Improved PAC-Bayesian Bounds for Linear Regression

Vera Shalaeva, Alireza Fakhrizadeh Esfahani, Pascal Germain, Mihaly Petreczky

▶ **To cite this version:**

Vera Shalaeva, Alireza Fakhrizadeh Esfahani, Pascal Germain, Mihaly Petreczky. Improved PAC-Bayesian Bounds for Linear Regression. AAAI 2020 - Thirty-Fourth AAAI Conference on Artificial Intelligence, Feb 2020, New York, United States. hal-02396556

HAL Id: hal-02396556
https://inria.hal.science/hal-02396556

Submitted on 6 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Improved PAC-Bayesian Bounds for Linear Regression

Vera Shalaeva[1], Alireza Fakhrizadeh Esfahani[2], Pascal Germain[3], and Mihaly Petreczky[4]

[1,3]The MODAL project-team, INRIA Lille Nord-Europe, France
{vera.shalaeva, pascal.germain}@inria.fr
[2,4]Univ. Lille, CNRS, Centrale Lille, UMR 9189,
CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, France
alireza.fakhrizadeh@univ-lille.fr, mihaly.petreczky@centralelille.fr

Abstract

In this paper, we improve the PAC-Bayesian error bound for linear regression derived in Germain et al. [10]. The improvements are two-fold. First, the proposed error bound is tighter, and converges to the generalization loss with a well-chosen temperature parameter. Second, the error bound also holds for training data that are not independently sampled. In particular, the error bound applies to certain time series generated by well-known classes of dynamical models, such as ARX models.

1 Introduction

When facing a machine learning problem, one must be careful to avoid overfitting the training dataset. Indeed, it is well known that minimizing the empirical prediction error is not sufficient to generalize to future observations. This is especially important for sensitive "AI" applications that are nowadays tackled by many industries (self-driving vehicles, health diagnosis, personality profiling, to name a few). Statistical learning theories study the generalization properties of learning algorithms. For the prediction problems, they provide guarantees on the "true" error of machine learning predictors (*i.e.*, the probability of erroneously predicting the labels of not seen yet samples).

The PAC-Bayesian learning theory, initiated by David McAllester (1999, 2003)—see Guedj [13] for a recent survey—, has the particularity of providing computable "non-vacuous" generalization bounds on popular machine learning algorithms, such as neural networks [8] and SVMs [3]. Moreover, as its name suggests, PAC-Bayesian framework bridges the *frequentist* Probably Approximately Correct theory and the *Bayesian* inference. This topic is namely discussed in Zhang [26], Grünwald [12], Alquier et al. [2], Germain et al. [10], Sheth and Khardon [23].

In this paper, we build on a result of Germain et al. [10], which analyses the Bayesian linear regression from a PAC-Bayesian perspective, leading to generalization bounds for the squared loss. We improve the preceding results in two directions. First, our new generalization bound is tighter than the one of Germain et al. [10], and converges to the generalization loss for proper parameters (see Section 3). Second, our result holds for training data that are not independently sampled (see Section 4). The latter result is directly applicable to the problem of learning dynamical systems from time series data, in particular, to learning ARX models. ARX models are a popular class of dynamical systems with a rich literature [17, 14] due to their relative simplicity and modelling power. Note that ARX models can be viewed as a simple yet non-trivial subclass of recurrent neural network regressions. For example, just like general recurrent neural networks, ARX models have a memory, *i.e.*, they are able to remember past input data.

Noteworthy, Alquier and Wintenberger [1] proposed PAC-Bayesian oracle inequalities to perform model selection on different time series (weakly dependent processes and causal Bernoulli shifts). Thus, their work is complementary to ours, as it relies on different assumptions and focuses on other types of error bounds.

2 PAC-Bayesian Learning

Let us consider a supervised learning setting, where a learning algorithm is given a training set $S = \{(x_i, y_i)\}_{i=1}^n$ of size n. Each pair (x_i, y_i) links a *description* $x_i \in \mathcal{X}$ to a *label* $y_i \in \mathcal{Y}$. Typically, the description is encoded by a real-valued vector ($\mathcal{X} \subseteq \mathbb{R}^d$), and the label is a scalar ($\mathcal{Y} \subseteq \mathbb{N}$ for classification problems, or $\mathcal{Y} \subseteq \mathbb{R}$ for regression ones). Given S, the learning algorithm returns a prediction function $f : \mathcal{X} \to \mathcal{Y}$, also referred to as a *hypothesis*. We restrict attention to prediction functions/hypotheses that are measurable. The "quality" of the predictor f is usually assessed through a measurable loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$—such as the zero-one loss $\ell(y, y') = \mathbf{1}_{y \neq y'}$ in classification context, or the squared loss $\ell(y, y') = (y - y')^2$ in regression context, by evaluating the *empirical loss*

$$\widehat{\mathcal{L}}^{\ell}(f)(S) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)\,, \quad \text{for any } S.$$

PAC Learning. When facing a machine learning problem, one wants to use f to predict the label $y \in \mathcal{Y}$ from a description $x \in \mathcal{X}$ that does not belong to the training set S. A good predictor "generalize to unseen data". This is the object of study of the Probably Approximately Correct (PAC) approach [25].

In order to study the statistical behavior of the average loss, we introduce the following statistical framework. We fix a probability space $(\Omega, \mathbf{P}, \mathbf{F})$, where $\mathbf{F}$ is a σ-algebra over Ω and $\mathbf{P}$ is a probability measure on $\mathbf{F}$, see for example Bilingsley [4] for the terminology. We assume that there exist random variables $\mathbf{X}_i : \Omega \to \mathcal{X}$, $\mathbf{Y}_i : \Omega \to \mathcal{Y}$, $i = 1, 2, \ldots,$ such that the description-label pairs $\{(x_i, y_i)\}_{i=1}^n$ are samples from the first n variables $\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n$ and there exists $\omega \in \Omega$ such that $x_i = \mathbf{X}_i(\omega)$, $y_i = \mathbf{Y}_i(\omega)$. Moreover, we assume that $\mathbf{X}_i, \mathbf{Y}_i$ are identically distributed, *i.e.* $\mathbf{E}\, g(\mathbf{X}_i, \mathbf{Y}_i)$ does not depend on i for any measurable function g.

Notation 1 (E). *We will use* $\mathbf{E}$ *to denote expected value with respect to the measure* $\mathbf{P}$.

That is, in the sequel, boldface symbols $\mathbf{P}$ and $\mathbf{E}$ will denote the probability and the corresponding mathematical expectation for the data generating distribution, and we will use boldface to denote random variables on the probability space $(\Omega, \mathbf{P}, \mathbf{F})$ and simple font for their samples. As we will see later on, we will also use a probability measure and the corresponding mathematical expectation defined on the space of predictors, which will be denoted differently.

The *generalization loss* of a predictor f is then defined as

$$\mathcal{L}^{\ell}(f) = \mathbf{E}\, \ell(f(\mathbf{X}_i), \mathbf{Y}_i)\,,$$

and it expresses the average error for "unseen data". It is then of interest to compare this error with the average empirical error, where the average is taken over all possible samples. To this end, we define the random variable

$$\widehat{\mathbf{L}^{\ell}}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(\mathbf{X}_i), \mathbf{Y}_i)\,,$$

i.e., for any sample $S = \{(x_i, y_i) = (\mathbf{X}_i(\omega), \mathbf{Y}_i(\omega)\}_{i=1}^n$ of $\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n$, $\widehat{\mathcal{L}}^{\ell}(f)(S) = \widehat{\mathbf{L}^{\ell}}(f)(\omega)$ is a sample of the random variable $\widehat{\mathbf{L}^{\ell}}(f)$. By slight abuse of terminology, we will refer to $\widehat{\mathbf{L}^{\ell}}(f)$ as the *empirical loss* too. PAC theories provide upper bounds of the form

$$\mathbf{P}\left(\mathcal{L}^{\ell}(f) \leq \widehat{\mathbf{L}^{\ell}}(f) + \varepsilon\right) \geq 1 - \delta\,,$$

where $\delta \in (0, 1]$ acts as a "confidence" parameter; the whole challenge of the PAC theories is to derive the mathematical expression of ε. Among the various approaches proposed to achieve this goal (reviewed in Shalev-Shwartz and Ben-David [22]), we can mention VC-dimension, sample compression, Rademacher's complexity, algorithmic stability, and the PAC-Bayesian theory. In the current work, we stand in the PAC-Bayesian learning framework.

