measure π over $\mathcal{T}$ and a temperature $\beta > 0$. We define f_{EWA} by $f_{EWA} = \int \hat{f}_t \, d\rho(t)$, with

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

a probability measure over $\mathcal{T}$. The intuition behind this construction to favor low risk estimates. When the temperature goes to 0 this estimator becomes very similar to the one minimizing the risk estimates 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

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

to which a penalty $\mathrm{pen}(t)$ is added. We will consider simultaneously the case of a penalty that depends on f_0 through an upper bound of a kind of sup norm and the case of a penalty that does not depend on f_0.

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 $t \in \mathcal{T}$,

$$\|P_t f_0\|_2^2 \leq C^2 tr(P_t^2).$$

By construction, $\widetilde{\|f_0\|_\infty}$ is indeed smaller than the sup norm of any coefficients of f_0 in any basis appearing in the $\mathcal{T}$. 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 pen(t) and the temperature β so that an oracle inequality of type

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

holds, with ϵ and ϵ' small non-negative numbers possibly equal to 0 and price(t) a loss depending on the choice of pen(t) and β. 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 up to some controlled cost.

3 A general oracle inequality

Our main result is the following:

Theorem 1. Assume W is a centered sub-Gaussian noise with parameter σ^2. Assume $\{\hat{f}_t(Y) = P_t Y | P_t \in \mathcal{S}_n^+(\mathbb{R}), t \in \mathcal{T}\}$ are such that there exists $V > 0$ satisfying $\sup_{t \in \mathcal{T}} \|P_t\|_2 \leq V$.

Let π be a arbitrary prior measure on $\mathcal{T}$, $\beta > 4\sigma^2 V$ an arbitrary temperature and pen(t) a penalty so that $f_{EWA} = \int \hat{f}_t \, d\rho(t)$ with

$$d\rho(t) = \frac{\exp\left(-\frac{1}{\beta} \left[r_t + \text{pen}(t) \right] \right)}{\int \exp\left(-\frac{1}{\beta} \left[r_{t'} + \text{pen}(t') \right] \right) d\pi(t')} d\pi(t)$$

For any $\delta \in [0,1]$, if $\beta \geq 4\sigma^2 V(1+4\delta)$, 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 $t \in \mathcal{T}$,

$$\text{pen}(t) \geq \frac{4\sigma^2}{\beta - 4\sigma^2 V} \left(1 + (1-\delta)(1+2\gamma V)^2 \frac{\widetilde{\|f_0\|_\infty^2}}{\sigma^2} \right) tr(P_t^2) \sigma^2,$$

then

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

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

- Furthermore

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

with

$$\text{price}(t) = 2\sigma^2 \left(tr(P_t) + \frac{2\sigma^2(1-\delta)(1+2\gamma V)^2}{\beta - 4\sigma^2 V} \frac{\widetilde{\|f_0\|_\infty^2}}{\sigma^2} tr(P_t^2) \right)$$

$$\epsilon'(\nu) = \frac{1}{1-(1+\nu)\gamma} - 1$$

$$\epsilon(\nu) = \frac{(1+\nu)^2\gamma}{\nu(1-(1+\nu)\gamma)} = \frac{(1+\nu)^2}{\nu}\gamma(1+\epsilon'(\nu))$$

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

This theorem is similar to the one obtained by Dai et al. [2014]. It yields a sufficient condition on the penalty for oracle inequalities to hold both in probability and in expectation. It holds however under a milder sub-Gaussianity assumption on the noise and allows to take into account a sup norm information in the penalty as used for instance in Guedj and Alquier [2013] . Note that the result in expectation requires a penalty that is not necessary, at least in the Gaussian case, as shown by Dalalyan and Salmon [2012].

If we are authorized to use the upper bound of the sup norm, we may ensure that the penalty satisfies the lower bound condition with $\delta = 0$. In that case, $\gamma = 0$ and $\epsilon'(\nu) = 0, \epsilon(\nu) = 0$. Thus, there is no need to optimize ν and it suffices to notice that $N = (0, +\infty)$ to obtain that

Corollary 1. Under the assumptions of Theorem 1, if $\beta > 4\sigma^2 V$, if for any $t \in \mathcal{T}$,

$$\text{pen}(t) \geq \frac{4\sigma^4}{\beta - 4\sigma^2 V} \left(1 + \frac{\widetilde{\|f_0\|_\infty^2}}{\sigma^2} \right) tr(P_t^2),$$

then

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

$$\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\mu \in \mathcal{M}_+^1(\mathcal{T})} \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t) + \int (\text{pen}(t) + \text{price}(t))\, d\mu(t) + \beta \left(2KL(\mu, \pi) + \ln \frac{1}{\eta} \right)$$

