Prediction of quantiles by statistical learning and application to GDP forecasting

Pierre Alquier, Xiaoyin Li

JMLR: Workshop and Conference Proceedings vol (2012) 1–18 25th Annual Conference on Learning Theory

Prediction of quantiles by statistical learning and application to GDP forecasting

author names withheld

Editor: Under Review for COLT 2012

Abstract

In this paper, we tackle the problem of prediction and confidence intervals for time series using a statistical learning approach and quantile loss functions. In a first time, we show that the Gibbs estimator (also known as Exponentially Weighted aggregate) is able to predict as well as the best predictor in a given family for a wide set of loss functions. In particular, using the quantile loss function of Koenker and Bassett (1978), this allows to build confidence intervals. We apply these results to the problem of prediction and confidence regions for the French Gross Domestic Product (GDP) growth, with promising results.

Keywords: Statistical learning theory, time series prediction, quantile regression, GDP forecasting, PAC-Bayesian bounds, oracle inequalities, weak dependence, confidence intervals, business surveys.

1. Introduction

Motivated by economics problems, the prediction of time series is one of the most emblematic problem of statistics. Various methodologies are used that come from such various fields as parametric statistics, statistical learning, computer science or game theory.

In the parametric approach, one assumes that the time series is generated according to a parametric model, like ARMA or ARIMA processes, see e.g. Hamilton (1994); Brockwell and Davis (2009). Such an assumption is unrealistic in many applications. However, under this assumption, it is possible to estimate the parameters of the model, and to build confidence intervals on the prevision.

In the statistical learning point of view, one usually tries to avoid such restrictive parametric assumptions - see, e.g., Cesa-Bianchi and Lugosi (2006); Stoltz (2010) for the online approach dedicated to the prediction of individual sequences, and Modha and Masry (1998); Meir (2000); Alquier and Wintenberger (2012) for the batch approach. However, in this setting, a few attention has been paid to the construction of confidence intervals or to any quantification of the precision of the prediction. This is a major drawback in many applications.

In Biau and Patra (2011), a method was proposed for the online approach: the idea is to minimize the cumulated risk corresponding to the quantile loss function defined by Koenker and Bassett (1978). Some asymptotic results are provided.

In this paper, we propose to adapt this approach to the batch setting and provide nonasymptotic results. We also apply these results to build quarterly prediction and confidence regions for the French Gross Domestic Product (GDP) growth. Our approach is the

following. We assume that we are given a set of basic predictors - this is a usual approach in statistical learning, the predictors are sometimes referred as "experts", e.g. Cesa-Bianchi and Lugosi (2006). Following Alquier and Wintenberger (2012), we describe a procedure of aggregation, usually referred as Exponentially Weigthed Agregate (EWA), Dalalyan and Tsybakov (2008); Gerchinovitz (2011), or Gibbs estimator, Catoni (2004, 2007). It is interesting to note that this procedure is also related to aggregations procedure in online learning as the weighted majority algorithm of Littlestone and Warmuth (1994), see also Vovk (1990). We give a PAC-Bayesian inequality that ensures optimality properties for this procedure. In a few words, this inequality claims that our predictor performs as well as the best basic predictor up to a remainder of the order $\mathcal{K}/\sqrt{n}$ where n is the number of observations and $\mathcal{K}$ measures the complexity of the set of basic predictors. This result is very general, two conditions will be required: the time series must be weakly dependent in a sense that we will make more precise in Section 4, and we need to have a Lipshitz loss function. This includes, in particular, the quantile loss functions. This allows us to apply this result to our problem of economic forecasting.

The paper is organized as follows. Section 2 provides the notations used in the whole paper. Then, we give a description the Gibbs estimator in Section 3. The PAC-Bayesian inequality, Theorem 4.1, is given in Section 4, and the application to quantile losses and GDP forecasting in Section 5. Finally, the proof of Theorem 4.1 is postponed to the appendix.

2. The context

Let us assume that we observe $X_1, \dots, X_n$ from a $\mathbb{R}^p$-valued stationnary time series $X = (X_t)_{t\in\mathbb{Z}}$ defined on $(\Omega, \mathcal{A}, \mathbb{P})$. From now, $\|.\|$ will denote the Euclidian norm on $\mathbb{R}^p$. Fix an integer k and let us assume that we are given a family of predictors $\left\{ f_\theta : (\mathbb{R}^p)^k \to \mathbb{R}^p, \theta \in \Theta \right\}$: for any θ and any t, f_θ applied to the last past values $(X_{t-1}, \dots, X_{t-k})$ is a possible prediction of X_t. For the sake of simplicity, let us put for any $t \in \mathbb{Z}$ and any $\theta \in \Theta$,

$$\hat{X}_t^\theta = f_\theta(X_{t-1}, \dots, X_{t-k}).$$

We also assume that $\theta \mapsto f_\theta$ is linear. Note that we may want to include parametric models as well as non-parametric prediction. In order to deal with various family of predictors, we propose a model-selection type approach:

$$\Theta = \bigcup_{j=1}^{m} \Theta_j.$$

Example 2.1 *A first example is the linear auto-regressive class of predictors. We can take* $\theta = (\theta_0, \theta_1, \dots, \theta_k) \in \Theta = \mathbb{R}^{k+1}$ *and*

$$f_\theta(X_{t-1}, \dots, X_{t-k}) = \theta_0 + \sum_{j=1}^{k} \theta_j X_{t-j}.$$

In this case we deal with only one model, $m = 1$ *and* $\Theta = \Theta_1$.

Example 2.2 *We may generalize the previous example to non-parametric auto-regression, for example using a dictionnary of functions* $(\mathbb{R}^p)^k \to \mathbb{R}^p$, *say* $(\varphi_i)_{i=0}^{\infty}$. *Then we can fix* $m = n$, *and take* $\theta = (\theta_1, \dots, \theta_\ell) \in \Theta_j = \mathbb{R}^j$ *and*

$$f_\theta(X_{t-1}, \dots, X_{t-k}) = \sum_{i=1}^{j} \theta_i \varphi_i(X_{t-1}, \dots, X_{t-k}).$$

Finally, we have to define a quantitative criterion to evaluate the quality of the predictions. Let ℓ be a loss function. More precisely, we will assume that ℓ satisfies the following assumption.
Assumption LipLoss: ℓ is given by: $\ell(x, x') = g(x - x')$ for some convex function g satisfying $g \geq 0$, $g(0) = 0$ and g is K-Lipshitz.

Definition 2.1 *We put, for any* $\theta \in \Theta$,

$$R(\theta) = \mathbb{E}\left[\ell\left(\hat{X}_t^\theta, X_t\right)\right].$$

Note that because of the stationnarity, $R(\theta)$ does not depend on t.