PAC-Bayes. The PAC-Bayesian learning framework [18, 19] has the particularity of reconciling the PAC learning standpoint with the Bayesian paradigm. To be more precise, let us define a σ-algebra $\mathbb{F}$ on the set of predictors $\mathcal{F}$.[1]

Notation 2 ($\mathbb{E}_{f\sim\rho}$)**.** *If ρ is a probability distribution function on $\mathcal{F}$, in the sequel we denote by $\mathbb{E}_{f\sim\rho}$ the mathematical expectation with respect to the probability measure which corresponds to ρ.*

In the PAC-Bayesian paradigm, we consider a *prior* probability distribution π and a *posterior* probability distribution $\hat{\rho}$ over this σ-algebra. The prior must be chosen independently of the training set S, and the learning algorithm role is to output the posterior distribution, instead of a single predictor. The PAC-Bayesian bounds take the form[2]

$$\mathbf{P}\left(\mathop{\mathbb{E}}_{f\sim\hat{\rho}}\mathcal{L}^{\ell}(f) \leq \mathop{\mathbb{E}}_{f\sim\hat{\rho}}\widehat{\mathbf{L}}^{\ell}(f) + \varepsilon\right) \geq 1-\delta\,.$$

That is, in the PAC-Bayesian setting, the study focuses on the $\hat{\rho}$-averaged loss.[3] Typically, the term ε takes into account the prior via the Kullback-Leibler divergence:

$$\mathrm{KL}(\hat{\rho}\|\pi) = \mathop{\mathbb{E}}_{f\sim\hat{\rho}}\frac{\hat{\rho}(f)}{\pi(f)}\,.$$

Note that KL-diverence is defined only if $\hat{\rho}$ is absolutely continuous with respect to π.

In this paper, we build on the PAC-Bayesian theorem of Alquier et al. [2], which is also the starting point of Germain et al. [10] result improved in upcoming sections.

[1] Note that $\mathbb{F}$ is completely different from the σ-algebra $\mathbf{F}$ of the probability space for which the data generating random variables $\mathbf{X}_i, \mathbf{Y}_i$ are defined. This is not surprising, as the randomness of the data represents an assumption on the nature of the process which generates the data, while $\mathbb{F}$ will be used to define probability distributions, which express our subjective preferences for certain predictors, and which will be adjusted based on the observed data.

[2] Contrary to the example we give here, the relation between the expected empirical loss and the term ε might be non-linear. This is the case of the famous PAC-Bayes theorem of Seeger [21].

[3] The PAC-Bayesian literature also studies the stochastic *Gibbs predictor*, that perform each prediction on $x \in \mathcal{X}$ by drawing f according to $\hat{\rho}$ and outputting $f(x)$ (*e.g.*, Germain et al. [9]).

Theorem 3 (Alquier et al. [2])**.** *Given a set $\mathcal{F}$ of measurable hypotheses $\mathcal{X}\to\mathcal{Y}$, a measurable loss function $\ell : \mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$, a prior distribution π over $\mathcal{F}$, a $\delta\in(0,1]$, and a real number $\lambda>0$, $\forall\hat{\rho}$ over $\mathcal{F}$:*

$$\mathbf{P}\Bigg(\mathop{\mathbb{E}}_{f\sim\hat{\rho}}\mathcal{L}^{\ell}(f) \leq \mathop{\mathbb{E}}_{f\sim\hat{\rho}}\widehat{\mathbf{L}}^{\ell}(f) + \frac{1}{\lambda}\left[\mathrm{KL}(\hat{\rho}\|\pi) + \ln\frac{1}{\delta} + \Psi_{\ell,\pi}(\lambda,n)\right]\Bigg) \geq 1-\delta\,, \tag{1}$$

$$\textit{where}\qquad \Psi_{\ell,\pi}(\lambda,n) = \ln \mathop{\mathbb{E}}_{f\sim\pi}\mathbf{E}\, e^{\lambda\left(\mathcal{L}^{\ell}(f)-\widehat{\mathbf{L}}^{\ell}(f)\right)}\,. \tag{2}$$

For completeness, we provide a proof of Theorem 3 in Appendix A.1. This proof highlights that the result is obtained without assuming that the random variables $\mathbf{X}_i$, $\mathbf{Y}_i$ are mutually independent, unlike many "classical" PAC-Bayesian theorems. However, the *i.i.d.* assumption might be necessary to obtain a computable expression from Theorem 3, because it requires bounding the term $\Psi_{\ell,\pi}(\lambda,n)$ of Eq. (2). Indeed, since $\Psi_{\ell,\pi}(\lambda,n)$ relies on the unknown joint distribution of $\mathbf{X}_i, \mathbf{Y}_i$ for $i=1,2,\ldots$ its approximation needs assumption on the data.

Interestingly, given a training set S, obtaining the optimal posterior $\hat{\rho}^*$ minimizing the bound of Theorem 3, does not require evaluating $\Psi_{\ell,\pi}(\lambda,n)$, as this latter term is independent of both S and $\hat{\rho}$. Indeed, for fixed S, π and λ, minimizing the right-hand side of Eq. (1) amounts to solve $\hat{\rho}^* = \mathrm{argmin}_{\hat{\rho}}\left[\lambda\,\mathbb{E}_{f\sim\hat{\rho}}\widehat{\mathcal{L}}^{\ell}(f)(S) + \mathrm{KL}(\hat{\rho}\|\pi)\right]$, which is given by the *Gibbs posterior* [6, 2, 13]; for all $f\in\mathcal{F}$,

$$\hat{\rho}^*(f) = \frac{1}{Z}\pi(f)\exp\left(-\lambda\widehat{\mathcal{L}}^{\ell}(f)(S)\right), \tag{3}$$

where Z is a normalization constant. We refer to λ as a *temperature* parameter, as it controls the emphasis on the empirical loss minimization. The value of λ also directly impacts the value of the generalization bound, and the convergence properties of $\Psi_{\ell,\pi}(\lambda,n)$. In particular, if a non-negative loss is upper bounded by a value L (*i.e.*, $\ell(y,y')\in[0,L]$ for all $y,y'\in\mathcal{Y}$), and $\mathbf{X}_i, \mathbf{Y}_i$ are *i.i.d.*, we have, for any $f\in\mathcal{F}$ (we provide

the mathematical details in Appendix A):

$$\mathbf{E}\exp\left[\lambda\left(\mathcal{L}^{\ell}(f)-\widehat{\mathbf{L}^{\ell}}(f)\right)\right]\leq\exp\left[\frac{\lambda^2L^2}{8n}\right]. \quad (4)$$

Hence, we have $\Psi_{\ell,\pi}(\lambda,n)\leq\ln\mathbb{E}_{f\sim\pi}\,e^{\frac{\lambda^2L^2}{8n}}=\frac{\lambda^2L^2}{8n}$, from which the following result is obtained.

Corollary 4. *Given $\mathcal{F}$, π, a measurable and bounded loss function $\ell:\mathcal{Y}\times\mathcal{Y}\to[0,L]$, under* i.i.d. *observations, for $\delta\in(0,1]$ and $\lambda>0$, for any $\hat{\rho}$ over $\mathcal{F}$:*

$$\mathbf{P}\left(\mathop{\mathbb{E}}_{f\sim\hat{\rho}}\mathcal{L}^{\ell}(f)\leq\mathop{\mathbb{E}}_{f\sim\hat{\rho}}\widehat{\mathbf{L}^{\ell}}(f)+\frac{1}{\lambda}\left[\mathrm{KL}(\hat{\rho}\|\pi)+\ln\frac{1}{\delta}+\frac{\lambda^2L^2}{8n}\right]\right)\geq1-\delta\,.$$

Therefore, from Corollary 4, we obtain with $\lambda=\sqrt{n}$,

$$\mathop{\mathbb{E}}_{f\sim\hat{\rho}}\mathcal{L}^{\ell}(f)\leq\mathop{\mathbb{E}}_{f\sim\hat{\rho}}\widehat{\mathbf{L}^{\ell}}(f)+\frac{1}{\sqrt{n}}\left[\mathrm{KL}(\hat{\rho}\|\pi)+\ln\frac{1}{\delta}+\frac{L^2}{8}\right]. \quad (5)$$

In turn, with $\lambda=n$,

$$\mathop{\mathbb{E}}_{f\sim\hat{\rho}}\mathcal{L}^{\ell}(f)\leq\mathop{\mathbb{E}}_{f\sim\hat{\rho}}\widehat{\mathbf{L}^{\ell}}(f)+\frac{1}{n}\left[\mathrm{KL}(\hat{\rho}\|\pi)+\ln\frac{1}{\delta}\right]+\frac{L^2}{8}\,, \quad (6)$$

with probability at least 1-δ. The generalization bound given by Eq. (5) has the nice property that its value converges to the generalization loss (*i.e.*, the $\frac{1}{\sqrt{n}}\left[\,\cdot\,\right]$ term tends to 0 as n grows to infinity). However, the result of Eq. (6) does not converge: the bound suffers from an additive term $L^2/8$ even with large n.

