PAC-Bayesian Error Bound, via Rényi Divergence, for a Class of Linear Time-Invariant State-Space Models

Mihály Petreczky

CNRS, École Centrale Lille, University of Lille, CRIStAL

Joint work with D. Eringis, R. Wisniewski, J. Leth, Zh. Tan. from Department of Electronic Systems, Aalborg University, Denmark.

Learning problem for time-series

- $\mathbb{X} = \mathbb{R}^{n_u}$ – input-space, $\mathbb{Y} = \mathbb{R}^{n_y}$ output space.
- $\mathbf{x}(t) \in \mathbb{X}$ – input process, $\mathbf{y}(t) \in \mathbb{Y}$ – output process, $t \in \mathbb{Z}$ – time axis.
- Hypotheses:
  $\mathcal{H} \subseteq \{$ functions of the form $h : \bigcup_{k=1}^{\infty} (\mathbb{X} \times \mathbb{Y})^k \to \mathbb{Y}\}$.
  $h(\{\mathbf{x}(s), \mathbf{y}(s)\}_{s=0}^{t-1})$ – prediction of output $\mathbf{y}(t)$ based on past values of the inputs and outputs.
- Quadratic loss function: $\ell : \mathbb{Y} \times \mathbb{Y} \to [0, +\infty)$,

$$\ell(y, y') = \|y - y'\|_2^2$$

  $\ell(\mathbf{y}(t), h(\{\mathbf{x}(s), \mathbf{y}(s)\}_{s=0}^{t-1}))$ difference between the output predicted by $h \in \mathcal{H}$ and true output.
- True error for a hypothesis h: long-term prediction error

$$\mathcal{L}(h) = \lim_{t \to \infty} \mathbf{E}[\ell(\mathbf{y}(t), h(\{\mathbf{x}(s), \mathbf{y}(s)\}_{s=0}^{t-1}))]$$

Learning problem for time-series

Learning problem: based on samples of $\{(\mathbf{x}(t), \mathbf{y}(t))\}_{t=1}^{N}$ find $h_\star \in \mathcal{H}$ such that $\mathcal{L}(h_\star)$ is small.

Solution:

1. define the empirical error for hypothesis h:

$$\hat{\mathcal{L}}_N(h) = \frac{1}{N} \sum_{t=0}^{N-1} \ell(\mathbf{y}(t), h(\{\mathbf{x}(s), \mathbf{y}(s)\}_{s=0}^{t-1})).$$

2. let $h_\star$ be such that $\left(\hat{\mathcal{L}}_N(h_\star) + \text{ regularization term }\right)$ is small.

Question:

What can we say about the true error $\mathcal{L}(h_\star)$?

Assumptions: hypothesis class

Hypotheses are parametrised by $\theta \in \Theta$, i.e., $\theta \mapsto h_\theta \in \mathcal{H}$, and are realized by stable LTI (linear time-invariant) dynamical systems

$$\begin{aligned} \hat{\mathbf{s}}(t+1) &= \hat{A}_\theta \hat{\mathbf{s}}(t) + \hat{B}_\theta \mathbf{x}(t) + \hat{K}_\theta \mathbf{y}(t), \quad \hat{\mathbf{s}}(0) = 0 \\ h_\theta(\{\mathbf{x}(\tau), \mathbf{y}(\tau)\}_{\tau=0}^{t-1}) &:= \hat{C}_\theta \hat{\mathbf{s}}(t) \end{aligned} \tag{1}$$

$h_\theta(\{\mathbf{x}(\tau), \mathbf{y}(\tau)\}_{\tau=0}^{t-1})$ – the prediction of the current label $\mathbf{y}(t)$ based on the past values of inputs and labels.

Recurrent neural networks (RNNs) with a linear activation function, and classical autoregressive models (ARX, ARMAX) are included.

Assumptions: data generator

- Data is generated by a stable linear dynamical system driven by a sub-Gaussian zero mean i.i.d. noise $\mathbf{e}_g$,

$$\begin{aligned} \mathbf{s}_g(t+1) &= A_g \mathbf{s}_g(t) + B_g \mathbf{x}(t) + K_g \mathbf{e}_g(t) \\ \mathbf{y}(t) &= C_g \mathbf{s}_g(t) + \mathbf{e}_g(t) \end{aligned} \tag{2}$$

Data generator $\implies$ hypothesis $h_{\theta_{true}}$ with minimal true loss

$$\begin{aligned} &\mathbf{s}_g(t+1) = (A_g - K_g C_g)\mathbf{s}_g(t) + B_g \mathbf{x}(t) + K_g \mathbf{y}(t) \\ &h_{\theta_{true}}(\{\mathbf{x}(\tau), \mathbf{y}(\tau)\}_{\tau=0}^{t-1}) = C_g \mathbf{s}_g(t) \end{aligned}$$

Minimizing empirical loss $\implies$ approximating the data generator.