Example 2.3 *A first example is* $\ell(x, x') = \|x - x'\|$. *In this case, the Lipshitz constant* K *is* 1. *This example was studied in detail in Alquier and Wintenberger (2012). In Modha and Masry (1998); Meir (2000), the loss function is the quadratic loss* $\ell(x, x') = \|x - x'\|^2$. *Note that it also satisfies our Lipshitz condition, but only if we assume that the time series is bounded.*

Example 2.4 *When the time-series is real-valued, we can use a quantile loss function. The class of quantile loss functions is defined as*

$$\ell_\tau(x, y) = \begin{cases} \tau(x - y), & \text{if } x - y > 0 \\ -(1 - \tau)(x - y), & \text{otherwise} \end{cases}$$

where $\tau \in (0, 1)$. *It is motivated by the following remark: if* U *is a real-valued random variable, then any value* t^* *satisfying* $\mathbb{P}(U \leq t^*) = \tau$ *is a minimizer of of* $t \mapsto \mathbb{E}(\ell_\tau(X - t))$; *such a value is called quantile of order* τ *of* U. *This loss function was introduced by Koenker and Bassett (1978) for "quantile regression", since then it became a classical tool in statistics, see e.g. Koenker (2005) for a survey. Recently, Belloni and Chernozhukov (2011) used it in the context of high-dimensional regression with the LASSO and by Biau and Patra (2011) used it to build non-parametric confidence intervals on time-series.*

3. Gibbs estimator

We introduce in this section the Gibbs estimator. As already mentionned in the introduction, such aggregated estimators were used in learning theory under the name weighted majority aggregate, EWA...

Definition 3.1 *We define, for any $\theta \in \Theta$, the empirical risk*

$$r_n(\theta) = \frac{1}{n-k} \sum_{i=k+1}^{n} \ell\left(\hat{X}_i^{\theta}, X_i\right).$$

Let $\mathcal{T}$ be a σ-algebra on Θ and $\mathcal{T}_\ell$ be its restriction to Θ_ℓ for any $\ell \in \{1, \ldots, m\}$. Let $\mathcal{M}_+^1(\Theta)$ denote the set of all probability measures on $(\Theta, \mathcal{T})$. Let $\pi \in \mathcal{M}_+^1(\Theta)$. This probability measure is usually called the *prior* by analogy with Bayesian statistics. Actually, it will be used as a tool to control the complexity of the set of predictors Θ.

Remark 3.1 *In the case where $\Theta = \cup_j \Theta_j$ and the Θ_j are disjoint, we can write*

$$\pi(\mathrm{d}\theta) = \sum_{j=1}^{m} \mu_j \pi_j(\mathrm{d}\theta)$$

where $\mu_j := \pi(\Theta_j)$ and $\pi_j(\mathrm{d}\theta) := \pi(\mathrm{d}\theta)\mathbf{1}_{\Theta_j}(\theta)/\mu_j$. Note that π_j can be interpreted as a prior probability measure inside the model Θ_j and that the weights μ_j can be interpreted as a priori probability measure between the models.

Definition 3.2 *We put, for any $\lambda > 0$,*

$$\hat{\theta}_\lambda = \int_\Theta \theta \hat{\rho}_\lambda(\mathrm{d}\theta)$$

where

$$\hat{\rho}_\lambda(\mathrm{d}\theta) = \frac{e^{-\lambda r_n(\theta)}\pi(\mathrm{d}\theta)}{\int e^{-\lambda r_n(\theta')}\pi(\mathrm{d}\theta')}.$$

Remark 3.2 *Note that analogously to Bayesian estimator, the Gibbs estimator can is written as an integral on the parameter space. It can thus be computed by Monte Carlo methods, see Robert (1996); Marin and Robert (2007). This is the approach that we will use in this paper.*

Remark 3.3 *The choice of the parameter λ is discussed in the next section.*

4. Theoretical results

In this section, we provide a PAC-Bayesian oracle inequality for the Gibbs estimator. PAC-Bayesian were introduced in the context of supervised classification (using the 0/1-loss), see the seminal papers Shawe-Taylor and Williamson (1997); McAllester (1999). More general versions can be found in Catoni (2004, 2007). These results were generalized to different contexts and loss functions, see Alquier (2008) for a presentation with a general loss function. See also Audibert (2010) for a nice survey of the more recent advances. The idea is that the risk of the Gibbs estimator will be close to $\inf_\theta R(\theta)$ up to a small remainder. More precisely, we upper-bound it by

$$\inf_\rho \left\{ \int R(\theta)\rho(\mathrm{d}\theta) + \text{ remainder}(\rho, \pi) \right\}$$

where the inf is taken upon all the probability distributions on Θ.

In order to be able to control the prevision risk of our estimator $\hat{\theta}_\lambda$, $R(\hat{\theta}_\lambda)$, we will need some hypothesis. The first hypothesis concerns the dependence of the process, it uses the $\theta_{\infty,n}(1)$-coefficients of Dedecker et al. (2007). Such a condition is also used in Alquier and Wintenberger (2012), and is more general than the mixing conditions used in Meir (2000); Modha and Masry (1998).

Assumption WeakDep: we assume that the distribution $\mathbb{P}$ is such that the stationary process $(X_t)_{t\in\mathbb{Z}}$ is bounded, *ie* a.s. $\|X_0\|_\infty \leq \mathcal{B} < \infty$, and such that there is a constant $\mathcal{C}$ with $\theta_{\infty,k}(1) \leq \mathcal{C} < \infty$ for any k. We remind that for any σ-algebra $\mathfrak{S} \subset \mathcal{A}$, for any $q \in \mathbb{N}$, for any $(\mathbb{R}^p)^q$-valued random variable Z defined on $(\Omega, \mathcal{A}, \mathbb{P})$, we put

$$\theta_\infty(\mathfrak{S}, Z) = \sup_{f\in\Lambda_1^q} \left\| \mathbb{E}\left[f(Z)|\mathfrak{S}\right] - \mathbb{E}\left[f(Z)\right] \right\|_\infty$$

where

$$\Lambda_1^q = \left\{ f : (\mathbb{R}^p)^q \to \mathbb{R}, \quad \frac{|f(z_1,\dots,z_q) - f(z_1',\dots,z_q')|}{\sum_{j=1}^q \|z_j - z_j'\|} \leq 1 \right\},$$

and that

$$\theta_{\infty,k}(1) := \sup\left\{\theta_\infty(\sigma(X_t, t\leq p), (X_{j_1},\dots,X_{j_\ell})), \quad p < j_1 < \ldots < j_\ell, 1 \leq \ell \leq k\right\}.$$

Remark 4.1 *Some examples of processes satisfying* **WeakDep** *are provided, for example, in Alquier and Wintenberger (2012). It includes the large family of bounded causal Bernoulli shifts, that is bounded processes of the form*

$$X_t = H(\xi_t, \xi_{t-1}, \xi_{t-2}, \dots)$$

where the "innovations" ξ_t are iid and bounded and H satisfies a Lipshitz-type condition. In particular, this includes ARMA processes with bounded innovations. It also includes uniform φ-mixing processes, defined e.g. in Doukhan (1994); Rio (2000a), and some dynamical systems.