- Furthermore

$$\mathbb{E}\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\mu \in \mathcal{M}_+^1(\mathcal{T})} \int \mathbb{E}\|f_0 - \hat{f}_t\|_2^2 d\mu(t) + \int \left(\text{pen}(t) + \text{price}(t)\right) d\mu(t) + 2\beta KL(\mu, \pi)$$

with

$$\text{price}(t) = 2\sigma^2 \left(tr(P_t) + \frac{2}{\beta - 4\sigma^2 V} \widetilde{\|f_0\|_\infty^2} tr(P_t^2) \right).$$

As soon as $\delta > 0$, a simple calculation yields that for any $\beta \geq 4\sigma^2 V(1+4\delta)$, $0 < 2\gamma V \leq 1$. As a result, if $V > 0.5$, $(0, 2V-1) \subseteq N$. Furthermore if $\beta > 4\sigma^2 V(1+4\delta)$ and $V > 0.5$, then $2V - 1 \in N$. Else, if $\delta > 0$ and $0 < V \leq 0.5$, N is non-empty if and only if $\beta > 4\sigma^2 V + 2\sigma^2\delta(1+2V)^2$, which is a stronger condition than $\beta \geq 4\sigma^2 V(1+4\delta)$. Since V is an upper bound, it may be chosen greater than 0.5.

If $\delta = 1$, we obtain weak oracle inequalities that do not require the use of any side information:

Corollary 2. Under the assumptions of Theorem 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 $t \in \mathcal{T}$,

$$\text{pen}(t) \geq \frac{4\sigma^4}{\beta - 4\sigma^2 V} tr(P_t^2),$$

then

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

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

- Furthermore

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

with

$$\begin{aligned}\text{price}(t) &= 2\sigma^2 tr(P_t)\\ \epsilon'(\nu) &= \frac{1}{1-(1+\nu)\gamma} - 1\\ \epsilon(\nu) &= \frac{(1+\nu)^2\gamma}{\nu(1-(1+\nu)\gamma)} = \frac{(1+\nu)^2}{\nu}\gamma(1+\epsilon'(\nu))\end{aligned}$$

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

Note that the parameter γ allows us to obtain a weak oracle inequality. It links $\|(P_t - P_u)f_0\|_2^2$ to $\|(P_t - P_u)Y\|_2^2$.

Finally, assume that we let

$$\text{pen}(t) \geq \kappa tr(P_t^2)\sigma^2.$$

The previous corollary implies that a weak oracle inequality holds for any temperature greater than $20\sigma^2 V$ as soon as $\kappa \geq \frac{4\sigma^2}{\beta - 4\sigma^2 V}$. Corollary 1 implies that 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 1 shows that there is a continuum between those two cases as weak oracle inequalities, with smaller leading constant than the one of Corollary 2, 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 $4\sigma^2 V$.

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 2) 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 $\|f_0 - P_t f\|_2^2$ by $\widetilde{\|f_0\|^2}$ on the one hand, to obtain an exact oracle inequality, and by $\|f_0 - P_t Y\|_2^2$ 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 $\|f_0 - P_t f\|_2^2$ is no longer always smaller than $\|f_0 - P_t Y\|_2^2$ and providing a fine tuning optimization allowing the equality in the constraint on β and an optimization on the parameters ϵ.

4 Proof of the oracle inequalities

Theorem 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 [2007]:

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

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

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

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

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

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

Equivalently,

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

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:

$$\begin{aligned}\int \left(\|f_0 - \hat{f}_t\|_2^2 - r_t\right) d\rho(t) - \int \left(\|f_0 - \hat{f}_u\|_2^2 - r_u\right) d\mu(u) \\ \leq C_1 \int \|f_0 - \hat{f}_t\|_2^2 d\rho(t) + C_2 \int \|f_0 - \hat{f}_u\|_2^2 d\mu(u) \\ + C_3 \int tr(P_t^2) d\rho(t) + C_4 \int tr(P_u) d\mu(u) + C_5 \int tr(P_u^2) d\mu(u) \\ + \beta \left(KL(\rho, \pi) + KL(\mu, \pi) + \ln \frac{1}{\eta} \right)\end{aligned}$$

where C_1 to C_6 are known functions. Combining with the previous inequality and taking $\text{pen}(t) \geq C_3\, tr(P_t^2)$ gives

$$\begin{aligned}(1 - C_1) \int \|f_0 - \hat{f}_t\|_2^2 d\rho(t) - (1 + C_2) \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t) \\ \leq C_4 \int tr(P_u) d\mu(u) + C_5 \int tr(P_u^2) d\mu(u) + \int \text{pen}(t) d\mu(t) \\ + \beta \left(2KL(\mu, \pi) + \ln \frac{1}{\eta} \right).\end{aligned}$$

