



PAC-Bayesian Learning

An overview of theory, algorithms and current trends

Benjamin Guedj

<https://bguedj.github.io>
Inria Lille - Nord Europe

<https://bguedj.github.io> - 6PAC

⁶PAC: Making PAC Learning great-again

1. Active
2. Sequential
3. Structure-aware
4. Efficient
5. Ideal
6. Safe



Peter Grünwald
CWI, co-PI



Benjamin Guedj
Inria, co-PI



Emilie Kaufmann
Inria



Wouter Koolen
CWI

A mathematical theory of learning: towards AI

{Statistical, Machine} learning: devise automatic procedures to infer general rules from data.

Field of study about computers' ability to learn without being explicitly programmed (Arthur Samuel, 1959).

In the (rather not so?) long term: mimic the inductive functioning of the human brain to develop an artificial intelligence.

Big data / data science (somewhat annoying) hype: extremely dynamic field at the crossroads of Computer Science, Optimization and Statistics.

A hot topic at CWI and Inria in general and in Lille in particular: we are hiring!

Learning in a nutshell

Learning in a nutshell

Collect data $\mathcal{D}_n = (\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^n$ distributed as a random variable $(\mathbf{X}, \mathbf{Y}) \in \mathcal{X} \times \mathcal{Y}$. Data may be incomplete (unsupervised setting, missing input), collected sequentially / actively, etc.

Learning in a nutshell

Collect data $\mathcal{D}_n = (\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^n$ distributed as a random variable $(\mathbf{X}, \mathbf{Y}) \in \mathcal{X} \times \mathcal{Y}$. Data may be incomplete (unsupervised setting, missing input), collected sequentially / actively, etc.

Goal: use \mathcal{D}_n to build up $\hat{\phi}$ such that $\hat{\phi}(\mathbf{X}) \approx \mathbf{Y}$. Learning is to be able to generalize!

Learning in a nutshell

Collect data $\mathcal{D}_n = (\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^n$ distributed as a random variable $(\mathbf{X}, \mathbf{Y}) \in \mathcal{X} \times \mathcal{Y}$. Data may be incomplete (unsupervised setting, missing input), collected sequentially / actively, etc.

Goal: use \mathcal{D}_n to build up $\hat{\phi}$ such that $\hat{\phi}(\mathbf{X}) \approx \mathbf{Y}$. Learning is to be able to generalize!

For some loss function $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$, let

$$R: \hat{\phi} \mapsto \mathbb{E} \ell \left(\hat{\phi}(\mathbf{X}), \mathbf{Y} \right) \quad \text{and} \quad r_n: \hat{\phi} \mapsto \frac{1}{n} \sum_{i=1}^n \ell \left(\hat{\phi}(\mathbf{X}_i), \mathbf{Y}_i \right)$$

denote the risk (unknown) and empirical risk (known), respectively.

Typical goals: probabilistic bounds on R , algorithm based on r_n .
Under classical assumptions, $r_n \rightarrow R$.

Bayesian learning in a nutshell

Bayesian learning in a nutshell

Let \mathcal{F} be a set of candidate functions equipped with a probability measure π (prior). Let f be the (known) density of the (assumed) distribution of (\mathbf{X}, \mathbf{Y}) , and define the posterior

$$\hat{\rho}(\cdot) \propto f(\mathbf{X}, \mathbf{Y}|\cdot)\pi(\cdot).$$

Model-based learning (may be parametric or nonparametric).

Bayesian learning in a nutshell

Let \mathcal{F} be a set of candidate functions equipped with a probability measure π (prior). Let f be the (known) density of the (assumed) distribution of (\mathbf{X}, \mathbf{Y}) , and define the posterior

$$\hat{\rho}(\cdot) \propto f(\mathbf{X}, \mathbf{Y}|\cdot)\pi(\cdot).$$

Model-based learning (may be parametric or nonparametric).

- ▶ MAP $\hat{\phi} \in \arg \max_{\phi \in \mathcal{F}} \hat{\rho}(\phi)$.
- ▶ Mean $\hat{\phi} = \mathbb{E}_{\hat{\rho}} \phi = \int_{\mathcal{F}} \phi \hat{\rho}(d\phi)$.
- ▶ Realization $\hat{\phi} \sim \hat{\rho}$.
- ▶ ...