Relation with Bayesian inference. Despite its lack of convergence, PAC-Bayesian theorem result of Eq. (6) is interesting for being closely linked to Bayesian inference. As discussed in Germain et al. [10] (based on earlier results of Zhang [26] and Grünwald [12]), maximizing the *Bayesian maximum likelihood* amounts to minimize the PAC-Bayes bound of Theorem 3 with $\lambda=n$, provided the Bayesian model parameters (typically denoted θ in the literature) are carefully reinterpreted as predictors (each θ is mapped to a regressor f_θ), and the considered loss function ℓ is the *negative log likelihood* (roughly[4], $\ell_{\mathrm{nll}}\big(y,f_\theta(x)\big)=-\ln p(y|x,\theta)$, where $p(y|x,\theta)$ is a Bayesian likelihood). That is, in these particular conditions, the posterior promoted by the celebrated *Bayesian rule* (*i.e.*, $p(\theta|X,Y)=\frac{p(\theta)p(Y|X,\theta)}{p(Y|X)}$, where $p(\theta)$ is the prior) aligns with the Gibbs posterior of Eq. (3).

Based on this observation, Germain et al. [10] extends Theorem 3 to Bayesian linear regression—for which the loss is unbounded–, as discussed in the next section.

3 Bounds for Bayesian Linear Regression

In the Bayesian literature [5, 20, ...], it is common to model a linear regression problem by assuming that $\mathcal{X}=\mathbb{R}^d$, $\mathcal{Y}=\mathbb{R}$. The input-output pairs $\mathbf{X}_i,\mathbf{Y}_i$ satisfy the following assumptions.

Assumption 5.

(a) *the inputs $\mathbf{X}_i$ are such that $\mathbf{X}_i\sim\mathcal{N}(\mathbf{0},\sigma_{\mathbf{x}}^2\mathbf{I})$, and $\mathbf{X}_i,\mathbf{X}_j$ are independent for $i\neq j$.*

(b) *the labels are given by $\mathbf{Y}_i=\mathbf{w}^*\cdot\mathbf{X}_i+\mathbf{e}_i$, where $\mathbf{e}_i\sim\mathcal{N}(0,\sigma_{\mathbf{e}}^2)$ and $\mathbf{e}_i,\mathbf{e}_j$ are independent for $i\neq j$.*

Here, we consider that $\sigma_{\mathbf{e}}>0$ is fixed, and we want to estimate the weight vector parameters $\mathbf{w}^*\in\mathbb{R}^d$. Thus, the likelihood function of $\mathbf{Y}_i$ given $\mathbf{X}_i$, $\mathbf{w}^*\in\mathbb{R}^d$ is given by

$$p(\mathbf{Y}_i|\mathbf{X}_i,\mathbf{w})=\mathcal{N}(\mathbf{Y}_i|\mathbf{w}\cdot\mathbf{X}_i,\sigma_{\mathbf{e}}^2)=(2\pi\sigma_{\mathbf{e}}^2)^{-\frac{1}{2}}e^{\left(\frac{1}{2\sigma_{\mathbf{e}}^2}(\mathbf{Y}_i-\mathbf{w}\cdot\mathbf{X}_i)^2\right)}.$$

Therefore, the corresponding negative log-likelihood loss function is proportional to the *squared loss* of a linear regressor $f_{\mathbf{w}}(\mathbf{x})=\mathbf{w}\cdot\mathbf{x}$:

$$\ell_{\mathrm{sqr}}(f_{\mathbf{w}}(\mathbf{X}_i),\mathbf{Y}_i)=(\mathbf{Y}_i-\mathbf{w}\cdot\mathbf{X}_i)^2\,. \quad (7)$$

[4] We omit several details here to concentrate on the general idea. We refer the reader to Germain et al. [10] for the whole picture.

3.1 Previous theorem

Considering a family of linear predictors, $\mathcal{F}_d=\{f_{\mathbf{w}}|\mathbf{w}\in\mathbb{R}^d\}$, Germain et al. [10] proposed a generalization bound for Bayesian linear regression under the following assumptions. To get a generalization bound for a squared loss in form of Eq. (1), one needs to compute the term $\Psi_{\ell_{\text{sqr}},\pi}(\lambda,n)$ or upper bound it. The following is the initial PAC-Bayesian bound for unbounded squared loss proposed by Germain et al. [10].

Theorem 6 (Germain et al. [10])**.** *Given $\mathcal{F}_d, \ell_{\text{sqr}}$, and δ defined above, given a prior distribution π over $\mathcal{F}_d$ which is a zero mean Gaussian with covariance $\sigma_\pi^2\mathbf{I}$, i.e., $\pi(f_{\mathbf{w}})=\mathcal{N}(\mathbf{w}|\mathbf{0},\sigma_\pi^2\,\mathbf{I})$, under Assumption 5, for constants $c\geq 2\sigma_{\mathbf{x}}^2\sigma_\pi^2$, and $\lambda\in(0,\frac{1}{c})$, for any posterior distribution $\hat{\rho}$ over $\mathcal{F}_d$:*

$$\mathbf{P}\left(\mathop{\mathbb{E}}_{f_{\mathbf{w}}\sim\hat{\rho}}\mathcal{L}^{\ell}(f_{\mathbf{w}})\leq\mathop{\mathbb{E}}_{f_{\mathbf{w}}\sim\hat{\rho}}\widehat{\mathbf{L}}^{\ell}(f_{\mathbf{w}})+\frac{1}{\lambda}\left[\mathrm{KL}(\hat{\rho}\|\pi)+\ln\frac{1}{\delta}\right]+\frac{\frac{1}{2}(d+\|\mathbf{w}^*\|^2)c+(1-\lambda c)\sigma_{\mathbf{e}}^2}{1-\lambda c}\right)\geq 1-\delta\,. \tag{8}$$

Theorem 6 expresses the result with λ stated explicitly, while Germain et al. [10]—see Appendix A.4 therein—were focusing on the case $\lambda=n$. Here, we observe that the bound does not converge; regardless the choice of λ, the last term of Eq. (8) is not negligible.

Note that PAC-Bayesian guarantees for similar Bayesian models has also been proposed by other authors, under different set of assumptions, either bounded loss [23] or non-random inputs [7].

3.2 Improved theorem

The first contribution of this paper is an improvement of Theorem 6.

Theorem 7. *Given $\mathcal{F}_d, \ell_{\text{sqr}}$ defined above, under Assumption 5, for any $\delta\in(0,1]$, $\lambda>0$, for any prior distribution π over $\mathcal{F}_d$, and for any posterior distribution $\hat{\rho}$ over $\mathcal{F}_d$, the following holds:*

$$\mathbf{P}\left(\mathop{\mathbb{E}}_{f_{\mathbf{w}}\sim\hat{\rho}}\mathcal{L}^{\ell}(f_{\mathbf{w}})\leq\mathop{\mathbb{E}}_{f_{\mathbf{w}}\sim\hat{\rho}}\widehat{\mathbf{L}}^{\ell}(f_{\mathbf{w}})+\frac{1}{\lambda}\left[\mathrm{KL}(\hat{\rho}\|\pi)+\ln\frac{1}{\delta}+\Psi_{\ell_{\text{sqr}},\pi}(\lambda,n)\right]\right)\geq 1-\delta\,, \tag{9}$$

where

$$\Psi_{\ell_{\text{sqr}},\pi}(\lambda,n)=\ln\mathop{\mathbb{E}}_{f_{\mathbf{w}}\sim\pi}\frac{\exp\left(\lambda v_{\mathbf{w}}\right)}{\left(1+\frac{\lambda v_{\mathbf{w}}}{\frac{n}{2}}\right)^{\frac{n}{2}}} \tag{10}$$

$$\leq\ln\mathop{\mathbb{E}}_{f_{\mathbf{w}}\sim\pi}\exp\left(\frac{\lambda^2 v_{\mathbf{w}}^2}{\frac{n}{2}}\right), \tag{11}$$

and $\quad v_{\mathbf{w}}=\sigma_{\mathbf{x}}^2\parallel\mathbf{w}^*-\mathbf{w}\parallel_2^2+\sigma_{\mathbf{e}}^2\,.$