Assumption Lip: for any $\theta \in \Theta$ we assume that there are coefficients $a_j(\theta)$ for $1 \leq j \leq k$ satisfying, for any x_1, ..., x_k and y_1, ..., y_k, the relation

$$\|f_\theta(x_1,\dots,x_k) - f_\theta(y_1,\dots,y_k)\| \leq \sum_{j=1}^k a_j(\theta)\|x_j - y_j\|.$$

We define $L := \sup_{\theta\in\Theta}\sum_{j=1}^k a_j(\theta)$ and assume that this value is finite.

Theorem 4.1 (PAC-Bayesian Oracle Inequality) *Let us assume that assumptions* **LipLoss**, **WeakDep** *and* **Lip** *are satisfied. Then, for any $\lambda > 0$, for any $\varepsilon > 0$,*

$$\mathbb{P}\left\{ R\left(\hat{\theta}_\lambda\right) \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta)} \left[\int R\mathrm{d}\rho + \frac{2\lambda\kappa^2}{n\left(1-\frac{k}{n}\right)^2} + \frac{2\mathcal{K}(\rho,\pi) + 2\log\left(\frac{2}{\varepsilon}\right)}{\lambda} \right] \right\} \geq 1-\varepsilon$$

where $\kappa = \kappa(K, L, \mathcal{B}, \mathcal{C}) := K(1+L)(\mathcal{B}+\mathcal{C})/\sqrt{2}$ *and where we remind that* $\mathcal{K}(\rho, \pi)$ *is the Kullback divergence between* ρ *and* π*, defined by*

$$\mathcal{K}(\rho,\pi) = \begin{cases} \int \log\left[\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta)\right] \rho(\mathrm{d}\theta) \text{ if } \rho \ll \pi, \\ \\ +\infty \text{ otherwise.} \end{cases}$$

Remark 4.2 *The choice of* λ *in practice may be a problem. In Catoni (2003, 2007) a general method is proposed to optimize the bound with respect to* λ*. However, while adapted in the iid case, this method is more difficult to use in the context of time series as it would require the knowledge of* κ*, and so the knowledge of* $\theta_{\infty,n}(1)$ *- or at least the knowledge of an explicit upper bound for* $\theta_{\infty,n}(1)$*. In practice, however, some empirical calibration seems to give good results, as shown in Section 5.*

Remark 4.3 *We want to mention that, at the price of a much more technical analysis, this result can be extended to the case where the* X_t *are not assumed to be bounded. In the iid case, it is possible to obtain results under the existence of moments of order* 4 *only, see Audibert and Catoni; Catoni. In the context of time series, the results in Alquier and Wintenberger (2012) require subGaussian tails for* X_t*, but suffer a* $\log(n)$ *loss in the learning rate.*

5. Application to French GDP and quantile prediction

We now in this section an application to data published by the INSEE (*Institut National de la Statistique et des Etudes Economiques*, the French national bureau of statistics).

5.1. Uncertainty in GDP forecasting

Every quarter t, economic forecasters at INSEE are asked a prediction for the quarterly growth rate of the French Gross Domestic Product (GDP). Since it involve a lot of information, the "true value" of the growth rate $\log(\mathrm{GDP}_t/\mathrm{GDP}_{t-1})$ is only known after two years, but *flash estimates* of the growth rate, say $\Delta\mathrm{GDP}_t$, are published 45 days after the end of the current quarter t. One of the most relevant economic information available at time t to the forecaster, apart from past GDP observations, are *business surveys*. Indeed, they are a rich source of information, for at least two reasons. First, they are rapidly available, on a monthly basis. Moreover, they provide information coming directly from the true economic decision makers.

A business survey is traditionally a fixed questionnaire of ten questions sent monthly to a panel of companies. This process is described in Devilliers (1984). INSEE publishes a composite indicator called the *French business climate indicator*: it summarises information of the whole survey. This indicator is defined in Clavel and Minodier (2009), see also Dubois and Michaux (2006). All these values are available from the INSEE website

`http://www.insee.fr/`

Note that a quite similar approach is used in other countries, see also Biau et al. (2008) for a prediction of the European Union GDP based on EUROSTATS data (EUROSTAT is the EU bureau of statistics).

It is however well known among economic forecasters that interval confidence or density forecasts are to be given with the prediction, in order to provide an idea of the uncertainty of the prediction. The ASA and the NBER started using density forecasts in 1968, see Diebold et al. (1997); Tay and Wallis (2000) for historical surveys on density forecasting. The Central Bank of England and INSEE, among others, provide their prediction with a "fan chart", Britton et al. (1998). However, it is interesting to note that the methdology used is often very crude, see the criticism in Cornec (2010); Dowd (2004). For example, until 2012, the fan chart provided by the INSEE led to the construction of confidence intervals with constant length. But there is an empirical evidence that it is more difficult to forecast economic quantities during crisis (e.g. the subprime crisis in 2008). The Central Bank of England fan chart is not reproducible as it includes subjective information. Recently, Cornec (2010) proposed a clever density forecasting method based on quantile regressions that gives satisfying results in practice. However, this method did not receive any theoretical support up to our knowledge.

Here, we use the Gibbs estimator described in the previous sections to build a forecasting of ΔGDP_t, using the quantile loss function. This allows to return a prediction: the forecasted median, for $\tau = 0.5$, that is theoretically supported. This also allows to provide various confidence intervals corresponding to various quantiles.

5.2. Application of Theorem 4.1

At each quarter t, the objective is to predict the flash estimate of GDP growth, ΔGDP_t. As described previouly, the available information is $\Delta\text{GDP}_{t'}$ for $t' < t$ and $I_{t'}$ for $t' < t$, where for notational convenience, I_{t-1} is the climate indicator available to the INSEE at time t (it is the mean of the climate indicator at month 3 of quarter $t-1$ and at month 1 and 2 of quarter t). The observation period is 1988-Q1 (1st quarter of 1988) to 2011-Q3.

We define $X_t = (\Delta\text{GDP}_t, I_t)' \in \mathbb{R}^2$. As we are not interested by the prevision of I_t but only by the prediction of the GDP growth, the loss function will only take into account ΔGDP_t. We use the quantile loss function of Example 2.4:

$$\ell_\tau((\Delta\text{GDP}_t, I_t), (\Delta'\text{GDP}_t, I'_t)) = \begin{cases} \tau\left(\Delta\text{GDP}_t - \Delta'\text{GDP}_t\right), & \text{if } \Delta\text{GDP}_t - \Delta'\text{GDP}_t > 0 \\ -\left(1-\tau\right)\left(\Delta\text{GDP}_t - \Delta'\text{GDP}_t\right), & \text{otherwise.} \end{cases}$$

In order to clearly know what is the value τ we are dealing with, we will now add a subscript τ in the notation of the prevision risk:

$$R^\tau(\theta) := \mathbb{E}\left[\ell_\tau\left(\Delta\text{GDP}_t, f_\theta(X_{t-1}, X_{t-2})\right)\right].$$

We also let r_n^τ denote the associated empirical risk.