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 The expository case of Gaussian noise and projection estimates

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

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

$$\text{pen}(t) \geq \frac{2\sigma^4}{\beta - 4\sigma^2} \left(1 + 2(1 - \delta) \frac{\widetilde{\|f_0\|_\infty^2}}{\sigma^2} \right) tr(P_t)$$

satisfies

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

$$\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\mu \in \mathcal{M}_+^1(\mathcal{T})} (1+2\epsilon) \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t)$$
$$+ (1+\epsilon) \int \left(\text{pen}(t) + \text{price}(t)\right) d\mu(t) + \beta(1+\epsilon) \left(2KL(\mu,\pi) + \ln \frac{1}{\eta}\right).$$

- Furthermore,

$$\mathbb{E}\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\mu \in \mathcal{M}_+^1(\mathcal{T})} (1+2\epsilon) \int \mathbb{E}\|f_0 - \hat{f}_t\|_2^2 d\mu(t)$$
$$+ (1+\epsilon) \int \left(\text{pen}(t) + \text{price}(t)\right) d\mu(t) + 2\beta(1+\epsilon) 2KL(\mu,\pi),$$

with

$$\text{price}(t) = 2\left(1 + \frac{2(1-\delta)\sigma^2}{\beta - 4\sigma^2} \frac{\widetilde{\|f_0\|_\infty^2}}{\sigma^2}\right) tr(P_t)\sigma^2 \quad \text{and} \quad \epsilon = \frac{4\sigma^2\delta}{\beta - 4\sigma^2(\delta+1)}.$$

Note that $\text{pen}(t) \geq \text{price}(t) + 2\sigma^2 \left(\frac{\sigma^2}{\beta - 4\sigma^2} - 1\right) tr(P_t)$, and the result may be further simplified using $\text{pen}(t) + \text{price}(t) \leq 2\left(\text{pen}(t) + \sigma^2 tr(P_t)\right)$.

As announced in the scheme of proof of the oracle inequalities (section 4), 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 $1-\eta$,

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

- Moreover,

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

The use of this lemma is detailed in section 5.2. We focus now on its proof 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 $\|f_0 - P_t f_0\|_2^2$ and $\|f_0 - P_t Y\|_2^2$. 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).

5.1 Proof of Lemma 2

Démonstration. For the sake of clarity, for any $t, u \in \mathcal{T}$, 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

$$\Delta_{t,u} = 2\left(W^\top (P_t - P_u) W + W^\top (P_t - P_u) f_0 - \sigma^2 tr(P_t - P_u)\right).$$

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

For any $\beta > 0$,

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

Following lemma 2.4 of Hsu et al. [2012], if $\beta > 4\sigma^2$,

$$\mathbb{E}\left[\exp\frac{\Delta_{t,u}}{\beta}\right] \leq \exp\frac{2\sigma^2}{\beta}\left(tr(P_u) + \frac{2\sigma^2 tr(P_t) + \|(P_t - P_u) f_0\|_2^2}{\beta - 4\sigma^2}\right). \quad (1)$$

Note that

$$\|(P_t - P_u) f_0\|_2^2 \leq 2\left(\|f_0 - P_t f_0\|_2^2 + \|f_0 - P_u f_0\|_2^2\right) \leq 2\left(\|f_0 - P_t Y\|_2^2 + \|f_0 - P_u Y\|_2^2\right)$$

and

$$\|(P_t - P_u)f_0\|_2^2 \leq 2\left(\|P_t f_0\|_2^2 + \|P_u f_0\|_2^2\right) \leq 2\widetilde{\|f_0\|_\infty^2}\left(tr(P_t) + tr(P_u)\right).$$

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

$$\mathbb{E}\exp\left[\frac{\Delta_{t,u}}{\beta} - \frac{2\sigma^2}{\beta}\left(tr(P_u) + \frac{2\sigma^2 tr(P_t)}{\beta - 4\sigma^2}\right) - \frac{4\sigma^2\delta}{\beta(\beta-4\sigma^2)}\left(\|f_0 - \hat{f}_t\|_2^2 + \|f_0 - \hat{f}_u\|_2^2\right) - \frac{4\sigma^2}{\beta(\beta-4\sigma^2)}(1-\delta)\widetilde{\|f_0\|_\infty^2}\left(tr(P_t) + tr(P_u)\right)\right] \leq 1.$$