Quasi-Bayesian learning in a nutshell

A.k.a generalized Bayes.

Quasi-Bayesian learning in a nutshell

A.k.a generalized Bayes.

Let \mathcal{F} be a set of candidates functions equipped with a probability measure π (prior). Let $\lambda > 0$, and define a quasi-posterior

$$\hat{\rho}_\lambda(\cdot) \propto \exp(-\lambda r_n(\cdot)) \pi(\cdot).$$

Model-free learning!

- ▶ MAQP $\hat{\phi}_\lambda \in \arg \max_{\phi \in \mathcal{F}} \hat{\rho}_\lambda(\phi)$.
- ▶ Mean $\hat{\phi}_\lambda = \mathbb{E}_{\hat{\rho}_\lambda} \phi = \int_{\mathcal{F}} \phi \hat{\rho}_\lambda(d\phi)$.
- ▶ Realization $\hat{\phi}_\lambda \sim \hat{\rho}_\lambda$.
- ▶ ...

Why quasi-Bayes?

Why quasi-Bayes?

One justification (there are others). Let \mathcal{K} denote the Kullback-Leibler divergence

$$\mathcal{K}(\rho, \pi) = \begin{cases} \int_{\mathcal{F}} \log \left(\frac{d\rho}{d\pi} \right) d\rho & \text{when } \rho \ll \pi, \\ +\infty & \text{otherwise.} \end{cases}$$

With the classical quadratic loss $\ell: (a, b) \mapsto (a - b)^2$,


$$\hat{\rho}_\lambda \in \arg \inf_{\rho \ll \pi} \left\{ \int_{\mathcal{F}} r_n(\phi) \rho(d\phi) + \frac{\mathcal{K}(\rho, \pi)}{\lambda} \right\}.$$

Statistical aggregation revisited

Statistical aggregation revisited

$$\begin{aligned}\hat{\phi}_\lambda &:= \mathbb{E}_{\hat{\rho}_\lambda} \phi = \int_{\mathcal{F}} \phi \hat{\rho}_\lambda(d\phi) \\ &= \int_{\mathcal{F}} \phi \exp(-\lambda r_n(\phi)) \pi(d\phi) \\ &= \sum_{i=1}^{\#\mathcal{F}} \underbrace{\frac{\exp(-\lambda r_n(\phi_i)) \pi(\phi_i)}{\sum_{j=1}^{\#\mathcal{F}} \exp(-\lambda r_n(\phi_j)) \pi(\phi_j)}}_{\omega_{\lambda,i}} \phi_i, \quad \text{if } |\mathcal{F}| < +\infty.\end{aligned}$$

This is the celebrated exponentially weighted aggregate (EWA).

 G. (2013). Agrégation d'estimateurs et de classificateurs : théorie et méthodes, *Ph.D. thesis, Université Pierre & Marie Curie*

PAC learning in a nutshell

PAC learning in a nutshell

Probably Approximately Correct (PAC) oracle inequalities /
generalization bounds and empirical bounds.

 Valiant (1984). A theory of the learnable, *Communications of the ACM*

PAC learning in a nutshell

Probably Approximately Correct (PAC) oracle inequalities / generalization bounds and empirical bounds.

▣ Valiant (1984). A theory of the learnable, *Communications of the ACM*

Let $\hat{\phi}$ be a learning algorithm. For any $\epsilon > 0$,

$$\mathbb{P} \left(R(\hat{\phi}) \leq \spadesuit \left\{ r_n(\hat{\phi}) + \Delta(n, d, \phi, \epsilon) \right\} \right) \geq 1 - \epsilon,$$

$$\mathbb{P} \left(R(\hat{\phi}) - R^* \leq \spadesuit \inf_{\phi \in \mathcal{F}} \left\{ R(\phi) - R^* + \Delta(n, d, \phi, \epsilon) \right\} \right) \geq 1 - \epsilon,$$

where $\spadesuit \geq 1$ and $R^* = \inf_{\phi \in \mathcal{F}} R(\phi)$.

Key argument: concentration inequalities (e.g., Bernstein) + duality formula (Csiszár, Catoni).

The PAC-Bayesian theory

The PAC-Bayesian theory

...consists in producing PAC bounds for quasi-Bayesian learning algorithms.