Following Cornec (2010); Li (2010) we consider predictors of the form:

$$f_\theta(X_{t-1}, X_{t-2}) = \theta_0 + \theta_1 \Delta\text{GDP}_{t-1} + \theta_2 I_{t-1} + \theta_3 (I_{t-1} - I_{t-2}) |I_{t-1} - I_{t-2}| \tag{1}$$

where $\theta = (\theta_0, \theta_1, \theta_2, \theta_3) \in \Theta(B)$. For any $B > 0$ we define

$$\Theta(B) = \left\{ \theta = (\theta_0, \theta_1, \theta_2, \theta_3) \in \mathbb{R}^4, \|\theta\|_1 = \sum_{i=0}^{3} |\theta_i| \leq B \right\}.$$

These predictors of Equation 1 correspond to the model used in Cornec (2010) for forecasting, one of the conclusions of Cornec (2010); Li (2010) is that these family of predictors allow to obtain a forecasting as precise as the INSEE one.

For technical reason that will become clear in the proofs, if one wants to achieve a prediction performance comparable to the best $\theta \in \Theta(B)$, it is more convenient to define the prior π as the uniform probability distribution on some slightly larger set, e.g. $\Theta(B+1)$. We will let Π_B denote this distribution. We let $\hat{\rho}^\tau_{B,\lambda}$ and $\hat{\theta}^\tau_{B,\lambda}$ denote repectively the associated agregation distribution and the associated estimator, defined in Definition 3.2.

Remark that in this framework, Assumption **Lip** is satisfied with $L = B + 1$, and the loss function is K-Lipshitz with $K = 1$ so Assumption **LipLoss** is also satisfied.

Theorem 5.1 *Let us fix $\tau \in (0,1)$. Let us assume that Assumption* **WeakDep** *is satisfied, and that $n \geq \max\left(10, \kappa^2/(3\mathcal{B}^2)\right)$. Let us fix $\lambda = \sqrt{3n}/\kappa$. Then, with probability at least $1-\varepsilon$ we have*

$$R^\tau(\hat{\theta}^\tau_{B,\lambda}) \leq \inf_{\theta\in\Theta(B)} \left\{ R^\tau(\theta) + \frac{2\sqrt{3}\kappa}{\sqrt{n}} \left[2.25 + \log\left(\frac{(B+1)\mathcal{B}\sqrt{n}}{\kappa} \right) + \frac{\log\left(\frac{1}{\varepsilon}\right)}{3} \right] \right\}.$$

A detailed proof is given in the appendix.

The choice of λ proposed in the theorem may be a problem as in practice we will not know κ. Note that from the proof, it is obvious that in any case, for n large enough, when $\lambda = \sqrt{n}$ we still have a bound

$$R^\tau(\hat{\theta}^\tau_{B,\lambda}) \leq \inf_{\theta\in\Theta(B)} \left\{ R^\tau(\theta) + \frac{C(B,\mathcal{B},\kappa,\varepsilon)}{\sqrt{n}} \right\}.$$

However, in practice, we will work in an online setting: at each date t we compute the Gibbs estimator based on the observations from 1 to t and use it to predict the GDP and its quantiles at time $t+1$. Let $\hat{\theta}^\tau_{B,\lambda}[t]$ denote this estimator. We propose the following empirical approach: we define a set of values $\Lambda = \{2^k, k \in \mathbb{N}\} \cap \{1, ..., n\}$. At each step t, we compute $\hat{\theta}^\tau_{B,\lambda}[t]$ for each $\lambda \in \Lambda$ and use for prediction $\hat{\theta}^\tau_{B,\lambda(t)}[t]$ where $\lambda(t)$ is defined by

$$\lambda(t) = \arg\min_{\lambda\in\Lambda} \sum_{j=3}^{t-1} \ell_\tau(\Delta GDP_j, f_{\hat{\theta}^\tau_{B,\lambda}[j]}(X_{j-1}, X_{j-2})),$$

namely, the value that is currently the best for online prediction. This choice leads to good numerical results.

In practice, the choice of B has less importance. As soon as B is large enough, the estimator does not really depend on B, only the theoretical bound does. As a consequence we take $B = 100$ in our experiments.

5.3. Implementation

We use the importance sampling method to compute $\hat{\theta}^\tau_{B,\lambda}[t]$ (see, e.g., Robert (1996)). We draw an iid sample T_1, ..., T_N of vectors in $\mathbb{R}^4$, from the distribution $\mathcal{N}(\hat{\theta}^\tau, vI)$ where $v > 0$ and $\hat{\theta}^\tau$ si simply the τ-quantile regression estimator of θ in (1), as computed by the "quantile

regression package" of the R software R Development Core Team (2008). Let $g(\cdot)$ denote the density of this distribution. Then, by the law of large numbers we can approximate

$$\sum_{i=1}^{N} \frac{T_i \exp\left[-\lambda r_t(T_i)\right] \mathbf{1}_{\Theta(B+1)}(T_i)}{g(T_i) \sum_{j=1}^{N} \frac{\exp[-\lambda r_t(T_j)] \mathbf{1}_{\Theta(B+1)}(T_j)}{g(T_j)}} \xrightarrow[N\to\infty]{a.s.} \hat{\theta}^{\tau}_{B,\lambda}[t].$$

Remark that this is particularly convenient as we only simulate the sample T_1, ..., T_N once and we can use the previous formula to approximate $\hat{\theta}^{\tau}_{B,\lambda}[t]$ for several different values of τ.

5.4. Results

The results are shown in Figure 1 for prediction, $\tau = 0.5$, in Figure 2 for confidence interval of order 50%, i.e. $\tau = 0.25$ and $\tau = 0.75$ (left) and for confidence interval of order 90%, i.e. $\tau = 0.05$ and $\tau = 0.95$ (right). We report only the results for the period 2000-Q1 to 2011-Q3 (using the period 1988-Q1 to 1999-Q4 for learning).

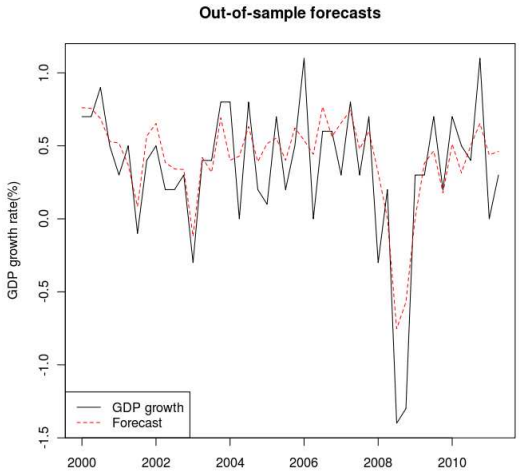

Figure 1: French GDP online prediction using the quantile loss function with $\tau = 0.5$.