Along the same lines as Alquier and Lounici [2011], we first integrate according to the prior π and use Fubini's theorem,

$$\mathbb{E}\int\int \exp\frac{1}{\beta}\left[\Delta_{t,u} - 2\sigma^2\left(tr(P_u) + \frac{2\sigma^2 tr(P_t)}{\beta - 4\sigma^2}\right) - \frac{4\sigma^2\delta}{\beta - 4\sigma^2}\left(\|f_0 - \hat{f}_t\|_2^2 + \|f_0 - \hat{f}_u\|_2^2\right) - \frac{4\sigma^2}{\beta - 4\sigma^2}(1-\delta)\widetilde{\|f_0\|_\infty^2}\left(tr(P_t) + tr(P_u)\right)\right] d\pi(t)d\pi(u) \leq 1,$$

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

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

before applying Jensen's inequality

$$\mathbb{E}\exp\frac{1}{\beta}\left[\int\int \Delta_{t,u}d\rho(t)d\mu(u) - \frac{4\delta\sigma^2}{\beta - 4\sigma^2}\left(\int \|f_0 - \hat{f}_t\|_2^2 d\rho(t) + \int \|f_0 - \hat{f}_u\|_2^2 d\mu(u)\right) - \frac{4\sigma^2}{\beta - 4\sigma^2}\left(\sigma^2 + (1-\delta)\widetilde{\|f_0\|_\infty^2}\right)\int tr(P_t)d\rho(t) - \beta\left(KL(\rho,\pi) + KL(\mu,\pi) + \ln\frac{1}{\eta}\right) - 2\sigma^2\left(1 + \frac{2(1-\delta)\widetilde{\|f_0\|_\infty^2}}{\beta - 4\sigma^2}\right)\int tr(P_u)d\mu(u)\right] \leq \eta. \quad (2)$$

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

$$\mathbb{P}\left[\int\int \Delta_{t,u}d\rho(t)d\mu(u) \leq \frac{4\delta\sigma^2}{\beta - 4\sigma^2}\left(\int \|f_0 - \hat{f}_t\|_2^2 d\rho(t) + \int \|f_0 - \hat{f}_u\|_2^2 d\mu(u)\right) + \frac{4\sigma^2}{\beta - 4\sigma^2}\left(\sigma^2 + (1-\delta)\widetilde{\|f_0\|_\infty^2}\right)\int tr(P_t)d\rho(t) + \beta\left(KL(\rho,\pi) + KL(\mu,\pi) + \ln\frac{1}{\eta}\right) + \frac{2\sigma^2}{n}\left(1 + \frac{2(1-\delta)n\widetilde{\|f_0\|_\infty^2}}{\beta n - 4\sigma^2}\right)\int tr(P_u)d\mu(u)\right] \geq 1 - \eta.$$

The result in expectation is obtained by Equation (2) with $\eta = 1$:

$$\mathbb{E}\exp\frac{1}{\beta}\left[\int\int \Delta_{t,u}d\rho(t)d\mu(u) - \frac{4\delta\sigma^2}{\beta-4\sigma^2}\left(\int \|f_0-\hat{f}_t\|_2^2 d\rho(t) + \int \|f_0-\hat{f}_u\|_2^2 d\mu(u)\right)\right.$$
$$-\frac{4\sigma^2}{\beta-4\sigma^2}\left(\sigma^2+(1-\delta)\widetilde{\|f_0\|_\infty^2}\right)\int tr(P_t)d\rho(t) - \beta\left(KL(\rho,\pi)+KL(\mu,\pi)\right)$$
$$\left.-2\sigma^2\left(1+\frac{2(1-\delta)\widetilde{\|f_0\|_\infty^2}}{\beta-4\sigma^2}\right)\int tr(P_u)d\mu(u)\right] \leq 1,$$

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

5.2 Proof of Theorem 2

We follow the scheme of proof given in section 4 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,