Proof. We get the complexity term in form of Eq. (10) by simplifying the general form given in Eq. (2), and using assumptions on inputs and a prior distribution.

$$\begin{aligned}
&\Psi_{\ell_{\text{sqr}},\pi}(\lambda,n)\\
&=\ln\mathop{\mathbb{E}}_{f_{\mathbf{w}}\sim\pi}\mathbf{E}\exp\left[\lambda\left(\mathcal{L}^{\ell}(f_{\mathbf{w}})-\widehat{\mathbf{L}}^{\ell}(f_{\mathbf{w}})\right)\right]\\
&=\ln\mathop{\mathbb{E}}_{f_{\mathbf{w}}\sim\pi}\exp\left(\lambda\mathcal{L}^{\ell}(f_{\mathbf{w}})\right)\mathbf{E}\exp\left(-\lambda\widehat{\mathbf{L}}^{\ell}(f_{\mathbf{w}})\right)\\
&=\ln\mathop{\mathbb{E}}_{f_{\mathbf{w}}\sim\pi}\exp\left(\lambda\mathcal{L}^{\ell}(f_{\mathbf{w}})\right)\underbrace{\mathbf{E}\exp\left(-\tfrac{\lambda}{n}\sum_{i=1}^{n}(\mathbf{Y}_i-\mathbf{w}\cdot\mathbf{X}_i)^2\right)}_{(\clubsuit)}.
\end{aligned}$$

Note that random variable $\mathbf{Y}_i-\mathbf{w}\cdot\mathbf{X}_i=(\mathbf{w}^*-\mathbf{w})\mathbf{X}_i+\mathbf{e}_i$ has zero expectation

$$\mathbf{E}(\mathbf{Y}_i-\mathbf{w}\cdot\mathbf{X}_i)=(\mathbf{w}^*-\mathbf{w})\,\mathbf{E}\,\mathbf{X}_i+\mathbf{E}\,\mathbf{e}_i=0\,,$$

and its second moment, denoted $v_{\mathbf{w}}$, which by definition equals $\mathcal{L}^{\ell}(f_{\mathbf{w}})$, is

$$\begin{aligned}
\mathcal{L}^{\ell}(f_{\mathbf{w}})&=\mathbf{E}(\mathbf{Y}_i-\mathbf{w}\cdot\mathbf{X}_i)^2\\
&=\mathbf{E}\left[(\mathbf{w}^*-\mathbf{w})\mathbf{X}_i^T\mathbf{X}_i(\mathbf{w}^*-\mathbf{w})\right]\\
&\quad+\mathbf{E}\left[2(\mathbf{w}^*-\mathbf{w})\mathbf{X}_i\mathbf{e}_i+\mathbf{e}_i^2\right]\\
&=\sigma_{\mathbf{x}}^2\parallel\mathbf{w}^*-\mathbf{w}\parallel_2^2+\sigma_{\mathbf{e}}^2\,.
\end{aligned}$$

Hence, $\frac{\mathbf{Y}_i - \mathbf{w}\cdot\mathbf{X}_i}{\sqrt{v_\mathbf{w}}} \sim \mathcal{N}(0,1)$ is a normalized random variable, and its squared sum follows Chi-squared distribution law. Note that the term (♣) of the function $\Psi_{\ell_{\text{sqr}},\pi}(\lambda, n)$ in the form

$$\mathbf{E}\exp\left(-\frac{\lambda v_\mathbf{w}}{n}\sum_{i=1}^{n}\left(\frac{\mathbf{Y}_i - \mathbf{w}\cdot\mathbf{X}_i}{\sqrt{v_\mathbf{w}}}\right)^2\right),$$

corresponds to the moment generating function (MGF) of a Chi-squared distribution, *i.e.* $(1-2t)^{\frac{n}{2}}$ with $t = -\frac{\lambda v_\mathbf{w}}{n}$.

By replacing the term (♣) by Chi-Squared MGF and $\mathcal{L}^{\,\ell}(f_\mathbf{w})$ by $v_\mathbf{w}$, we get the complexity term in form of Eq. (10).

Eq. (11) is obtained by lower bounding the denominator of Eq. (10) by using the inequality $(1+\frac{a}{b})^b > \exp(\frac{ab}{a+b})$, for $a, b > 0$:

$$\Psi_{\ell,\pi}(\lambda, n) = \ln \mathop{\mathbb{E}}_{f_\mathbf{w}\sim\pi} \frac{\exp(\lambda v_\mathbf{w})}{\exp\left(\frac{\lambda v_\mathbf{w}\frac{n}{2}}{\lambda v_\mathbf{w} + \frac{n}{2}}\right)}$$

$$= \ln \mathop{\mathbb{E}}_{f_\mathbf{w}\sim\pi} \exp\left(\frac{\lambda^2 v_\mathbf{w}^2}{\lambda v_\mathbf{w} + \frac{n}{2}}\right) \leq \ln \mathop{\mathbb{E}}_{f_\mathbf{w}\sim\pi} \exp\left(\frac{\lambda^2 v_\mathbf{w}^2}{\frac{n}{2}}\right). \quad \square$$

We are interested in the convergence properties of the right side of Eq. (9). This will highly depend on the choice of λ.

- If λ is fixed and does not depend on n, and the latter approaches to ∞, we get

  $$\mathop{\mathbb{E}}_{f_\mathbf{w}\sim\hat{\rho}} \mathcal{L}^{\,\ell}(f_\mathbf{w}) \leq \mathop{\mathbb{E}}_{f_\mathbf{w}\sim\hat{\rho}} \widehat{\mathbf{L}}^{\ell}(f_\mathbf{w}) + \frac{1}{\lambda}\left[\mathrm{KL}(\hat{\rho}\|\pi) + \ln\frac{1}{\delta}\right]$$

  The term $\Psi_{\ell,\pi}(\lambda, n)$ amounts to 0, since the expression under the expectation of Eq. (10) will converge to 1 due to the fact that

  $$\exp(\lambda v_\mathbf{w}) = \lim_{n\to\infty}\left(1 + \frac{\lambda v_\mathbf{w}}{\frac{n}{2}}\right)^{\frac{n}{2}}.$$

  Hence, an empirical error converges to the generalization error with sufficiently large value of the parameter λ, and small divergence between prior and posterior distributions.

- If λ is considered as a function of n, then we can obtain convergence of the right side of the Eq. (9) to the left side with a well-chosen temperature parameter. Let λ be $n^{\frac{1}{d}}\ln(\frac{1}{\delta})$, then from Eq. (9) and (11), we have

  $$\mathop{\mathbb{E}}_{f_\mathbf{w}\sim\hat{\rho}} \mathcal{L}^{\,\ell}(f_\mathbf{w}) \leq \mathop{\mathbb{E}}_{f_\mathbf{w}\sim\hat{\rho}} \widehat{\mathbf{L}}^{\ell}(f_\mathbf{w}) + \frac{\mathrm{KL}(\hat{\rho}\|\pi)}{n^{\frac{1}{d}}\ln(\frac{1}{\delta})}$$
  $$+ \frac{1}{n^{\frac{1}{d}}} + \frac{1}{n^{\frac{1}{d}}}\ln \mathop{\mathbb{E}}_{f_\mathbf{w}\sim\pi} \exp\left(\frac{2n^{\frac{2}{d}}\ln(\frac{1}{\delta})^2 v_\mathbf{w}^2}{n}\right).$$

  If the amount of training examples $n\to\infty$, then the bound converges to generalization loss.

3.3 Theorems comparison

The new bound given by Theorem 7 is always tighter than the previous one of Theorem 6. Indeed, the fraction of Eq. (10) is upper bounded by its numerator $\exp(\lambda v_\mathbf{w})$. The latter is the exact same expression as in the derivation of Germain et al. [10] (Supp. Material A4, p.11, line 4), which lead us to the prior bound shown in Eq. (8). Moreover, the new bound converges to zero for well-chosen temperature parameter λ as the number of training observations goes to infinity. For these reasons, the result of Theorem 7 is strictly stronger than those of Theorem 6.