Note that we can compare the ability of our predictor $\widehat{\theta}^{0.5}_{B,\lambda}$ with the predictor used in Li (2010) that relies on a least square estimation of (1), that we will denote by $\hat{\theta}^{*}$. Interestingly, both are quite similar but $\widehat{\theta}^{0.5}_{B,\lambda}$ is a bit more precise. We remind that

$$\begin{aligned} \text{mean abs. pred. error} &= \tfrac{1}{n}\textstyle\sum_{t=1}^{n} \left| \Delta GDP_t - f_{\hat{\theta}^{0.5}_{B,\lambda(t)}[t]}(X_{t-1}, X_{t-2}) \right| \\ \text{mean quad. pred. error} &= \tfrac{1}{n}\textstyle\sum_{t=1}^{n} \left[\Delta GDP_t - f_{\hat{\theta}^{0.5}_{B,\lambda(t)}[t]}(X_{t-1}, X_{t-2}) \right]^2. \end{aligned}$$

Predictor	Mean absolute prevision error	Mean quadratic prevision error
$\widehat{\theta}^{0.5}_{B,\lambda}$	0.22360	0.08033
$\widehat{\theta}_{\star}$	0.24174	0.08178

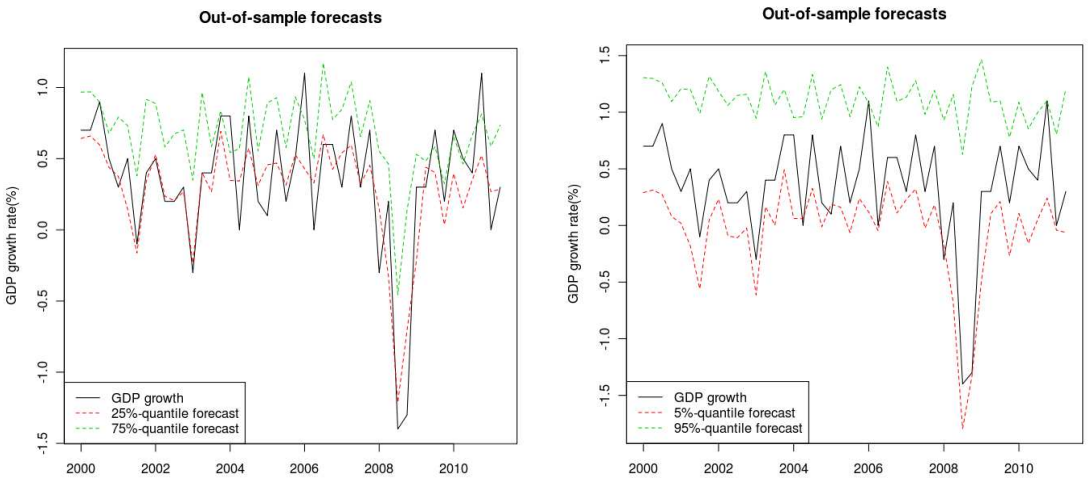

Figure 2: French GDP online 50%-confidence intervals (left) and 90%-confidence intervals (right).

We also report the frequency of realizations of the GDP falling above the predicted τ-quantile for each τ. Note that this quantity should be close to τ.

Estimator	Frequency
$\widehat{\theta}_{B,\lambda}^{0.05}$	0.065
$\widehat{\theta}_{B,\lambda}^{0.25}$	0.434
$\widehat{\theta}_{B,\lambda}^{0.5}$	0.608
$\widehat{\theta}_{B,\lambda}^{0.75}$	0.848
$\widehat{\theta}_{B,\lambda}^{0.95}$	0.978

It can be seen that our method behaves quite well in practice. As the INSEE did, we miss the value of the 2008 crisis. However, it is interesting to note that our confidence interval shows that our prediction at this date is less reliable than the previous ones: so, at this time, the forecaster could have been aware of some problems in their predictions.

6. Conclusion

We proposed some theoretical results to extend learning theory to the context of weakly dependent time series. The method showed good results on an application to GDP forecasting. It would also be interesting to give theoretical results on the online risk of our method, e.g. following tools in Catoni (2004); Gerchinovitz (2011). From both theoretical and practical perspective, an adaptation with respect to the dependence coefficient $\theta_{\infty,n}(1)$ would also be really interesting but is probably a more difficult objective.

References

P. Alquier. Pac-bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics*, 17(4):279–304, 2008.

P. Alquier and O. Wintenberger. Model selection for weakly dependent time series forecasting. *Bernoulli* (to appear), available on arXiv:0902.2924, 2012.

J.-Y. Audibert. Pac-bayesian aggregation and multi-armed bandits. HDR Université Paris Est, 2010.

J.-Y. Audibert and O. Catoni. Robust linear least squares regression. Preprint arXiv:1010.0074, to appear in the Annals of Statistics.

A. Belloni and V. Chernozhukov. L1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

G. Biau and B. Patra. Sequential quantile prediction of time series. *IEEE Transactions on Information Theory*, 57:1664–1674, 2011.

G. Biau, O. Biau, and L. Rouvière. Nonparametric forecasting of the manufacturing output growth with firm-level survey data. *Journal of Business Cycle Measurement and Analysis*, 3:317–332, 2008.

E. Britton, P. Fisher, and J. Whitley. The inflation report projections: Understanding the fan chart. *Bank of England Quarterly Bulletin*, 38(1):30–37, 1998.

P. Brockwell and R. Davis. *Time Series: Theory and Methods (2nd Edition)*. Springer, 2009.

O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. Annales de l'IHP, to appear.

O. Catoni. A pac-bayesian approach to adaptative classification. *Preprint Laboratoire de Probabilités et Modèles Aléatoires*, 2003.

O. Catoni. *Statistical Learning Theory and Stochastic Optimization, Lecture Notes in Mathematics (Saint-Flour Summer School on Probability Theory 2001, ed. J. Picard)*. Springer, 2004.

O. Catoni. *PAC-Bayesian Supervised Classification (The Thermodynamics of Statistical Learning)*, volume 56 of *Lecture Notes-Monograph Series*. IMS, 2007.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, 2006.

L. Clavel and C. Minodier. A monthly indicator of the french business climate. Documents de Travail de la DESE, 2009.

M. Cornec. Constructing a conditional gdp fan chart with an application to french business survey data. 30th CIRET Conference, New York, 2010.

A. Dalalyan and A. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72:39–61, 2008.

J. Dedecker, P. Doukhan, G. Lang, J. R. León, S. Louhichi, and C. Prieur. *Weak Dependence, Examples and Applications*, volume 190 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin, 2007.

M. Devilliers. Les enquêtes de conjoncture. In *Archives et Documents*, number 101. INSEE, 1984.

F. X. Diebold, A. S. Tay, and K. F. Wallis. Evaluating density forecasts of inflation: the survey of professional forecasters. Discussion Paper No.48, ESRC Macroeconomic Modelling Bureau, University of Warwick and Working Paper No.6228, National Bureau of Economic Research, Cambridge, Mass., 1997.

P. Doukhan. *Mixing*, volume 85 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1994.

K. Dowd. The inflation fan charts: An evaluation. *Greek Economic Review*, 23:99–111, 2004.

E. Dubois and E. Michaux. étalonnages à l'aide d'enquêtes de conjoncture: de nouvaux résultats. In *Économie et Prévision*, number 172. INSEE, 2006.