While PAC bounds focus on estimators $\hat{\theta}_n$ that are obtained as functionals of the sample and for which the risk R is small, the PAC-Bayesian approach studies an aggregation distribution $\hat{\rho}_n$ that depends on the sample, for which $\int R(\theta)\hat{\rho}_n(d\theta)$ is small.

The PAC-Bayesian theory

...consists in producing PAC bounds for quasi-Bayesian learning algorithms.

While PAC bounds focus on estimators $\hat{\theta}_n$ that are obtained as functionals of the sample and for which the risk R is small, the PAC-Bayesian approach studies an aggregation distribution $\hat{\rho}_n$ that depends on the sample, for which $\int R(\theta)\hat{\rho}_n(d\theta)$ is small.

▣ Shawe-Taylor and Williamson (1997). A PAC analysis of a Bayes estimator, *COLT*

▣ McAllester (1998). Some PAC-Bayesian theorems, *COLT*

▣ McAllester (1999). PAC-Bayesian model averaging, *COLT*

▣ Catoni (2004). *Statistical Learning Theory and Stochastic Optimization*, Springer

▣ Audibert (2004). Une approche PAC-bayésienne de la théorie statistique de l'apprentissage, *Ph.D. thesis*, Université Pierre & Marie Curie

▣ Catoni (2007). PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning, IMS

▣ Dalalyan and Tsybakov (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity, *Machine Learning*

A flexible and powerful framework (1/2)

A flexible and powerful framework (1/2)

- Alquier and Wintenberger (2012). Model selection for weakly dependent time series forecasting, *Bernoulli*
- Seldin, Laviolette, Cesa-Bianchi, Shawe-Taylor and Auer (2012). PAC-Bayesian inequalities for martingales, *IEEE Transactions on Information Theory*
- Alquier and Biau (2013). Sparse Single-Index Model, *Journal of Machine Learning Research*
- G. and Alquier (2013). PAC-Bayesian Estimation and Prediction in Sparse Additive Models, *Electronic Journal of Statistics*
- Alquier and G. (2017). An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization, *Mathematical Methods of Statistics*
- Dziugaite and Roy (2017). Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data, *UAI*
- Dziugaite and Roy (2018). Data-dependent PAC-Bayes priors via differential privacy, *NIPS*

A flexible and powerful framework (2/2)

- ▣ Rivasplata, Parrado-Hernandez, Shawe-Taylor, Sun and Szepesvari (2018). PAC-Bayes bounds for stable algorithms with instance-dependent priors, *arXiv preprint*
- ▣ G. and Robbiano (2018). PAC-Bayesian High Dimensional Bipartite Ranking, *Journal of Statistical Planning and Inference*
- ▣ Li, G. and Loustau (2018). A Quasi-Bayesian perspective to Online Clustering, *Electronic Journal of Statistics*
- ▣ G. and Li (2018). Sequential Learning of Principal Curves: Summarizing Data Streams on the Fly, *arXiv preprint*

A flexible and powerful framework (2/2)

- ▣ Rivasplata, Parrado-Hernandez, Shawe-Taylor, Sun and Szepesvari (2018). PAC-Bayes bounds for stable algorithms with instance-dependent priors, *arXiv preprint*
- ▣ G. and Robbiano (2018). PAC-Bayesian High Dimensional Bipartite Ranking, *Journal of Statistical Planning and Inference*
- ▣ Li, G. and Loustau (2018). A Quasi-Bayesian perspective to Online Clustering, *Electronic Journal of Statistics*
- ▣ G. and Li (2018). Sequential Learning of Principal Curves: Summarizing Data Streams on the Fly, *arXiv preprint*

Towards (almost) no assumptions to derive powerful results

- ▣ Bégin, Germain, Laviolette and Roy (2016). PAC-Bayesian bounds based on the Rényi divergence, *AISTATS*
- ▣ Alquier and G. (2018). Simpler PAC-Bayesian bounds for hostile data, *Machine Learning*

Existing implementation: PAC-Bayes in the real world

Existing implementation: PAC-Bayes in the real world

▶ (Transdimensional) MCMC

📖 G. and Alquier (2013). PAC-Bayesian Estimation and Prediction in Sparse Additive Models, *Electronic Journal of Statistics*

📖 Alquier and Biau (2013). Sparse Single-Index Model, *Journal of Machine Learning Research*

📖 Li, G. and Loustau (2018). A Quasi-Bayesian perspective to Online Clustering, *Electronic Journal of Statistics*