PAC-Bayesian bounds with Rényi divergence

Theorem (Main contribution)

For all $\delta \in [0, 0.5)$, *for any prior probability density* π *on* Θ

$$\mathbf{P}\Big(\forall \rho \text{ probability density on } \Theta,\ \rho \ll \pi : \underbrace{E_{\theta\sim\rho}\mathcal{L}(\theta)}_{\text{true error}} \leq \underbrace{E_{\theta\sim\rho}\hat{\mathcal{L}}_N(\theta)}_{\text{empirical error}} + r_N(\pi,\rho,\delta)\Big) > 1 - 2\delta \quad (3)$$

$$r_N(\pi,\rho,\delta) \triangleq \frac{K}{\sqrt{\delta N}} \bar{\mathcal{D}}_2(\rho|\pi)\left[G_1 + \frac{4}{\sqrt{N}} G_2\right]$$

- **P** – probability on data.
- $E_{\theta\sim\rho}$ – expectation over all parameters (hypotheses) using density ρ.
- $\bar{\mathcal{D}}_2(\rho|\pi) \triangleq \left(E_{\theta\sim\pi}\left(\frac{\rho(\theta)}{\pi(\theta)}\right)^2\right)^{\frac{1}{2}}$ – Rényi divergence, i.e., a sort of distance, between the posterior ρ and the prior π.

Discussion on the bound

- $O\left(\frac{1}{\sqrt{N}}\right)$ bound, converges to zero
- G_1, G_2 – quadratic in the ℓ_1-norm of the data generator (A_g, B_g, K_g, C_g) and of the ℓ_1 the hypothesis class $(A_\theta, B_\theta, K_\theta, C_\theta, D_\theta)$
- K depends on the variance of the noise of the data generator.
- ℓ_1-norms depend on the stability (robustness) of the hypotheses and data generator.
  More stability $\implies$ smaller generalization gap.
- Dependence on $\frac{1}{\sqrt{\delta}}$ instead of $\ln(\frac{1}{\delta})$.

Learning using the PAC-Bayesian bound

(1) find a posterior $\rho = \hat{\rho}_N$ which minimizes

$$E_{\theta\sim\rho}[\hat{\mathcal{L}}_N(\theta)] + \frac{K}{\sqrt{\delta N}}\bar{\mathcal{D}}_2(\rho|\pi)\left[G_1 + \frac{4}{\sqrt{N}}G_2\right]$$

(2) $\theta_\star$ is one of the following:

- $\theta_\star$ random sample from $\hat{\rho}_N$, or
- most likely model, i.e. $\theta_\star = \sup_{\theta\in\Theta}\hat{\rho}_N(\theta)$, or
- $\theta_\star$ is the mean model: $E_{\theta\sim\hat{\rho}_N}\theta$.

PAC-Bayesian bound (3) $\implies$ high probability bounds

- on the generalization gap $\mathcal{L}(\theta_\star) - \hat{\mathcal{L}}_N(\theta_\star)$
- on the parameter estimation error $\theta_\star - \theta_{true}$.

Numerical example

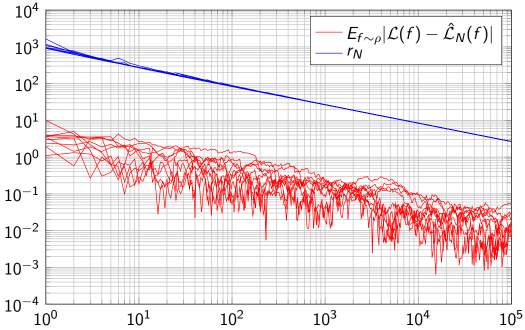

Figure: Results of a synthetic example, the case of $\mathbf{w} = \mathbf{u}$, 10 different realisations of data, $r_N = r_N(\rho, \pi)$

- The data is generated by (2) with 2 states, such that $n_u = n_y = 1$, $\mathbf{e}_g(t) \sim \mathcal{N}(0, Q_e)$,
- hypotheses: linear systems with two states.

Novelty, revelance, & conclusions

- PAC-Bayesian bounds for i.i.d. data using KL divergence [1] is a classical topic. Using Rényi divergence [2, 3] allows to cover additional cases.
- We have extended prior results to dynamical systems in state-space form and non i.i.d. data.
  Our results extend the bounds for autoregressive models from [4, 2].
- Stability is the key: it makes the data weakly dependent.
- Future research: evaluate the bounds on realistic parametrizations and data sets.

Alquier, P.
User-friendly introduction to PAC-Bayes bounds. 2021.
arXiv:2110.11216.

Alquier, P. and Guedj, B. Simpler PAC-Bayesian Bounds for Hostile Data. Machine Learning, 107(5):887–902, 2018.

Bégin, L., Germain, P., Laviolette, F., and Roy, J.-F.
Pac-bayesian bounds based on the Rényi divergence. Artificial Intelligence and Statistics, 435–444. PMLR, 2016.

Alquier, P.; and Wintenberger, O.
Model selection for weakly dependent time series forecasting.
Bernoulli, 18(3): 883 – 913, 2012.