4 Extension to the non *i.i.d.* case

In this section we will study the case when the observed data are no longer sampled independently from the underlying distribution.

4.1 The learning problem and its relationship with time series

We consider the same learning problem as in Section 3, but we modify Assumption 5 by no longer assuming that $\mathbf{X}_i$ are *i.i.d.* random variables, more precisely, we assume the following:

Assumption 8. *We assume Part (b) of Assumption 5 and we assume that* $\mathbf{X}_i \sim \mathcal{N}(\mathbf{0}, Q_x)$ *for some positive definite matrix* $Q_x > 0$.

It then follows that $\mathbf{Y}_i$ are also identically distributed, $\mathbf{Y}_i \sim \mathcal{N}(0, \sigma_y^2)$, where

$$\sigma_y^2 = \mathbf{w}^{*T} Q_x \mathbf{w}^* + \sigma_{\mathbf{e}}^2 I \,.$$

Note that from the assumption that $\mathbf{X}_i$ are identically distributed it follows that $\mathcal{L}^{\ell}(f_{\mathbf{w}})$ does not depend on i and

$$\mathcal{L}^{\ell}(f_{\mathbf{w}}) = (\mathbf{w}^* - \mathbf{w})^T Q_x (\mathbf{w}^* - \mathbf{w}) + \sigma_{\mathbf{e}}^2 \,.$$

A particular instance of the learning problem above is the problem of learning ARX models, which is a well-studied problem in control theory and econometrics [17, 14]. For the sake of simplicity, we will deal only with the scalar input, scalar output case. Consider stationary zero mean discrete-time stochastic processes $\mathbf{y}_t, \mathbf{u}_t$, $t \in \mathbb{Z}$, $t > 0$.

Assume that there exist real numbers $\{a_i, b_i\}_{i=1}^k$ and a stochastic process $\mathbf{e}_t$ such that

$$\mathbf{y}_t = \sum_{i=1}^{k} a_i \mathbf{y}_{t-i} + \sum_{i=1}^{k} b_i \mathbf{u}_{t-i} + \mathbf{e}_t \,, \qquad (12)$$

where $\mathbf{e}_t$ is assumed to be an *i.i.d.* sequence of random variables such that $\mathbf{e}_t \in \mathcal{N}(0, \sigma^2)$ and $\mathbf{e}_t$ is uncorrelated with $\mathbf{y}_s, \mathbf{u}_s$ for $s < t$. Consider the polynomial $\mathbf{a}(z) = z^k - \sum_{i=1}^k a_i z^{k-i-1}$. If $\mathbf{a}(z)$ has all its complex roots inside the unit disc, and $\mathbf{u}_t$ is a stationary, then it is well known [14] that there $\mathbf{y}_t$ is the unique stationary process which satisfies Eq. (12).

Moreover, if $\mathbf{u}_t$ is a jointly Gaussian process, then the $\mathbf{y}_t$ and the parameters $(\{a_i, b_i\}_{i=1}^k, \sigma^2)$ together with the joint distribution of $\mathbf{u}_t$ determine the distribution of $\mathbf{y}_t$ uniquely [14].

Intuitively, the learning problem is to try to compute a prediction $\hat{\mathbf{y}}_t$ of $\mathbf{y}_t$ based on past values $\{\mathbf{y}_{t-l}, \mathbf{u}_{t-l}\}_{l=1}^{\infty}$ of the input and output processes. In the literature [14, 17] one typically would like to minimize the prediction error $\mathbf{E}[(\mathbf{y}_t - \hat{\mathbf{y}}_t)^2]$ In principle, this generalization error may depend on t. However, if we assume that the predictor f uses only the last L observations and it is of the form $\hat{\mathbf{y}}_t = \sum_{i=1}^L \hat{a}_i \mathbf{y}_{t-i} + \sum_{i=1}^L \hat{b}_i \mathbf{u}_{t-i}$, then by stationarity of $\mathbf{y}_t, \mathbf{u}_t$, $t \in \mathbb{Z}$, the predictor will not depend on t. Furthermore, if $\mathbf{y}_t, \mathbf{u}_t$ come from an ARX model Eq. (12) and they are Gaussian, then it can be shown [14] under some mild assumptions that the best possible predictor is necessarily of the above form with $L = k$, and in fact, we should take $\hat{a}_i = a_i$, $\hat{b}_i = b_i$, $i = 1, \ldots, k$, and in this case the generalization error $\mathbf{E}[(\mathbf{y}_t - \hat{\mathbf{y}}_t)^2] = \sigma^2$. For this reason, in the literature [17, 14] the learning problem is often formulated as the problem of estimating the parameters of the true model (Eq. (12)). It is well known that for ARX models, the latter point of view is essentially equivalent to finding the predictor for which the generalization error $\mathbf{E}[(\mathbf{y}_t - \hat{\mathbf{y}}_t)^2]$ is the smallest.

This allows us to recast the learning problem into our framework for linear regression as follows. For every $i = 1, 2, \ldots$, define

$$\begin{aligned} &\mathbf{Y}_i = \mathbf{y}_{i+k} \,, \\ &\mathbf{X}_i = \begin{bmatrix} \mathbf{y}_{i+k-1} & \ldots & \mathbf{y}_{i-1} & \mathbf{u}_{i+k-1} & \ldots & \mathbf{u}_{i-1} \end{bmatrix}^T , \\ &\mathbf{w}^* = \begin{bmatrix} a_1 & \ldots & a_k & b_1 & \ldots & b_k \end{bmatrix} , \mathbf{e}_t = \mathbf{e}_{i+k} \,. \end{aligned}$$

It then follows that $\mathbf{X}_i, \mathbf{Y}_i, \mathbf{e}_i$ satisfy Assumption 8.

4.2 PAC-Bayesian approach for linear regression with possibly dependent observations

In this section we discuss the extension of Theorem 7 to the case when the observations are not independently sampled.

Although Theorem 3 holds even when $(\mathbf{X}_i, \mathbf{Y}_i)$ are not *i.i.d.*, the proof of Theorem 7 relies heavily on the independence of $\mathbf{X}_i$, $i = 1, \ldots, n$. More precisely, let us recall from the proof of Theorem 7 the empirical prediction error variables

$$\mathbf{Z}_{\mathbf{w},i} = \mathbf{Y}_i - \mathbf{w} \cdot \mathbf{X}_i = (\mathbf{w}^* - \mathbf{w}) \cdot \mathbf{X}_i + \mathbf{e}_i \,. \qquad (13)$$

The proof of Theorem 7 relied on $\mathbf{Z}_{\mathbf{w},i}$, $i = 1, \ldots, n$ being independent and identically distributed zero mean Gaussian random variables. In our case, the variables $\mathbf{Z}_{\mathbf{w},i}$ are still zero mean Gaussian variables which are identically distributed, but they no longer independent. Hence, we have to take into account the joint distribution of $\{\mathbf{Z}_{\mathbf{w},i}\}_{i=1}^n$, which in turn depends on the joint distribution of $\{\mathbf{X}_i\}_{i=1}^n$.

In order to deal with this phenomenon, we will define the *joint covariance matrix* $Q_{X,n}$ of the random variable $\mathbf{X}_{1:n} = \left[\mathbf{X}_1^T, \quad \ldots, \quad \mathbf{X}_n^T\right]$ as follows:

$$Q_{X,n} = \mathbf{E}[\mathbf{X}_{1:n}\mathbf{X}_{1:n}^T]\,,$$

i.e., the (i,j)th $d\times d$ block matrix element of $Q_{X,n}$ is $\mathbf{E}[\mathbf{X}_i\mathbf{X}_j^T]$. We can then formulate the following bound.