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

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

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

Furthermore, using

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

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

$$\|f_0 - f_{EWA}\|_2^2 \leq \inf_{\mu\in\mathcal{M}_+^1(\mathcal{T})} \left(1 + \frac{8\sigma^2\delta}{\beta - 4\sigma^2(\delta+1)}\right) \int \|f_0 - \hat{f}_t\|_2^2 d\mu(t)$$
$$+ \left(1 + \frac{4\sigma^2\delta}{\beta - 4\sigma^2(\delta+1)}\right) 2\sigma^2 \left(1 + \frac{2(1-\delta)\widetilde{\|f_0\|_\infty^2}}{\beta - 4\sigma^2}\right) \int tr(P_t) d\mu(t)$$
$$+ \left(1 + \frac{4\sigma^2\delta}{\beta - 4\sigma^2(\delta+1)}\right) \int \mathrm{pen}(t) d\mu(t) + \beta \left(1 + \frac{4\sigma^2\delta}{\beta - 4\sigma^2(\delta+1)}\right) \left(2KL(\mu,\pi) + \ln\frac{1}{\eta}\right).$$

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

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

6 Appendix : Proofs in the sub-Gaussian case

6.1 Proof of Theorem 1

The proof follows from the scheme described in section 4. 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 P_t is a symmetric positive semi-definite matrix, there exists $V > 0$ such that $\sup_{t\in\mathcal{T}} \|P_t\|_2 \leq V$ and W is a centered sub-Gaussian noise. For any $t, u \in \mathcal{T}$, 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 $1-\eta$,

$$\int\int \Delta_{t,u} d\rho(t)d\mu(u) \leq (1+\nu)\gamma \int \|P_t Y - f_0\|_2^2 d\rho(t)$$
$$+ \frac{4\sigma^2}{\beta - 4\sigma^2 V}\left(\sigma^2 + (1-\delta)(1+2\gamma V)^2 \widetilde{\|f_0\|_\infty^2}\right)\int tr(P_t^2)d\rho(t)$$
$$+ 2\sigma^2\left(\int tr(P_u)d\mu(u) + \frac{2(1-\delta)(1+2\gamma V)^2}{\beta - 4\sigma^2 V}\widetilde{\|f_0\|_\infty^2}\int tr(P_u^2)d\mu(u)\right)$$
$$+ \left(1+\frac{1}{\nu}\right)\gamma\int \|P_u Y - f_0\|_2^2 d\mu(u) + \beta\left(KL(\rho,\pi) + KL(\mu,\pi) + \ln\frac{1}{\eta}\right)$$

- Moreover,

$$\mathbb{E}\left[\int\int \Delta_{t,u} d\rho(t)d\mu(u)\right] \leq \mathbb{E}\Big[(1+\nu)\gamma \int \|P_t Y - f_0\|_2^2 d\rho(t)$$
$$+ \frac{4\sigma^2}{\beta - 4\sigma^2 V}\left(\sigma^2 + (1-\delta)(1+2\gamma V)^2 \widetilde{\|f_0\|_\infty^2}\right)\int tr(P_t^2)d\rho(t)$$
$$+ 2\sigma^2\left(\int tr(P_u)d\mu(u) + \frac{2(1-\delta)(1+2\gamma V)^2}{\beta - 4\sigma^2 V}\widetilde{\|f_0\|_\infty^2}\int tr(P_u^2)d\mu(u)\right)$$
$$+ \left(1+\frac{1}{\nu}\right)\gamma\int \|P_u Y - f_0\|_2^2 d\mu(u) + \beta\left(KL(\rho,\pi) + KL(\mu,\pi)\right)\Big]$$

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

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

Taking $\mathrm{pen}(t) \geq \frac{4\sigma^2}{\beta - 4\sigma^2 V}\left(\sigma^2 + (1-\delta)(1+2\gamma V)^2 \widetilde{\|f_0\|_\infty^2}\right) tr(P_t^2)$ and $\nu \in N = \{\nu > 0 | (1+\nu)\gamma < 1\}$, such that the inequality stays informative,

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

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

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

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 $\|P_t f_0 - f_0\|_2^2$ (see Equation (1)). Since P_t are not projections, $\|P_t f_0 - f_0\|_2^2 \leq \|P_t Y - f_0\|_2^2$ does not hold any more. The presence of $\|P_t Y - f_0\|_2^2$ allows us to obtain a weak oracle inequality. To overcome this difficulty, $\|(P_t - P_u)Y\|_2^2$ is introduced and for an arbitrary $\gamma \geq 0$, we try to control $\Delta_{t,u} - \gamma\|(P_t - P_u)Y\|_2^2$.