S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. In *Proceedings of COLT'11*, 2011.

J. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

R. Koenker. *Quantile Regression*. Cambridge University Press, Cambridge, 2005.

R. Koenker and G. Jr. Bassett. Regression quantiles. *Econometrica*, 46:33–50, 1978.

S. Kullback. *Information theory and statistics*. Wiley, New York, 1959.

X. Li. Agrégation de prédicteurs appliquée à la conjoncture. Rapport de stage de M2 - Université Paris 6, effectué à l'INSEE sous la direction de Matthieu Cornec, 2010.

N. Littlestone and M.K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

J.-M. Marin and C. P. Robert. *Bayesian Core: A practical approach to computational Bayesian analysis*. Springer, 2007.

D. A. McAllester. Pac-bayesian model averaging. In *Procs. of of the 12th Annual Conf. On Computational Learning Theory, Santa Cruz, California (Electronic)*, pages 164–170. ACM, New-York, 1999.

R. Meir. Nonparametric time series prediction through adaptive model selection. *Machine Learning*, 39:5–34, 2000.

D. S. Modha and E. Masry. Memory-universal prediction of stationary random processes. *IEEE transactions on information theory*, 44(1):117–133, 1998.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, 2008.

E. Rio. *Théorie asymptotique des processus aléatoires faiblement dépendants*, volume 31 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2000a.

E. Rio. Ingalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *Comptes Rendus de l'Académie des Sciences de Paris, Srie I*, 330:905–908, 2000b.

C. P. Robert. *Méthods de Monte Carlo par chaines de Markov*. Economica (Paris), 1996.

Y. Seldin, F. Laviolette, N. Cesa-Bianchi, P. Auer, and J. Shawe-Taylor. Pac-bayesian inequalities for martingales. 2011.

J. Shawe-Taylor and R. Williamson. A pac analysis of a bayes estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory, COLT'97*, pages 2–9. ACM, 1997.

G. Stoltz. Agrégation séquentielle de prédicteurs : méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique. *Journal de la SFDS*, 151(2):66–106, 2010.

A. S. Tay and K. F. Wallis. Density forecasting: a survey. *Journal of Forecasting*, 19: 235–254, 2000.

V.G. Vovk. Aggregating strategies. In *Proceedings of the 3rd Annual Workshop on Computational Learning Theory (COLT)*, pages 372–283, 1990.

O. Wintenberger. Deviation inequalities for sums of weakly dependent time series. *Electronic Communications in Probability*, 15:489–503, 2010.

Appendix A. Proofs

A.1. Some preliminary lemmas

First, we remind Rio's Hoeffding type inequality.

Lemma 1 (Rio Rio (2000b)) *Let h be a function $(\mathbb{R}^p)^n \to \mathbb{R}$ such that*

$$\forall (x_1,\dots,x_n,y_1,\dots,y_n) \in (\mathbb{R}^p)^{2n}, |h(x_1,\dots,x_n) - h(y_1,\dots,y_n)| \leq \sum_{i=1}^{n} \|x_i - y_i\|. \qquad (2)$$

Then for any $t > 0$ we have

$$\mathbb{E}\left(e^{t\{\mathbb{E}[h(X_1,\dots,X_n)] - h(X_1,\dots,X_n)\}}\right) \leq e^{\frac{t^2 n(\mathcal{B}+\theta_{\infty,n}(1))^2}{2}}.$$

Note that others Hoeffding and Bernstein type inequalities could be used to obtain PAC-Bounds in the context of time series. The monographs Doukhan (1994); Rio (2000a) provide nice review of the results available for mixing time series. Note however that weak dependence assumptions are usually more general, some inequalities are provided in Dedecker et al. (2007), a nice review and new results are given in Wintenberger (2010). See also the martingale approach in Seldin et al. (2011). However, Lemma 1 is particularly convenient in this setting, and leads to particularly general hypothesis.

Using Lemma 1, we can prove the following lemma.

Lemma 2 *Let us assume that Assumptions* **LipLoss**, **WeakDep** *and* **Lip** *are satisfied. For any $\lambda > 0$, for any $\theta \in \Theta$, we have*

$$\mathbb{E}\left(e^{\lambda[R(\theta)-r_n(\theta)]}\right) \leq e^{\frac{\lambda^2\kappa^2}{n\left(1-\frac{k}{n}\right)^2}} \text{ and } \mathbb{E}\left(e^{\lambda[r_n(\theta)-R(\theta)]}\right) \leq e^{\frac{\lambda^2\kappa^2}{n\left(1-\frac{k}{n}\right)^2}},$$

where we remind that $\kappa = K(1+L)(\mathcal{B}+\mathcal{C})/\sqrt{2}$.

Proof Let us fix $\lambda > 0$ and $\theta \in \Theta$. Let us define the function h by:

$$h(x_1, \ldots, x_n) = \frac{1}{K(1+L)} \sum_{i=k+1}^{n} \ell(f_\theta(x_{i-1}, \ldots, x_{i-k}), x_i).$$

We now check that h satisfies (2),

$$\begin{aligned}
&\left|h\left(x_1, \ldots, x_n\right) - h\left(y_1, \ldots y_n\right)\right| \\
&\quad \leq \frac{1}{K(1+L)} \sum_{i=k+1}^{n} \left|\ell(f_\theta(x_{i-1}, \ldots, x_{i-k}), x_i) - \ell(f_\theta(y_{i-1}, \ldots, y_{i-k}), y_i)\right| \\
&\quad \leq \frac{1}{K(1+L)} \sum_{i=k+1}^{n} \left|g(f_\theta(x_{i-1}, \ldots, x_{i-k}) - x_i) - g(f_\theta(y_{i-1}, \ldots, y_{i-k}) - y_i)\right| \\
&\quad \leq \frac{1}{1+L} \sum_{i=k+1}^{n} \left\|\left(f_\theta(x_{i-1}, \ldots, x_{i-k}) - x_i\right) - \left(f_\theta(y_{i-1}, \ldots, y_{i-k}) - y_i\right)\right\|
\end{aligned}$$

where we used Assumption **LipLoss** for the last inequality. So we have

$$\begin{aligned}
&\left|h\left(x_1, \ldots, x_n\right) - h\left(y_1, \ldots y_n\right)\right| \\
&\quad \leq \frac{1}{1+L} \sum_{i=k+1}^{n} \left(\left\|f_\theta(x_{i-1}, \ldots, x_{i-k}) - f_\theta(y_{i-1}, \ldots, y_{i-k})\right\| + \left\|x_i - y_i\right\|\right) \\
&\quad \leq \frac{1}{1+L} \sum_{i=k+1}^{n} \left(\sum_{j=1}^{k} a_j(\theta)\|x_{i-j} - y_{i-j}\| + \|x_i - y_i\|\right) \\
&\quad \leq \frac{1}{1+L} \sum_{i=1}^{n} \left(1 + \sum_{j=1}^{k} a_j(\theta)\right) \|x_i - y_i\|
\end{aligned}$$