📖 G. and Robbiano (2018). PAC-Bayesian High Dimensional Bipartite Ranking, *Journal of Statistical Planning and Inference*

Existing implementation: PAC-Bayes in the real world

▶ (Transdimensional) MCMC

▣ G. and Alquier (2013). PAC-Bayesian Estimation and Prediction in Sparse Additive Models, *Electronic Journal of Statistics*

▣ Alquier and Biau (2013). Sparse Single-Index Model, *Journal of Machine Learning Research*

▣ Li, G. and Loustau (2018). A Quasi-Bayesian perspective to Online Clustering, *Electronic Journal of Statistics*

▣ G. and Robbiano (2018). PAC-Bayesian High Dimensional Bipartite Ranking, *Journal of Statistical Planning and Inference*

▶ Stochastic optimization

▣ Alquier and G. (2017). An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization, *Mathematical Methods of Statistics*

▣ G. and Li (2018). Sequential Learning of Principal Curves: Summarizing Data Streams on the Fly, *arXiv preprint*

Existing implementation: PAC-Bayes in the real world

▶ (Transdimensional) MCMC

▣ G. and Alquier (2013). PAC-Bayesian Estimation and Prediction in Sparse Additive Models, *Electronic Journal of Statistics*

▣ Alquier and Biau (2013). Sparse Single-Index Model, *Journal of Machine Learning Research*

▣ Li, G. and Loustau (2018). A Quasi-Bayesian perspective to Online Clustering, *Electronic Journal of Statistics*

▣ G. and Robbiano (2018). PAC-Bayesian High Dimensional Bipartite Ranking, *Journal of Statistical Planning and Inference*

▶ Stochastic optimization

▣ Alquier and G. (2017). An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization, *Mathematical Methods of Statistics*

▣ G. and Li (2018). Sequential Learning of Principal Curves: Summarizing Data Streams on the Fly, *arXiv preprint*

▶ Variational Bayes

▣ Alquier, Ridgway and Chopin (2016). On the properties of variational approximations of Gibbs posteriors, *Journal of Machine Learning Research*

(intermediary) take-home message

(intermediary) take-home message

PAC-Bayesian learning is a flexible and powerful machinery.

(intermediary) take-home message

PAC-Bayesian learning is a flexible and powerful machinery.

- + little to no assumptions (teaser for second part)
- + flexibility: works as long as you can define a loss
- + generalization properties: state-of-the-art PAC risk bounds
- + model-free learning

(intermediary) take-home message

PAC-Bayesian learning is a flexible and powerful machinery.

- + little to no assumptions (teaser for second part)
- + flexibility: works as long as you can define a loss
- + generalization properties: state-of-the-art PAC risk bounds
- + model-free learning

- still perceived as a black box and suffers from lack of interpretability
- implementation plagued with the same issues as "classical" Bayesian learning (speed / high dim / ...)

A unified PAC-Bayesian framework

 Alquier and G. (2018)

Simpler PAC-Bayesian Bounds for hostile data

Machine Learning

Motivation: towards an agnostic learning theory

PAC-Bayesian bounds are a key justification in stat/ML for using Bayesian-flavored learning algorithms in several settings.

high dimensional bipartite ranking, non-negative matrix factorization, sequential learning of principal curves, online clustering, single-index models, high dimensional additive regression, domain adaptation, neural networks, ...

Conversely, they are also used to elicit new learning algorithms.

Motivation: towards an agnostic learning theory

PAC-Bayesian bounds are a key justification in stat/ML for using Bayesian-flavored learning algorithms in several settings.

high dimensional bipartite ranking, non-negative matrix factorization, sequential learning of principal curves, online clustering, single-index models, high dimensional additive regression, domain adaptation, neural networks, ...

Conversely, they are also used to elicit new learning algorithms.

Most of these bounds rely on heavy and unrealistic assumptions: e.g., boundedness of the loss function, independence. Hardly met when working on real data!

We relaxed these constraints and provide unprecedented PAC-Bayesian learning bounds for dependent and/or heavy-tailed data, a.k.a *hostile data*.

▶ skip context

Context: PAC bounds for heavy-tailed random variables

Calm before the storm (< 2015)

PAC-Bayesian bounds for unbounded losses, under strong exponential moments assumptions.