Démonstration. A simple calculation yields

$$\Delta_{t,u} - \gamma\|(P_t - P_u)Y\|_2^2 = W^\top (2I - \gamma(P_t - P_u)^\top)(P_t - P_u)W$$
$$+ 2W^\top (I - \gamma(P_t - P_u)^\top)(P_t - P_u)f_0 - 2\sigma^2 tr(P_t - P_u) - \gamma\|(P_t - P_u)f_0\|_2^2.$$

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

$$\mathbb{E}\exp\left(\frac{\Delta_{t,u}}{\beta} - \frac{\gamma}{\beta}\|(P_t - P_u)Y\|_2^2\right)$$
$$\leq \mathbb{E}\left[\exp\frac{2}{\beta}\left(W^\top P_t W + W^\top(I - \gamma(P_t - P_u))(P_t - P_u)f_0\right)\right]$$
$$\times \exp\frac{-1}{\beta}\left(2\sigma^2 tr(P_t - P_u) + \gamma\|(P_t - P_u)f_0\|_2^2\right).$$

The first step is to bring us back to the Gaussian case, using W's sub-Gaussianity and an idea of Hsu et al. [2012]. Let Z be a standard Gaussian random variable, independent of W. Then,

$$\mathbb{E}\exp\left(\frac{2}{\sqrt{\beta}}W^\top\sqrt{P_t}Z + \frac{2}{\beta}W^\top(I - \gamma(P_t - P_u))(P_t - P_u)f_0\right)$$
$$= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{2}{\sqrt{\beta}}W^\top\sqrt{P_t}Z + \frac{2}{\beta}W^\top(I - \gamma(P_t - P_u))(P_t - P_u)f_0\right)\big|W\right]\right]$$
$$= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{2}{\sqrt{\beta}}W^\top\sqrt{P_t}Z\right)\big|W\right]\exp\left(\frac{2}{\beta}W^\top(I - \gamma(P_t - P_u))(P_t - P_u)f_0\right)\right]$$
$$= \mathbb{E}\exp\frac{2}{\beta}\left(W^\top P_t W + W^\top(I - \gamma(P_t - P_u))(P_t - P_u)f_0\right).$$

On the other hand,

$$\mathbb{E}\left[\exp\frac{2}{\beta}\left(W^\top P_t W + W^\top(I-\gamma(P_t-P_u))(P_t-P_u)f_0\right)\right]$$
$$= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{2}{\sqrt{\beta}}W^\top\sqrt{P_t}Z + \frac{2}{\beta}W^\top(I-\gamma(P_t-P_u))(P_t-P_u)f_0\right)|Z\right]\right].$$

Since W is sub-Gaussian with parameter σ,

$$\mathbb{E}\left[\exp\frac{2}{\beta}\left(W^\top P_t W + W^\top(I-\gamma(P_t-P_u))(P_t-P_u)f_0\right)\right]$$
$$\leq \mathbb{E}\exp\left(\frac{\sigma^2}{2}\left\|\frac{2}{\sqrt{\beta}}\left(\sqrt{P_t}Z + \frac{1}{\sqrt{\beta}}(I-\gamma(P_t-P_u))(P_t-P_u)f_0\right)\right\|_2^2\right)$$

Hence,

$$\mathbb{E}\exp\left(\frac{\Delta_{t,u}}{\beta} - \frac{\gamma}{\beta}\|(P_t-P_u)Y\|_2^2\right)$$
$$\leq \mathbb{E}\left[\exp\frac{2\sigma^2}{\beta}\left(Z^\top P_t Z + \frac{2}{\sqrt{\beta}}Z^\top\sqrt{P_t}(I-\gamma(P_t-P_u))(P_t-P_u)f_0\right)\right]$$
$$\times\exp\left(\frac{2\sigma^2}{\beta^2}\|(I-\gamma(P_t-P_u))(P_t-P_u)f_0\|_2^2 - \frac{2\sigma^2}{\beta}tr(P_t-P_u) - \frac{\gamma}{\beta}\|(P_t-P_u)f_0\|_2^2\right).$$