Theorem 9. *Let ρ_n be the minimal eigenvalue of $Q_{X,n}$ and assume that $\rho_n>0$. Under Assumption 8, for any prior distribution π over $\mathcal{F}_d$, any $\delta \in (0,1]$, any real number $\lambda > 0$, and for any posterior distribution $\hat\rho$ over $\mathcal{F}_d$, we have*

$$\mathbf{P}\left(\mathop{\mathbb{E}}_{f_{\mathbf{w}}\sim\hat\rho}\mathcal{L}^{\ell}(f_{\mathbf{w}}) \le \mathop{\mathbb{E}}_{f_{\mathbf{w}}\sim\hat\rho}\widehat{\mathbf{L}^{\ell}}(f_{\mathbf{w}}) + \frac{1}{\lambda}\left[\mathrm{KL}(\hat\rho\|\pi) + \ln\frac{1}{\delta} + \hat\Psi_{\ell,\pi}(\lambda,n)\right]\right) \ge 1-\delta\,, \tag{14}$$

where

$$\hat\Psi_{\ell,\pi}(\lambda,n) = \ln \mathop{\mathbb{E}}_{f_{\mathbf{w}}\sim\pi}\frac{\exp(\lambda v_{\mathbf{w}})}{\left(1+\frac{\lambda\rho_{n,\mathbf{w}}}{\frac{n}{2}}\right)^{\frac{n}{2}}} \tag{15}$$

$$\le \ln \mathop{\mathbb{E}}_{f_{\mathbf{w}}\sim\pi}\exp\left(\frac{\lambda^2 v_{\mathbf{w}}\rho_{n,\mathbf{w}}}{\frac{n}{2}} + \lambda(v_{\mathbf{w}}-\rho_{n,\mathbf{w}})\right), \tag{16}$$

with $v_{\mathbf{w}} = (\mathbf{w}^*-\mathbf{w})^T Q_x(\mathbf{w}^*-\mathbf{w}) + \sigma_{\mathbf{e}}^2$,
and $\rho_{n,\mathbf{w}} = \rho_n(\mathbf{w}^*-\mathbf{w})^T(\mathbf{w}^*-\mathbf{w}) + \sigma_{\mathbf{e}}^2$.

Remark 10 (Comparison with the *i.i.d.* case)**.** *If $\mathbf{X}_i$, $i=1,2,\ldots$, are independent and $Q_{\mathbf{X}} = \sigma_{\mathbf{x}}^2 I_d$, then $Q_{X,n}$ is diagonal, with the diagonal elements being $\sigma_{\mathbf{x}}^2$. In this case, $\rho_n = \sigma_{\mathbf{x}}^2$ and $\rho_{n,\mathbf{w}} = v_{wb}$ and hence the statement of Theorem 9 boils down to that of Theorem 7.*

Before presenting the proof of Theorem 9 some discussion is in order.

Recall that one of the advantages of the error bound of Theorem 7 was that it converged to zero as $n\to\infty$. The question arises if this is the case for the error bound of Theorem 9. In order to answer this question we need to investigate the dependence on n of the smallest eigenvalue ρ_n of the covariance matrix $Q_{X,n}$, since ρ_n is used in the error bound of Theorem 9. To this end, note that $Q_{X,n}$ is a positive semi-definite matrix, and hence by the properties of minimal eigenvalues of positive semi-definite matrices [11] $\rho_n r^T r \le r^T Q_{X,n} r$. From Södeström and Stoica [24](Chapter 5, page 135) it follows that $\rho_n \ge \rho_{n-1}$, *i.e.*, ρ_n is a monotonically increasing sequence. In particular, as $\rho_n \le \rho_1$ and $Q_{X,1} = Q_X$, $\rho_1\|\mathbf{w}-\mathbf{w}^*\|_2^2 \le (\mathbf{w}-\mathbf{w}^*)^T Q_x(\mathbf{w}-\mathbf{w}^*)$ and hence $\rho_{n,\mathbf{w}} \le v_{\mathbf{w}}$. This means that the right-hand side of Eq. (15) is not smaller than the right-hand side of Eq. (10), and Eq. (16) is not smaller than Eq. (11).

That is, the error bounds of Theorem 9 are not smaller than those of Theorem 7. Moreover, $\rho_n \ge 0$ since it is an eigenvalue of the positive definite matrix $Q_{X,n}$. In particular, $\rho_* = \lim_{n\to\infty}\rho_n = \inf_n \rho_n$ exists.

Then we get the following corollary of Theorem 9, by noticing that since $\rho_n \ge \rho_*$, $\frac{\exp(\lambda v_{\mathbf{w}})}{\left(1+\frac{\lambda\rho_{n,\mathbf{w}}}{\frac{n}{2}}\right)^{\frac{n}{2}}} \le \frac{\exp(\lambda v_{\mathbf{w}})}{\left(1+\frac{\lambda\rho_{*,\mathbf{w}}}{\frac{n}{2}}\right)^{\frac{n}{2}}}$.

Corollary 11. *Assume $\rho_* > 0$. For any prior π over $\mathcal{F}_d$, any $\delta\in(0,1]$, and any $\lambda>0$, and any $\hat\rho$ over $\mathcal{F}_d$, Eq. (14) remains true if we replace $\hat\Psi_{\ell,\pi}$ by $\tilde\Psi_{\ell,\pi}$, where*

$$\hat\Psi_{\ell,\pi}(\lambda,n) \le \tilde\Psi_{\ell,\pi}(\lambda,n) = \ln \mathop{\mathbb{E}}_{f_{\mathbf{w}}\sim\pi}\frac{\exp(\lambda v_{\mathbf{w}})}{\left(1+\frac{\lambda\rho_{*,\mathbf{w}}}{\frac{n}{2}}\right)^{\frac{n}{2}}}\,,$$

with $v_{\mathbf{w}} = (\mathbf{w}^*-\mathbf{w})^T Q_x(\mathbf{w}^*-\mathbf{w}) + \sigma_{\mathbf{e}}^2$ *and* $\rho_{*,\mathbf{w}} = \rho_*(\mathbf{w}^*-\mathbf{w})^T(\mathbf{w}^*-\mathbf{w}) + \sigma_{\mathbf{e}}^2$.

Corollary 11 gives a PAC-Bayesian bound, asymptotic behavior of which is easy to study. Indeed, since $1+\frac{\lambda\rho_{*,\mathbf{w}}}{n/2}$ increases with n and it converges to $\exp(\lambda\rho_{*,\mathbf{w}})$ as $n\to\infty$, the error bound $\tilde\Psi_{\ell,\pi}(\lambda,n)$ will decrease with n and

$$\lim_{n\to\infty}\tilde\Psi_{\ell,\pi,}(\lambda,n) = \ln \mathop{\mathbb{E}}_{f_{\mathbf{w}}\sim\pi}\exp\left(\lambda(v_{\mathbf{w}}-\rho_{*,\mathbf{w}})\right). \tag{17}$$

That is, contrary to the *i.i.d.* case in Theorem 9, PAC-Bayesian error bound of Corollary 11 decreases with n, but it will not converge to 0, rather, it will be bounded from above by the right-hand side of

Eq. (17). Note that $v_{\mathbf{w}} - \rho_{*,\mathbf{w}} = (\mathbf{w} - \mathbf{w}^*)^T(Q_x - \rho_* I_d)(\mathbf{w} - \mathbf{w}^*)$. The latter is a monotonically increasing function of $Q_x - \rho_* I_d$: the smaller this difference is, the close the right-hand side of Eq. (17) to zero. The difference $Q_x - \rho_* I_d$ is zero in the *i.i.d.* case, and can be seen as a kind of measure of the degree of dependence of $\mathbf{X}_i$, $i = 1, 2, \ldots,$.

Note that Theorem 9 and Corollary 11 are meaningful only for $\rho_n > 0$ and $\rho_* > 0$.

For time series assumption that $\rho_* > 0$ is equivalent to $Q_{x,n} > mI_{nd}$ for all n for some m. This property is mild modification of the well-known property of *informativity of the data set* $\{\mathbf{y}_t, \mathbf{u}_t\}_{t=1}^{\infty}$ [17]. This can be seen by an easy modification of the argument of Södeström and Stoica [24](Chapter 5, page 122, proof of Property 1). In turn, informativity of the data set is a standard assumption made in the literature [17], and it is required for learning ARX models. Note that under mild assumptions on $\mathbf{u}_t$, from Ljung [17][Theorem 2.3] it then follows that the $\widehat{\mathbf{L}^{\ell}}(f_{\mathbf{w}}) \to \mathcal{L}^{\ell}(f_{\mathbf{w}})$ as $n \to \infty$ with probability one. That is, even though the law of large numbers does not apply in this case, we still know that the empirical loss converges to the generalization error as $n \to \infty$.