$$\leq \sum_{i=1}^{n} \|x_i - y_i\|$$

where we used Assumption **Lip**. So we can apply Lemma 1. Note that $h(X_1, \ldots, X_n) = \frac{n-k}{K(1+L)} r_n(\theta)$, $\mathbb{E}(h(X_1, \ldots, X_n)) = \frac{n-k}{K(1+L)} R(\theta)$ and we choose $t = K(1+L)\lambda/(n-k)$, we obtain:

$$\mathbb{E}\left(e^{\lambda[R(\theta)-r_n(\theta)]}\right) \leq e^{\frac{\lambda^2 K^2 (1+L)^2 (\mathcal{B}+\theta_{\infty,n}(1))^2}{2n\left(1-\frac{k}{n}\right)^2}}$$
$$\leq e^{\frac{\lambda^2 K^2 (1+L)^2 (\mathcal{B}+\mathcal{C})^2}{2n\left(1-\frac{k}{n}\right)^2}}$$

because of Assumption **WeakDep**. This ends the proof of the first inequality. The reverse inequality is obtained by replacing the function h by $-h$. ■

We also remind the following classical result concerning the Kullback divergence function.

Lemma 3 (Legendre transform of the Kullback divergence function) *For any $\pi \in \mathcal{M}_+^1(E)$, for any measurable function $h : E \to \mathbb{R}$ such that $\pi[\exp(h)] < +\infty$ we have:*

$$\pi[\exp(h)] = \exp\left(\sup_{\rho \in \mathcal{M}_+^1(E)} \left(\rho[h] - \mathcal{K}(\rho, \pi)\right)\right), \tag{3}$$

with convention $\infty - \infty = -\infty$. Moreover, as soon as h is upper-bounded on the support of π, the supremum with respect to ρ in the right-hand side is reached for the Gibbs measure $\pi\{h\}$ defined by

$$\pi\{h\}(\mathrm{d}x) = \frac{e^{h(x)}\pi(\mathrm{d}x)}{\pi[\exp(h)]}.$$

Actually, it seems that in the case of discrete probabilities, this result was already known by Kullback (Problem 8.28 of Chapter 2 in Kullback (1959)). For a complete proof in the general case, we refer the reader for example to Catoni (2003, 2007). We are now ready to state the following key result.

Lemma 4 *Let us assume that Assumptions* **LipLoss**, **WeakDep** *and* **Lip** *are satisfied. Let us fix $\lambda > 0$. Let k be defined as in Lemma 2. Then,*

$$\mathbb{P}\left\{\begin{array}{l} \forall \rho \in \mathcal{M}_+^1(\Theta), \\ \int R\mathrm{d}\rho \leq \int r_n \mathrm{d}\rho + \frac{\lambda\kappa^2}{n\left(1-\frac{k}{n}\right)^2} + \frac{\mathcal{K}(\rho,\pi)+\log\left(\frac{2}{\varepsilon}\right)}{\lambda} \\ \quad \textit{and} \\ \int r_n \mathrm{d}\rho \leq \int R\mathrm{d}\rho + \frac{\lambda\kappa^2}{n\left(1-\frac{k}{n}\right)^2} + \frac{\mathcal{K}(\rho,\pi)+\log\left(\frac{2}{\varepsilon}\right)}{\lambda} \end{array}\right\} \geq 1 - \varepsilon.$$

Proof Let us fix $\theta > 0$ and $\lambda > 0$, and apply the first inequality of Lemma 2. We have:

$$\mathbb{E}\left(e^{\lambda\left[R(\theta)-r_n(\theta)-\frac{\lambda\kappa^2}{n\left(1-\frac{k}{n}\right)^2}\right]}\right) \leq 1,$$

and we multiply this result by $\varepsilon/2$ and integrate it with respect to $\pi(\mathrm{d}\theta)$. Fubini's Theorem gives:

$$\mathbb{E}\left(\int e^{\lambda[R(\theta)-r_n(\theta)]-\frac{\lambda^2\kappa^2}{n\left(1-\frac{k}{n}\right)^2}-\log\left(\frac{2}{\varepsilon}\right)}\pi(\mathrm{d}\theta)\right)\leq\frac{\varepsilon}{2}.$$

We apply Lemma 3 and we get:

$$\mathbb{E}\left(e^{\sup_\rho\left\{\lambda\int[R(\theta)-r_n(\theta)]\rho(\mathrm{d}\theta)-\frac{\lambda^2\kappa^2}{n\left(1-\frac{k}{n}\right)^2}-\log\left(\frac{2}{\varepsilon}\right)-\mathcal{K}(\rho,\pi)\right\}}\right)\leq\frac{\varepsilon}{2}.$$

As $e^x\geq\mathbf{1}_{\mathbb{R}_+}(x)$, we have:

$$\mathbb{P}\left\{\sup_\rho\left\{\lambda\int\left[R(\theta)-r_n(\theta)\right]\rho(\mathrm{d}\theta)-\frac{\lambda^2\kappa^2}{n\left(1-\frac{k}{n}\right)^2}-\log\left(\frac{2}{\varepsilon}\right)-\mathcal{K}(\rho,\pi)\right\}\geq 0\right\}\leq\frac{\varepsilon}{2}.$$

Now, we follow the same proof again but starting with the second inequality of Lemma 2. We obtain:

$$\mathbb{P}\left\{\sup_\rho\left\{\lambda\int\left[r_n(\theta)-R(\theta)\right]\rho(\mathrm{d}\theta)-\frac{\lambda^2\kappa^2}{n\left(1-\frac{k}{n}\right)^2}-\log\left(\frac{2}{\varepsilon}\right)-\mathcal{K}(\rho,\pi)\right\}\geq 0\right\}\leq\frac{\varepsilon}{2}.$$

A union bound ends the proof. ■

A.2. Proof of Theorems 4.1 and 5.1

Proof [Proof of Theorem 4.1] Remark that **LipLoss**, **WeakDep** and **Lip** are satisfied. We apply the first inequality of Lemma 4. We obtain that with probability at least $1-\varepsilon$, we are on the event

$$\forall\rho\in\mathcal{M}_+^1(\Theta),\quad\begin{cases}\int R\mathrm{d}\rho\leq\int r_n\mathrm{d}\rho+\frac{\lambda\kappa^2}{n\left(1-\frac{k}{n}\right)^2}+\frac{\mathcal{K}(\rho,\pi)+\log\left(\frac{2}{\varepsilon}\right)}{\lambda}\\ \text{and}\\ \int r_n\mathrm{d}\rho\leq\int R\mathrm{d}\rho+\frac{\lambda\kappa^2}{n\left(1-\frac{k}{n}\right)^2}+\frac{\mathcal{K}(\rho,\pi)+\log\left(\frac{2}{\varepsilon}\right)}{\lambda}.\end{cases}\tag{4}$$