The expectation is similar to the one obtained in the Gaussian case: the exponential of some quadratic form. The same recipe is applied. Since P_t is positive semi-definite, there exist an orthogonal matrix U and a diagonal matrix D such that $P_t = U^\top DU$. 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. [2012]. Then, for any $\beta > 4\sigma^2 V$, any $\gamma \geq 0$,

$$\mathbb{E}\exp\left(\frac{\Delta_{t,u}}{\beta} - \frac{\gamma}{\beta}\|(P_t-P_u)Y\|_2^2\right)$$
$$\leq \exp\frac{2\sigma^2}{\beta}\left(tr(P_t) + \frac{2\sigma^2}{\beta(\beta-4\sigma^2V)}\left(\beta tr(P_t^2) + 2\left\|\sqrt{P_t}(I-\gamma(P_t-P_u))(P_t-P_u)f_0\right\|_2^2\right)\right)$$
$$\times\exp\left(\frac{2\sigma^2}{\beta^2}\|(I-\gamma(P_t-P_u))(P_t-P_u)f_0\|_2^2 - \frac{2\sigma^2}{\beta}tr(P_t-P_u) - \frac{\gamma}{\beta}\|(P_t-P_u)f_0\|_2^2\right).$$

Consequently,

$$\begin{aligned}
&\mathbb{E}\exp\left(\frac{\Delta_{t,u}}{\beta}+\frac{\gamma}{\beta}\left(\|(P_t-P_u)f_0\|_2^2-\|(P_t-P_u)Y\|_2^2\right)\right)\\
&\leq \exp\frac{2\sigma^2}{\beta}\left(tr(P_u)+\frac{2\sigma^2}{\beta-4\sigma^2V}tr(P_t^2)\right)\\
&\quad\times\exp\left(\frac{2\sigma^2}{\beta^2}\left(\frac{4\sigma^2V}{\beta-4\sigma^2V}(1+2\gamma V)^2+(1+2\gamma V)^2\right)\|(P_t-P_u)f_0\|_2^2\right).\\
&\leq \exp\frac{2\sigma^2}{\beta}\left(tr(P_u)+\frac{2\sigma^2}{\beta-4\sigma^2V}tr(P_t^2)+\frac{(1+2\gamma V)^2}{\beta-4\sigma^2V}\|(P_t-P_u)f_0\|_2^2\right).
\end{aligned}$$

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

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

  $$\|(P_t-P_u)f_0\|_2^2\leq 2\left(\|P_tf_0\|_2^2+\|P_uf_0\|_2^2\right)\leq 2\left(tr(P_t^2)+tr(P_u^2)\right)\widetilde{\|f_0\|_\infty^2}$$

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

  $$\begin{aligned}
  &\mathbb{E}\exp\left(\frac{\Delta_{t,u}}{\beta}+\frac{\gamma}{\beta}\left(\|(P_t-P_u)f_0\|_2^2-\|(P_t-P_u)Y\|_2^2\right)\right)\\
  &\leq \exp\frac{2\sigma^2}{\beta}\left(tr(P_u)+\frac{2\sigma^2}{\beta-4\sigma^2V}tr(P_t^2)+\frac{(1+2\gamma V)^2(1-\delta)}{\beta-4\sigma^2V}\|(P_t-P_u)f_0\|_2^2\right)\\
  &\times\exp\left(\frac{2\sigma^2(1+2\gamma V)^2\delta}{\beta(\beta-4\sigma^2V)}\|(P_t-P_u)f_0\|_n^2\right)\\
  &\leq \exp\frac{2\sigma^2}{\beta}\left(tr(P_u)+\frac{2\sigma^2}{\beta-4\sigma^2V}tr(P_t^2)+\frac{(1+2\gamma V)^2\delta}{\beta-4\sigma^2V}\|(P_t-P_u)f_0\|_2^2\right)\\
  &\times\exp\left(\frac{4\sigma^2(1+2\gamma V)^2(1-\delta)}{\beta(\beta-4\sigma^2V)}\left(tr(P_t^2)+tr(P_u^2)\right)\widetilde{\|f_0\|_\infty^2}\right).
  \end{aligned}$$

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

  $$\begin{aligned}
  \frac{2\sigma^2(1+2\gamma V)^2\delta}{\beta-4\sigma^2V}\|(P_t-P_u)f_0\|_2^2-\gamma\left(\|(P_t-P_u)f_0\|_2^2-\|(P_t-P_u)Y\|_2^2\right)&\\
  \leq C_1\|P_tY-f_0\|_2^2+C_2\|P_uY-f_0\|_2^2&
  \end{aligned}$$

  for some non-negative constants C_1 and C_2 and with $\delta>0$. Since for any $\nu>0$, $\|(P_t-P_u)Y\|_2^2\leq(1+\nu)\|P_tY-f_0\|_2^2+\left(1+\frac{1}{\nu}\right)\|P_uY-f_0\|_2^2$, it suffices that

  $$\frac{2\sigma^2(1+2\gamma V)^2\delta}{\beta-4\sigma^2V}\|(P_t-P_u)f_0\|_2^2-\gamma\|(P_t-P_u)f_0\|_2^2\leq 0.$$