Proof of Theorem 9. The proof follows the same lines as that of Theorem 9. From Theorem 3 it follows that

$$\mathbf{P}\left(\mathop{\mathbb{E}}_{f_{\mathbf{w}}\sim\hat{\rho}} \mathcal{L}^{\ell}(f_{\mathbf{w}}) \leq \mathop{\mathbb{E}}_{f_{\mathbf{w}}\sim\hat{\rho}} \widehat{\mathbf{L}^{\ell}}(f_{\mathbf{w}}) + \frac{1}{\lambda}\left[\mathrm{KL}(\hat{\rho}\|\pi) + \ln\frac{1}{\delta} + \Psi_{\ell,\pi}(\lambda, n)\right]\right) \geq 1 - \delta\,. \tag{18}$$

Consider the random variable $Z_{\mathbf{w},i}$ defined in Eq. (13). Just like in the proof of Theorem 7,

$$\Psi_{\ell,\pi}(\lambda, n) = \ln \mathop{\mathbb{E}}_{f_{\mathbf{w}}\sim\pi} \mathbf{E} \exp\left[\lambda\left(\mathcal{L}^{\ell}(f_{\mathbf{w}}) - \widehat{\mathbf{L}^{\ell}}(f_{\mathbf{w}})\right)\right] \tag{19}$$

$$= \ln \mathop{\mathbb{E}}_{f_{\mathbf{w}}\sim\pi} \left\{\exp\left(\lambda\mathcal{L}^{\ell}(f_{\mathbf{w}})\right) \mathbf{E} \exp\left(-\frac{\lambda}{n}\sum_{i=1}^{n} \mathbf{Z}_{\mathbf{w},i}^2\right)\right\}.$$

And, it can be shown that $\mathbf{Z}_{\mathbf{w},i}$ is zero mean Gaussian with variance $\mathbf{E}[\mathbf{Z}_{\mathbf{w},i}^2] = v_{\mathbf{w}}$. In the proof of Theorem 7 we used the fact that under its assumptions $\{\mathbf{Z}_{\mathbf{w},i}\}_{i=1}^n$ were mutually independent and identically distributed and hence $\frac{\lambda v_{\mathbf{w}}}{n}\sum_{i=1}^n \frac{\mathbf{Z}_{\mathbf{w},i}}{v_{\mathbf{w}}^2}$ had χ^2 distribution. In our case, $\mathbf{Z}_{\mathbf{w},i}$ are not independent. In order to get around this issue, we define the random variable $\mathbf{Z}_{\mathbf{w},1:n}$ and its covariance matrix $Q_{\mathbf{w},n}$:

$$\mathbf{Z}_{\mathbf{w},1:n} = \begin{bmatrix}\mathbf{Z}_{\mathbf{w},i} & \ldots & \mathbf{Z}_{\mathbf{w},n}\end{bmatrix}^T,$$
$$Q_{\mathbf{w},n} = \mathbf{E}[\mathbf{Z}_{\mathbf{w},1:n}\mathbf{Z}_{\mathbf{w},1:n}^T]\,.$$

It is easy to see that $Q_{\mathbf{w},n} = D_{\mathbf{w}}^T Q_{X,n} D_{\mathbf{w}} + \sigma_{\mathbf{e}}^2 I_n$, where

$$D_{\mathbf{w}} = \mathrm{diag}(\underbrace{(\mathbf{w} - \mathbf{w}^*)I_d, \ldots, (\mathbf{w} - \mathbf{w}^*)I_d}_{n \text{ times}})\,.$$

Notice that $r^T Q_{X,n} r \geq \rho_n r^T r$ for all $r \in \mathbb{R}^d$ by Golub and Van Loan [11]. Then, for any $z \in \mathbb{R}^n$, by taking $r = D_{\mathbf{w}} z$, it follows that

$$\begin{aligned} z^T Q_{\mathbf{w},n} z &= (D_{\mathbf{w}} z)^T Q_{X,n}(D_{\mathbf{w}} z) + \sigma_{\mathbf{e}}^2 z^T z \\ &\geq \rho_n (D_{\mathbf{w}} z)^T (D_{\mathbf{w}} z) + \sigma_{\mathbf{e}}^2 z^T z = \rho_{n,\mathbf{w}}\,. \end{aligned} \tag{20}$$

where we used that $\|D_{\mathbf{w}} z\|_2^2 = \|\mathbf{w} - \mathbf{w}^*\|_2^2 \|z\|_2^2$. Define

$$\mathbf{S} = Q_{\mathbf{w},n}^{-1/2} \mathbf{Z}_{\mathbf{w},1:n}\,.$$

and let $\mathbf{S}_i$ be the ith entry of $\mathbf{S}$, *i.e.*, $\mathbf{S} = \begin{bmatrix}\mathbf{S}_1 & \ldots & \mathbf{S}_n\end{bmatrix}^T$. Then from Eq. (20) it follows that

$$\begin{aligned} \sum_{i=1}^n \mathbf{Z}_{\mathbf{w},i}^2 &= \mathbf{Z}_{\mathbf{w},1:n}^T Q_{\mathbf{w},n}^{-1/2} Q_{\mathbf{w},n} Q_{\mathbf{w},n}^{-1/2} \mathbf{Z}_{\mathbf{w},1:n} \\ &= \mathbf{S}^T Q_{\mathbf{w},n} \mathbf{S} \geq \mathbf{S}^T \mathbf{S} \rho_{n,\mathbf{w}} = \Big(\sum_{i=1}^n \mathbf{S}_i^2\Big)\rho_{n,\mathbf{w}}\,. \end{aligned}$$

It then follows that

$$\exp\left(-\frac{\lambda}{n}\sum_{i=1}^n \mathbf{Z}_{\mathbf{w},i}^2\right) \leq \exp\left(-\frac{\lambda}{n}\rho_{n,\mathbf{w}}\sum_{i=1}^n \mathbf{S}_i^2\right). \tag{21}$$

Notice now that $\mathbf{S}$ is Gaussian and zero mean, with covariance $\mathbf{E}[\mathbf{S}\mathbf{S}^T] = Q_{\mathbf{w},n}^{-1/2}\,\mathbf{E}[\mathbf{Z}_{\mathbf{w},1:n}\mathbf{Z}_{\mathbf{w},1:n}^T]Q_{\mathbf{w},n}^{-1/2} = I_n$. That is, the random variables $\mathbf{S}_i$ are normally

distributed and $\mathbf{S}_i, \mathbf{S}_j$ are independent, and therefore $\sum_{i=1}^n \mathbf{S}_i^2$ has χ^2 distribution. Hence,

$$\mathbf{E}\left[\exp\left(-\frac{\lambda\rho_{n,\mathbf{w}}}{n}\sum_{i=1}^{n}\mathbf{S}_i^2\right)\right] = \frac{1}{(1+\frac{\lambda\rho_{n,\mathbf{w}}}{\frac{n}{2}})^{\frac{n}{2}}}\,.$$

Combining this with Eq. (21) and (19), Eq. (18) implies Eq. (15). By using the inequality $\left(1+\frac{a}{b}\right)^b > e^{\frac{ab}{a+b}}$ for $a, b>0$ with $a=\lambda\rho_{n,\mathbf{w}}$ and $b=\frac{n}{2}$, Eq. (16) follows from Eq. (15). □

4.3 Related works

Note that PAC bounds for learning time series has been explored in the literature by Kuznetsov and Mohri (2017, 2018). Their approach is based on covering numbers and Rademacher complexity instead of PAC-Bayes analysis, but in contrast to the current paper, Kuznetsov and Mohri's work allows for non-stationary time series.

Alquier and Wintenberger [1] includes a PAC-Bayesian analysis in their model selection procedure for time series. Among other differences, they provide oracle inequalities type of bounds, whereas our analysis provides generalization bounds relying on the empirical loss.

5 Conclusion

We have presented an improved PAC-Bayesian error bound for linear regression and extended this error bound to the case of non *i.i.d.* observations. Thus, the obtained bound applies to the learning problem of time series using ARX models, which can be viewed as a simple yet non-trivial subclass of recurrent neural network regressions. For this reason, we are hopeful that the results of Section 4 could potentially lead to PAC-Bayesian bounds for recurrent neural networks.