 Catoni (2004). Statistical Learning Theory and Stochastic Optimization, Springer

Context: PAC bounds for heavy-tailed random variables

The next big thing (≥ 2015)

- ▶ PAC bounds for the (penalized) ERM without an exponential moments assumption with the small-ball property

📖 Mendelson (2015). Learning without concentration, *Journal of ACM* 📖 Lecué and Mendelson (2016). Regularization and the small-ball method, *The Annals of Statistics* 📖 Grünwald and Mehta (2016). Fast Rates for Unbounded Losses, *arXiv preprint*

- ▶ Robust loss functions

📖 Catoni (2016). PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design, *arXiv preprint*

- ▶ Median-of-means tournaments for estimating the mean without an exponential moments assumption.

📖 Devroye, Lerasle, Lugosi and Oliveira (2016). Sub-Gaussian mean estimators, *The Annals of Statistics*

📖 Lugosi and Mendelson (2018). Risk minimization by median-of-means tournaments, *Journal of the*

European Mathematical Society 📖 Lugosi and Mendelson (2017). Regularization, sparse recovery, and

median-of-means tournaments, *arXiv preprint* 📖 Lecué and Lerasle (2017). Learning from MoM's

principles: Le Cam's approach, *arXiv preprint*

Context: PAC bounds for dependent observations

PAC(-Bayesian) bounds have been provided by a series of papers. However all these works relied on concentration inequalities or limit theorems for time series for which boundedness or exponential moments assumption are crucial.

📖 Mohri and Rostamizadeh (2010). Stability bounds for stationary ϕ -mixing and β -mixing processes, *Journal of Machine Learning Research*

📖 Ralaivola, Szafranski and Stempfel (2010). Chromatic PAC-Bayes bounds for non-iid data: Applications to ranking and stationary β -mixing processes, *Journal of Machine Learning Research*

📖 Seldin, Lavolette, Cesa-Bianchi, Shawe-Taylor and Auer (2012). PAC-Bayesian inequalities for martingales, *IEEE Transactions on Information Theory*

📖 Alquier and Wintenberger (2012). Model selection for weakly dependent time series forecasting, *Bernoulli*

📖 Agarwal and Duchi (2013). The generalization ability of online algorithms for dependent data, *IEEE Transactions on Information Theory*

📖 Kuznetsov and Mohri (2014). Generalization bounds for time series prediction with non-stationary processes, *ALT*

Disclaimer

The strategy I'm about to describe yields, at best, the same rates as those existing in known settings.

Disclaimer

The strategy I'm about to describe yields, at best, the same rates as those existing in known settings.

However we designed a unified framework to derive PAC-Bayesian bounds for settings where even no PAC learning bounds were available (such as heavy-tailed time series).

Notation

Loss function ℓ , observations $(X_1, Y_1), \dots, (X_n, Y_n)$, family of predictors $(f_\theta, \theta \in \Theta)$.

Observations are not required to be independent nor identically distributed. Let $\ell_i(\theta) = \ell[f_\theta(X_i), Y_i]$, and define the (empirical) risk as

$$r_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ell_i(\theta),$$
$$R(\theta) = \mathbb{E}[r_n(\theta)].$$

Key quantities

Key quantities

Definition

For any function g , let

$$\mathcal{M}_{\phi_p, n} = \int \mathbb{E} (|r_n(\theta) - R(\theta)|^p) \pi(d\theta).$$

Key quantities

Definition

For any function g , let

$$\mathcal{M}_{\phi_p, n} = \int \mathbb{E} (|r_n(\theta) - R(\theta)|^p) \pi(d\theta).$$

Definition

Let f be a convex function with $f(1) = 0$. Csiszár's f -divergence between ρ and π is defined by

$$D_f(\rho, \pi) = \int f\left(\frac{d\rho}{d\pi}\right) d\pi$$

when $\rho \ll \pi$ and $D_f(\rho, \pi) = +\infty$ otherwise.

Note that $\mathcal{K}(\rho, \pi) = D_{x \log(x)}(\rho, \pi)$ and $\chi^2(\rho, \pi) = D_{x^2-1}(\rho, \pi)$.