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 $\nu > 0$,

$$\begin{aligned}
&\mathbb{E}\exp\left(\frac{\Delta_{t,u}}{\beta} - \frac{\gamma}{\beta}\left((1+\nu)\|P_t Y - f_0\|_2^2 + \left(1+\frac{1}{\nu}\right)\|P_u Y - f_0\|_2^2\right)\right) \\
&\leq \exp\frac{2\sigma^2}{\beta}\left(tr(P_u) + \frac{2\sigma^2}{\beta - 4\sigma^2 V}tr(P_t^2) + \frac{2(1+2\gamma V)^2(1-\delta)}{\beta - 4\sigma^2 V}\left(tr(P_t^2) + tr(P_u^2)\right)\widetilde{\|f_0\|_\infty^2}\right).
\end{aligned}$$

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 $\eta \in (0,1]$,

$$\begin{aligned}
\mathbb{E}\exp\frac{1}{\beta}\Big[\int\int \Delta_{t,u}d\rho(t)d\mu(u) &- (1+\nu)\gamma\int\|P_tY - f_0\|_2^2 d\rho(t) \\
&- \frac{4\sigma^2}{\beta - 4\sigma^2 V}\left(\sigma^2 + (1-\delta)(1+2\gamma V)^2\widetilde{\|f_0\|_\infty^2}\right)\int tr(P_t^2)d\rho(t) \\
&- 2\sigma^2\left(\int tr(P_u)d\mu(u) + \frac{2(1-\delta)(1+2\gamma V)^2}{\beta - 4\sigma^2 V}\widetilde{\|f_0\|_\infty^2}\int tr(P_u^2)d\mu(u)\right) \\
&- \left(1+\frac{1}{\nu}\right)\gamma\int\|P_uY - f_0\|_2^2 d\mu(u) \\
&- \beta\left(KL(\rho,\pi) + KL(\mu,\pi) + \ln\frac{1}{\eta}\right)\Big] \leq \eta. \quad (3)
\end{aligned}$$

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 $\nu > 0$,

$$\begin{aligned}
&\mathbb{P}\Big[\int\int \Delta_{t,u}d\rho(t)d\mu(u) \leq (1+\nu)\gamma\int\|P_tY - f_0\|_2^2 d\rho(t) \\
&+ \frac{4\sigma^2}{\beta - 4\sigma^2 V}\left(\sigma^2 + (1-\delta)(1+2\gamma V)^2\widetilde{\|f_0\|_\infty^2}\right)\int tr(P_t^2)d\rho(t) \\
&+ 2\sigma^2\left(\int tr(P_u)d\mu(u) + \frac{2(1-\delta)(1+2\gamma V)^2}{\beta - 4\sigma^2 V}\widetilde{\|f_0\|_\infty^2}\int tr(P_u^2)d\mu(u)\right) \\
&+ \left(1+\frac{1}{\nu}\right)\gamma\int\|P_uY - f_0\|_2^2 d\mu(u) + \beta\left(KL(\rho,\pi) + KL(\mu,\pi) + \ln\frac{1}{\eta}\right)\Big] \geq 1-\eta.
\end{aligned}$$

The result in expectation comes from Equation (3) with $\eta = 1$, combined with the inequality $t \leq \exp(t) - 1$. □

References

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`.

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`.

P. Bellec. Concentration of quadratic forms and aggregation of affine estimators. *arXiv*, 2014. 1410.0436v1.

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`.

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

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`.

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.

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

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

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.

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.

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.

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.

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.

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.

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.

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.

A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In NaderH. Bshouty and Claudio Gentile, editors, *Learning Theory*, volume 4539 of *Lecture Notes in Computer Science*, pages 97–111. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-72925-9. doi: 10.1007/978-3-540-72927-3_9. URL http://dx.doi.org/10.1007/978-3-540-72927-3_9.

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. URL http://certis.enpc.fr/~dalalyan/Download/Dal_Tsyb2008.pdf.

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.

A. S. Dalalyan, M. Hebiri, K. Meziani, and J. Salmon. Learning heteroscedastic models by convex programming under group sparsity. In *ICML*, 2013. URL papers/ICML13_DHMS.pdf.

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.

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

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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. ISBN 978-3-540-40720-1. doi: 10.1007/978-3-540-45167-9_23. URL http://dx.doi.org/10.1007/978-3-540-45167-9_23.

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

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.

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

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

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

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.

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

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

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