6 Acknowledgement

This work is funded in part by CNRS project PEPS Blanc INS2I 2019 BayesReaForRNN, in part by CPER Data project, co-financed by European Union, European Regional Development Fund (ERDF), French State and the French Region of Hauts-de-France, and in part by the French project APRIORI ANR-18-CE23-0015.

A Mathematical details

A.1 Proof of Theorem 3

Proof. The PAC-Bayesian theorem is based on the following *Donsker-Varadhan's change of measure*.

For any measurable function $\phi : \mathcal{F} \to \mathbb{R}$, we have $\mathbb{E}_{f\sim\hat{\rho}}\,\phi(f) \le \mathrm{KL}(\hat{\rho}\|\pi) + \ln\left(\mathbb{E}_{f\sim\pi}\, e^{\phi(f)}\right)$. Thus, with $\phi(f)=\lambda\big(\mathcal{L}^{\ell}(f)-\widehat{\mathbf{L}}^{\ell}(f)\big)$, we obtain $\forall\,\hat{\rho}$ on $\mathcal{F}$:

$$\mathop{\mathbb{E}}_{f\sim\hat{\rho}} \lambda\left(\mathcal{L}^{\ell}(f) - \widehat{\mathbf{L}}^{\ell}(f)\right) \le \mathrm{KL}(\hat{\rho}\|\pi) + \ln\left(\mathop{\mathbb{E}}_{f\sim\pi} e^{\lambda\left(\mathcal{L}^{\ell}(f)-\widehat{\mathbf{L}}^{\ell}(f)\right)}\right). \quad (22)$$

Let's consider the random variable $\xi = \mathop{\mathbb{E}}_{f\sim\pi} e^{\lambda\left(\mathcal{L}^{\ell}(f)-\widehat{\mathbf{L}}^{\ell}(f)\right)}$. By the Markov inequality, we have

$$\mathbf{P}\left(\xi \le \frac{1}{\delta}\,\mathbf{E}\,\xi\right) \ge 1-\delta\,,$$

which, combined with Eq. (22), gives

$$\mathbf{P}\left(\mathop{\mathbb{E}}_{f\sim\hat{\rho}} \lambda\left(\mathcal{L}^{\ell}(f) - \widehat{\mathbf{L}}^{\ell}(f)\right) \le \mathrm{KL}(\hat{\rho}\|\pi) + \ln\left(\frac{1}{\delta}\,\mathbf{E}\,\xi\right)\right) \ge 1-\delta\,.$$

By rearranging the terms of above equation, we obtain the following equivalent form of the statement of the theorem:

$$\mathbf{P}\left(\mathop{\mathbb{E}}_{f\sim\hat{\rho}} \mathcal{L}^{\ell}(f) \le \mathop{\mathbb{E}}_{f\sim\hat{\rho}} \widehat{\mathbf{L}}^{\ell}(f) + \frac{1}{\lambda}\left[\mathrm{KL}(\hat{\rho}\|\pi) + \ln\left(\frac{1}{\delta}\,\mathbf{E}\,\xi\right)\right]\right) \ge 1-\delta\,.$$

To see that the inequality above is equivalent to the statement of the theorem, note that by Fubini's theorem,

$$\mathbf{E}\,\xi = \mathbf{E}\mathop{\mathbb{E}}_{f\sim\hat{\rho}} e^{\lambda\left(\mathcal{L}^{\ell}(f)-\widehat{\mathbf{L}}^{\ell}(f)\right)} = \mathop{\mathbb{E}}_{f\sim\hat{\rho}} \mathbf{E}\, e^{\lambda\left(\mathcal{L}^{\ell}(f)-\widehat{\mathbf{L}}^{\ell}(f)\right)},$$

and hence $\ln \mathbf{E}\,\xi = \Psi_{\ell,\pi}(\lambda, n)$. Moreover, $\ln(\frac{1}{\delta}\,\mathbf{E}\,\xi) = \ln\frac{1}{\delta} + \ln \mathbf{E}\,\xi$. □

A.2 Details leading to Eq. (4)

For any $f \in \mathcal{F}$:

$$\begin{aligned}
&\mathbf{E} \exp\left[\lambda\left(\mathcal{L}^{\ell}(f) - \widehat{\mathbf{L}^{\ell}}(f)\right)\right] \\
&= \mathbf{E}\, e^{\frac{\lambda}{n}\sum_{i=1}^{n}(\mathbf{E}\,\ell(f(\mathbf{X}_k),\mathbf{Y}_k) - \ell(f(\mathbf{X}_i),\mathbf{Y}_i))} \\
&= \mathbf{E} \prod_{i=1}^{n} e^{\frac{\lambda}{n}(\mathbf{E}\,\ell(f(\mathbf{X}_k),\mathbf{Y}_k) - \ell(f(\mathbf{X}_i),\mathbf{Y}_i))} \\
(\mathbf{X}_i, \mathbf{Y}_i \text{ i.i.d.}) &= \prod_{i=1}^{n} \mathbf{E}\, e^{\frac{\lambda}{n}(\mathbf{E}\,\ell(f(\mathbf{X}_k),\mathbf{Y}_k) - \ell(f(\mathbf{X}_i),\mathbf{Y}_i))} \\
(\text{Hoeff.}) &\leq \prod_{i=1}^{n} \exp\left[\frac{\lambda^2 L^2}{8n^2}\right] \\
&= \exp\left[\frac{\lambda^2 L^2}{8n}\right],
\end{aligned}$$

where the line (Hoeff.) is obtained from Hoeffding's lemma on the random variable $\left(\mathcal{L}^{\ell}(f) - \ell(f(\mathbf{X}_i), \mathbf{Y}_i)\right) \in [-\mathcal{L}^{\ell}(f), L - \mathcal{L}^{\ell}(f)]$, which has an expected value of zero.

References

[1] Pierre Alquier and Olivier Wintenberger. Model selection for weakly dependent time series forecasting. *Bernoulli*, 18(3):883–913, 2012.

[2] Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *JMLR*, 17(239): 1–41, 2016.

[3] Amiran Ambroladze, Emilio Parrado-Hernández, and John Shawe-Taylor. Tighter PAC-Bayes bounds. In *NIPS*, 2006.

[4] P. Bilingsley. *Probability and measure*. Wiley, 1986. ISBN 0471804789.

[5] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[6] Olivier Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*, volume 56. Inst. of Mathematical Statistic, 2007.

[7] Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61, 2008.

[8] Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *UAI*. AUAI Press, 2017.

[9] Pascal Germain, Alexandre Lacasse, Francois Laviolette, Mario Marchand, and Jean-Francis Roy. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. *JMLR*, 16, 2015.

[10] Pascal Germain, Francis R. Bach, Alexandre Lacoste, and Simon Lacoste-Julien. PAC-Bayesian theory meets bayesian inference. In *NIPS*, pages 1876–1884, 2016.

[11] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 2013. ISBN 9781421407944.

[12] Peter Grünwald. The safe Bayesian - learning the learning rate via the mixability gap. In *ALT*, 2012.

[13] Benjamin Guedj. A Primer on PAC-Bayesian Learning. *arXiv preprint arXiv:1901.05353*, 2019.

[14] E.J. Hannan and M. Deistler. *The Statistical Theory of Linear Systems*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1988. ISBN 9781611972191.

[15] Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, 2017.

[16] Vitaly Kuznetsov and Mehryar Mohri. Theory and algorithms for forecasting time series. *arXiv preprint arXiv:1803.05814*, 2018.

[17] L. Ljung. *System Identification: Theory for the user (2nd Ed.)*. PTR Prentice Hall., Upper Saddle River, USA, 1999.

[18] David McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

[19] David McAllester. Simplified PAC-Bayesian margin bounds. In *COLT*, pages 203–215, 2003.

[20] Kevin P Murphy. *Machine learning: a probabilistic perspective*. The MIT Press, 2012.

[21] Matthias Seeger. PAC-Bayesian generalization bounds for Gaussian processes. *JMLR*, 3:233–269, 2002.

[22] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014.

[23] Rishit Sheth and Roni Khardon. Excess risk bounds for the bayes risk using variational inference in latent gaussian models. In *NIPS*, pages 5151–5161, 2017.

[24] Torsten Södeström and Petre Stoica. *System Identification*. Prentice Hall, 1989. ISBN 0138812365.

[25] Leslie G. Valiant. A theory of the learnable. *Commununications of the ACM*, 27(11):1134–1142, 1984.

[26] Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Trans. Information Theory*, 52(4):1307–1321, 2006.