Main theorem

Fix $p > 1$, $q = \frac{p}{p-1}$ and $\delta \in (0, 1)$. With probability at least $1 - \delta$ we have for any distribution ρ

$$\left| \int R d\rho - \int r_n d\rho \right| \leq \left(\frac{\mathcal{M}_{\phi_q, n}}{\delta} \right)^{\frac{1}{q}} (D_{\phi_{p-1}}(\rho, \pi) + 1)^{\frac{1}{p}}.$$

Proof

Let $\Delta_n(\theta) := |r_n(\theta) - R(\theta)|$.

$$\begin{aligned} \left| \int R d\rho - \int r_n d\rho \right| &\leq \int \Delta_n d\rho = \int \Delta_n \frac{d\rho}{d\pi} d\pi \\ &\leq \left(\int \Delta_n^q d\pi \right)^{\frac{1}{q}} \left(\int \left(\frac{d\rho}{d\pi} \right)^p d\pi \right)^{\frac{1}{p}} \quad (\text{Hölder ineq.}) \\ &\leq \left(\frac{\mathbb{E} \int \Delta_n^q d\pi}{\delta} \right)^{\frac{1}{q}} \left(\int \left(\frac{d\rho}{d\pi} \right)^p d\pi \right)^{\frac{1}{p}} \quad (\text{Markov, w.p. } 1 - \delta) \\ &= \left(\frac{\mathcal{M}_{\phi_q, n}}{\delta} \right)^{\frac{1}{q}} (D_{\phi_p-1}(\rho, \pi) + 1)^{\frac{1}{p}}. \end{aligned}$$

Inspired by

 Bégin, Germain, Lavolette and Roy (2016). PAC-Bayesian bounds based on the Rényi divergence, *AISTATS*

We can compare $\int r_n d\rho$ (observable) to $\int R d\rho$ (unknown, the objective) in terms of

- ▶ the moment $\mathcal{M}_{\phi_q, n}$ (which depends on the distribution of the data)
- ▶ and the divergence $D_{\phi_p-1}(\rho, \pi)$ (which is a measure of the complexity of the set Θ).

We can compare $\int r_n d\rho$ (observable) to $\int R d\rho$ (unknown, the objective) in terms of

- ▶ the moment $\mathcal{M}_{\phi_q, n}$ (which depends on the distribution of the data)
- ▶ and the divergence $D_{\phi_p-1}(\rho, \pi)$ (which is a measure of the complexity of the set Θ).

Corollary: with probability at least $1 - \delta$, for any ρ ,

$$\int R d\rho \leq \int r_n d\rho + \left(\frac{\mathcal{M}_{\phi_q, n}}{\delta} \right)^{\frac{1}{q}} (D_{\phi_p-1}(\rho, \pi) + 1)^{\frac{1}{p}}.$$

Strong incitement to define our aggregation distribution $\hat{\rho}_n$ as the minimizer of the right-hand side!

Intersection

1. Computing the divergence term

- ▶ Discrete case

 - ▶ goto

- ▶ Continuous case

 - ▶ goto

2. Bounding the moments

- ▶ The iid case

 - ▶ goto

- ▶ The dependent case

 - ▶ goto

3. PAC-Bayes bounds to elicit new learning algorithms

 - ▶ goto

4. Conclusion

 - ▶ goto

Computing the divergence term (discrete case)

Assume $\text{Card}(\Theta) = K < \infty$ and that π is uniform on Θ . Fix $p > 1$, $q = \frac{p}{p-1}$ and $\delta \in (0, 1)$. With probability at least $1 - \delta$

$$R(\hat{\theta}_{\text{ERM}}) \leq \inf_{\theta \in \Theta} \{r_n(\theta)\} + K^{1-\frac{1}{p}} \left(\frac{\mathcal{M}_{\phi_q, n}}{\delta} \right)^{\frac{1}{q}}.$$

▶ back to intersection

Computing the divergence term (continuous case)

Assume that there exists $d > 0$ such that for any $\gamma > 0$,

$$\pi \left\{ \theta \in \Theta : \{r_n(\theta)\} \leq \inf_{\theta' \in \Theta} r_n(\theta') + \gamma \right\} \geq \gamma^d.$$

Fix $p > 1$, $q = \frac{p}{p-1}$, $\delta \in (0, 1)$ and

$$\pi_\gamma(d\theta) \propto \pi(d\theta) \mathbf{1} \left[r(\theta) - r_n(\hat{\theta}_{\text{ERM}}) \leq \gamma \right].$$