We apply the first inequality of (4) to $\hat{\rho}_\lambda(\mathrm{d}\theta)$. We obtain:

$$\mathbb{P}\left\{\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta)\leq\int r_n(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta)+\frac{\lambda\kappa^2}{n\left(1-\frac{k}{n}\right)^2}+\frac{1}{\lambda}\log\left(\frac{2}{\varepsilon}\right)+\frac{1}{\lambda}\mathcal{K}(\hat{\rho}_\lambda,\pi)\right\}\geq 1-\frac{\varepsilon}{2}.$$

According to Lemma 3 we have:

$$\int r_n(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta)+\frac{1}{\lambda}\mathcal{K}(\hat{\rho}_\lambda,\pi)=\inf_\rho\left(\int r(\theta)\rho(\mathrm{d}\theta)+\frac{1}{\lambda}\mathcal{K}(\rho,\pi)\right)$$

so we obtain

$$\mathbb{P}\left\{\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) \leq \inf_\rho \left[\int r_n(\theta)\rho(\mathrm{d}\theta) + \frac{\lambda\kappa^2}{n\left(1-\frac{k}{n}\right)^2} + \frac{\mathcal{K}(\rho,\pi)+\log\left(\frac{2}{\varepsilon}\right)}{\lambda}\right]\right\} \geq 1-\frac{\varepsilon}{2}. \quad (5)$$

We now want to bound from above $r(\theta)$ by $R(\theta)$. Applying the second inequality of (4) and plugging it into Inequality 5 gives

$$\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) \leq \inf_\rho \left[\int R\mathrm{d}\rho + \frac{2}{\lambda}\mathcal{K}(\rho,\pi) + \frac{2\lambda\kappa^2}{n\left(1-\frac{k}{n}\right)^2} + \frac{2}{\lambda}\log\left(\frac{2}{\varepsilon}\right)\right].$$

We end the proof by the remark that $\theta \mapsto R(\theta)$ is convex and so

$$\int R(\theta)\hat{\rho}_\lambda(\mathrm{d}\theta) \geq R\left(\int \theta\hat{\rho}_\lambda(\mathrm{d}\theta)\right) = R(\hat{\theta}_\lambda).$$

■

Proof [Proof of Theorem 5.1] We can apply Theorem 4.1 with $R = R_\tau$. We have, with probability at least $1-\varepsilon$,

$$R^\tau(\hat{\theta}^\tau_{B,\lambda}) \leq \inf_{\rho\in\mathcal{M}^1_+(\Theta)} \left[\int R^\tau\mathrm{d}\rho + \frac{2\lambda\kappa^2}{n\left(1-\frac{2}{n}\right)^2} + \frac{2\mathcal{K}(\rho,\pi)+2\log\left(\frac{2}{\varepsilon}\right)}{\lambda}\right].$$

Now, let us fix $\delta \in (0,1]$ and $\theta \in \Theta(B)$. We define the probability distribution $\rho_{\theta,\delta}$ as the uniform probability measure on the set:

$$\{T \in \mathbb{R}^4, \quad \|\theta - T\|_1 \leq \delta\}.$$

Note that $\rho_{\theta,\delta} \ll \pi_B$ as π_B is defined as uniform on $\Theta(B+1) \supset \Theta(B+\delta)$. Then:

$$R^\tau(\hat{\theta}^\tau_{B,\lambda}) \leq \inf_{\theta\in\Theta(B)}\inf_{\delta>0} \left[\int R^\tau\mathrm{d}\rho_{\theta,\delta} + \frac{2\lambda\kappa^2}{n\left(1-\frac{2}{n}\right)^2} + \frac{2\mathcal{K}(\rho_{\theta,\delta},\pi)+2\log\left(\frac{2}{\varepsilon}\right)}{\lambda}\right]. \quad (6)$$

Now, we have to compute or to upper-bound all the terms in the right-hand side of this inequality. First, note that:

$$\int R^\tau\mathrm{d}\rho_{\theta,\delta} = \int_{\{\|\theta-T\|_1\leq\delta\}} R^\tau(T)\mathrm{d}\rho_{\theta,\delta}(T) \leq R^\tau(\theta) + 2\mathcal{B}\delta\max(\tau,1-\tau) \leq R^\tau(\theta) + 2\mathcal{B}\delta. \quad (7)$$

Then, let us remark that:

$$\mathcal{K}(\rho_{\theta,\delta},\pi_B) = 3\log\left(\frac{B+1}{\delta}\right). \quad (8)$$

We plug (7) and (8) into (6) to obtain:

$$R^\tau(\hat{\theta}^\tau_{B,\lambda}) \leq \inf_\theta\inf_\delta \left\{R^\tau(\theta) + 2\left[\frac{\lambda\kappa^2}{n\left(1-\frac{2}{n}\right)^2} + \mathcal{B}\delta + \frac{3\log\left(\frac{B+1}{\delta}\right)+\log\left(\frac{2}{\varepsilon}\right)}{\lambda}\right]\right\}.$$

It can easily be seen that the minimum of the right-hand side w.r.t. δ is reached for $\delta = 3/(\mathcal{B}\lambda)$ (we will have to be careful with the choice of λ to ensure that $\delta < 1$), and so:

$$R^\tau(\hat{\theta}^\tau_{B,\lambda}) \leq \inf_\theta \left\{ R^\tau(\theta) + \frac{2\lambda\kappa^2}{n\left(1-\frac{2}{n}\right)^2} + \frac{6\log\left(\frac{(B+1)\mathcal{B}\lambda \mathrm{e}}{3}\right) + 2\log\left(\frac{2}{\varepsilon}\right)}{\lambda} \right\}.$$

We finally minimize the r.h.s. (roughly) with respect to λ to propose: $\lambda = \sqrt{3n}/\kappa$, this leads to:

$$R^\tau(\hat{\theta}^\tau_{B,\lambda}) \leq \inf_\theta \left\{ R^\tau(\theta) + \frac{2\sqrt{3}\kappa}{\sqrt{n}} \left[\frac{1}{\left(1-\frac{2}{n}\right)^2} + \log\left(\frac{(B+1)\mathcal{B}\mathrm{e}}{\kappa}\sqrt{\frac{n}{3}} \right) + \frac{\log\left(\frac{2}{\varepsilon}\right)}{3} \right] \right\}.$$

Remark that the condition $\delta < 1$ is satisfied as soon as $n > \kappa^2/(3\mathcal{B}^2)$. Also, when $n \geq 10$ we have:

$$\frac{1}{\left(1-\frac{2}{n}\right)^2} \leq \frac{25}{16}$$

and we can re-organize the terms to obtain:

$$R^\tau(\hat{\theta}^\tau_{B,\lambda}) \leq \inf_\theta \left\{ R^\tau(\theta) + \frac{2\sqrt{3}\kappa}{\sqrt{n}} \left[2.25 + \log\left(\frac{(B+1)\mathcal{B}\sqrt{n}}{\kappa} \right) + \frac{\log\left(\frac{1}{\varepsilon}\right)}{3} \right] \right\}.$$

■