With probability at least $1 - \delta$

$$\int R d\pi_\gamma \leq \inf_{\theta \in \Theta} \{r_n(\theta)\} + \left(\frac{\mathcal{M}_{\phi_q, n}}{\delta} \right)^{\frac{1}{1+\frac{d}{q}}} \left\{ \left(\frac{d}{q} \right)^{\frac{1}{1+\frac{d}{q}}} + \left(\frac{d}{q} \right)^{\frac{-\frac{d}{q}}{1+\frac{d}{q}}} \right\}.$$

Bounding the moment $\mathcal{M}_{\phi_q, n}$: the i.i.d case

Assume that

$$s^2 = \int \text{Var}[\ell_1(\theta)]\pi(d\theta) < +\infty$$

then

$$\mathcal{M}_{\phi_q, n} \leq \left(\frac{s^2}{n}\right)^{\frac{q}{2}}.$$

So

$$\int R d\rho \leq \int r_n d\rho + \frac{(D_{\phi_p-1}(\rho, \pi) + 1)^{\frac{1}{p}}}{\delta^{\frac{1}{q}}} \sqrt{\frac{s^2}{n}}.$$

This rate can not be improved without further assumptions.

▶ back to intersection

Bounding the moment $\mathcal{M}_{\phi_q, n}$: the i.i.d case

Assume $\text{Card}(\Theta) = K < +\infty$ and for any θ , $\ell_i(\theta)$ is sub-Gaussian with parameter σ^2 .

For any $\delta \in (0, 1)$, with probability at least $1 - \delta$

$$R(\hat{\theta}_{\text{ERM}}) \leq \inf_{\theta \in \Theta} \{r_n(\theta)\} + \sqrt{\frac{2e\sigma^2 \log\left(\frac{2K}{\delta}\right)}{n}}.$$

This rate can not be improved without further assumptions on the loss ℓ .

 Audibert (2009). Fast learning rates in statistical inference through aggregation, *The Annals of Statistics*

[▶ back to intersection](#)

Bounding the moment $\mathcal{M}_{\phi_q, n}$: the dependent case

Definition

The α -mixing coefficients between two σ -algebras \mathcal{F} and \mathcal{G} are defined by

$$\alpha(\mathcal{F}, \mathcal{G}) = \sup_{A \in \mathcal{F}, B \in \mathcal{G}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|.$$

Define

$$\alpha_j = \alpha[\sigma(X_0, Y_0), \sigma(X_j, Y_j)].$$

When the future of the series is strongly dependent of the past, α_j will remain constant or slowly decay. When the near future is almost independent of the past, then the α_j quickly decay to 0.

▶ back to intersection

Bounding the moment $\mathcal{M}_{\phi_q, n}$: the dependent case

Bounded case: assume $0 \leq \ell \leq 1$ and $(X_i, Y_i)_{i \in \mathbb{Z}}$ is a stationary process which satisfies $\sum_{j \in \mathbb{Z}} \alpha_j < \infty$. Then

$$\mathcal{M}_{\phi_2, n} \leq \frac{1}{n} \sum_{j \in \mathbb{Z}} \alpha_j.$$

Bounding the moment $\mathcal{M}_{\phi_q, n}$: the dependent case

Bounded case: assume $0 \leq \ell \leq 1$ and $(X_i, Y_i)_{i \in \mathbb{Z}}$ is a stationary process which satisfies $\sum_{j \in \mathbb{Z}} \alpha_j < \infty$. Then

$$\mathcal{M}_{\phi_2, n} \leq \frac{1}{n} \sum_{j \in \mathbb{Z}} \alpha_j.$$

Unbounded case: assume that $(X_i, Y_i)_{i \in \mathbb{Z}}$ is a stationary process. Let $1/r + 2/s = 1$ and assume

$$\sum_{j \in \mathbb{Z}} \alpha_j^{1/r} < \infty, \quad \int \{\mathbb{E}[\ell_i^s(\theta)]\}^{\frac{2}{s}} \pi(d\theta) < \infty.$$

Then

$$\mathcal{M}_{\phi_2, n} \leq \frac{1}{n} \left(\int \{\mathbb{E}[\ell_i^s(\theta)]\}^{\frac{2}{s}} \pi(d\theta) \right) \left(\sum_{j \in \mathbb{Z}} \alpha_j^{\frac{1}{r}} \right).$$

Example

Consider auto-regression with quadratic loss and linear predictors:
 $X_i = (1, Y_{i-1}) \in \mathbb{R}^2$, $\Theta = \mathbb{R}^2$ and $f_\theta(\cdot) = \langle \theta, \cdot \rangle$. Let

$$\nu = 32\mathbb{E} (Y_i^6)^{\frac{2}{3}} \sum_{j \in \mathbb{Z}} \alpha_j^{\frac{1}{3}} \left(1 + 4 \int \|\theta\|^6 \pi(d\theta) \right).$$

With probability at least $1 - \delta$ we have for any ρ

$$\int R d\rho \leq \int r_n d\rho + \sqrt{\frac{\nu[1 + \chi^2(\rho, \pi)]}{n\delta}}.$$

Up to our knowledge, first PAC(-Bayesian) bound in the case of a time series without any boundedness nor exponential moment assumption.

PAC-Bayesian bounds to elicit new learning algorithms

Reminder:

For $p > 1$ and $q = p/(p - 1)$, with probability at least $1 - \delta$ we have for any ρ

$$\int R d\rho \leq \int r_n d\rho + \left(\frac{\mathcal{M}_{\phi_q, n}}{\delta} \right)^{\frac{1}{q}} (D_{\phi_{p-1}}(\rho, \pi) + 1)^{\frac{1}{p}}.$$

▶ back to intersection

Definition

We define $\bar{r}_n = \bar{r}_n(\delta, \rho)$ as

$$\bar{r}_n = \min \left\{ u \in \mathbb{R}, \int [u - r_n(\theta)]_+^q \pi(d\theta) = \frac{\mathcal{M}_{\phi_q, n}}{\delta} \right\}.$$

Such a minimum always exists as the integral is a continuous function of u , is equal to 0 when $u = 0$ and $\rightarrow \infty$ when $u \rightarrow \infty$.

We then define

$$\frac{d\hat{\rho}_n}{d\pi}(\theta) = \frac{[\bar{r}_n - r_n(\theta)]_+^{\frac{1}{\rho-1}}}{\int [\bar{r}_n - r_n]_+^{\frac{1}{\rho-1}} d\pi}.$$

With probability at least $1 - \delta$,

$$\int R d\hat{\rho}_n \leq \bar{r}_n \leq \inf_{\rho} \left\{ \int R d\rho + 2 \left(\frac{\mathcal{M}_{\phi_q, n}}{\delta} \right)^{\frac{1}{q}} (D_{\phi_p-1}(\rho, \pi) + 1)^{\frac{1}{p}} \right\}.$$

With probability at least $1 - \delta$,

$$\int R d\hat{\rho}_n \leq \bar{r}_n \leq \inf_{\rho} \left\{ \int R d\rho + 2 \left(\frac{\mathcal{M}_{\phi_q, n}}{\delta} \right)^{\frac{1}{q}} \left(D_{\phi_{p-1}}(\rho, \pi) + 1 \right)^{\frac{1}{p}} \right\}.$$

Assume that there exists $d > 0$ such that for any $\gamma > 0$,

$$\pi \left\{ \theta \in \Theta : \{r_n(\theta)\} \leq \inf_{\theta' \in \Theta} r_n(\theta') + \gamma \right\} \geq \gamma^d.$$

With probability at least $1 - \delta$,

$$\int R d\hat{\rho}_n \leq \bar{r}_n \leq \inf_{\theta \in \Theta} \{r_n(\theta)\} + 2 \left(\frac{\mathcal{M}_{\phi_q, n}}{\delta} \right)^{\frac{1}{q+d}}.$$

Highlights

Highlights

NIPS 2017 Workshop

(Almost) 50 Shades of Bayesian Learning: PAC-Bayesian trends and insights

Long Beach Convention Center, California

December 9, 2017



- ▶ ⁶PAC
- ▶ 2 ANR-funded projects for the period 2019–2023
 - ▶ APRIORI: representation learning and deep neural networks, with PAC-Bayes
 - ▶ BEAGLE (PI): agnostic learning, with PAC-Bayes
- ▶ H2020 European Commission project PERF-AI: machine learning algorithms (including PAC-Bayes) applied to aviation



We are hiring!

Interns, Engineers, PhD students, Postdocs

Spread the word!