PAC-Bayes learning beyond bounded losses

Maxime Haddouche

Inria London

https://maximehaddouche.github.io/

Presented work

Article

PAC-Bayes Unleashed: Generalisation Bounds with Unbounded Losses

Maxime Haddouche [1], Benjamin Guedj [2,3,*], Omar Rivasplata [2] and John Shawe-Taylor [2]

`https://arxiv.org/pdf/2206.00024.pdf`

Summary

1. What is PAC-Bayes learning?
   - Introduction
   - Classical results
2. Modern extensions of PAC-Bayes theory
   - Towards tighter PAC-Bayesian bounds?
   - PAC-Bayes beyond the usual assumptions
   - Other expansions
3. Practical PAC-Bayes: towards concrete algorithms
4. PAC-Bayes learning with unbounded losses

What is PAC-Bayes learning?

- A branch of learning theory
- Emerged in the late 90s with the works of Shawe-Taylor and Williamson, 1997 and McAllester, 1998, 1999.
- Technical tools: measure theory, concentration inequalities, information theory. Also Catoni, 2007 used tools from statistical physics

What is PAC-Bayes learning?

- A branch of learning theory
- Emerged in the late 90s with the works of Shawe-Taylor and Williamson, 1997 and McAllester, 1998, 1999.
- Technical tools: measure theory, concentration inequalities, information theory. Also Catoni, 2007 used tools from statistical physics

For more precision see the recent surveys of:

1. Alquier 2021: `https://arxiv.org/abs/2110.11216`
2. Guedj 2019: `https://arxiv.org/abs/1901.05353`

Terminology

The two terms ’PAC’ and ’Bayes’ stand for the following.

- PAC is the acronym of ’Probably Approximately Correct’ and says that we ’probably’ have guarantees to have an ’approximately correct surrogate’ of the generalisation error i.e. the error we pay when exploit our trained algorithms on new data (out of the train set).
- ’Bayes’ says that we take inspiration from the Bayesian philopsophy. Indeed, PAC-Bayesian theory aims to construct distributions over the predictor space instead of a single point. It also exploits the idea of building a posterior distribution from a prior one (without using Bayes formula).

An usual framework

A *learning problem* is specified by tuple $(\mathcal{H}, \mathcal{Z}, \ell)$ where:

- $\mathcal{H}$ is the space of considered predictors
- $\mathcal{Z}$ is the data space. z can be an unlabeled data x or a couple (x, y) of a point with its label. We assume that μ is a distribution over $\mathcal{Z}$ which rules the distribution of our data.
- $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$ is a loss function i.e. the learning objective we want to minimise.

An usual framework (2)

- $S = (z_1, ...z_m)$ an iid dataset following μ.
- The generalisation risk for $h \in \mathbb{H}$: $R(h) = \mathbb{E}_{z\sim\mu}[\ell(h, z)]$.
- The empirical risk $R_m(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i)$.

What does PAC-Bayes do?

PAC-Bayes theory aims to design a meaningful distribution Q over $\mathcal{H}$.
A classical PAC-Bayes bound controls the *expected generalisation error*:

$$\mathbb{E}_{h\sim Q}[R(h)] := \mathbb{E}_{h\sim Q}\mathbb{E}_{z\sim \mu}[\ell(h,z)]$$

with regards to the *expected empirical error*:

$$\mathbb{E}_{h\sim Q}[R_m(h)] := \mathbb{E}_{h\sim Q}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(h,z_i)\right]$$

Classical examples: McAllester's bound (Guedj, 2019 Thm.1)

Assumptions: $\ell \in [0, 1]$, iid data, data-free prior

Theorem

For any prior distribution P, we have with probability $1 - \delta$ over the m-sample S, for any posterior distribution Q such that $Q \ll P$:

$$\mathbb{E}_{h\sim Q}\left[R(h)\right] \leqslant \mathbb{E}_{h\sim Q}\left[R_m(h)\right] + \sqrt{\frac{KL(Q,P) + \log(2\sqrt{m}/\delta)}{2m}},$$

where KL is the Kullback-Leibler divergence.

Classical examples: Catoni's bound (Catoni, 2007 Thm.1.2.6)

Assumptions: $\ell \in [0, 1]$, iid data, data-free prior

Theorem

For any prior distribution P, any $\lambda > 0$, we have with probability $1 - \delta$ over the m-sample S, for any posterior distribution Q such that $Q \ll P$:

$$\mathbb{E}_{h\sim Q}\left[R(h)\right] \leqslant \frac{1 - \exp\left\{-\frac{\lambda \mathbb{E}_{h\sim Q}[R_m(h)]}{m} - \frac{KL(Q||P) - \log(\delta)}{m}\right\}}{1 - \exp\left(-\frac{\lambda}{m}\right)},$$

where KL is the Kullback-Leibler divergence.

Towards tighter PAC-Bayes bounds?

A line of work has been dedicated to the discovering of novel PAC-Bayes bound with the hope of tightening classical results or provide variety in the nature of the terms involved, some of those are cited below.

Towards tighter PAC-Bayes bounds?

A line of work has been dedicated to the discovering of novel PAC-Bayes bound with the hope of tightening classical results or provide variety in the nature of the terms involved, some of those are cited below.

- (Tolstikhin and Seldin, 2013) proposed bounds involving correlation between the empirical risk and the KL term, they also introduced PAC-Bayes Bernstein inequalities.
- Thiemann et al. 2017 proposed a quasiconvex PAC-Bayesian bound which allowed to exploit directly optimisation routes from the convex world.

Towards tighter PAC-Bayes bounds? (2)

- Mhameddi et al. 2019 tightened Tolstikhin and Seldin's work by providing a bound involving new and refined variance terms.
- Bégin et al. 2016 proposed new PAC-Bayesian bounds based on Rényi divergences instead of the classical KL.

PAC-Bayes beyond the usual assumptions

The three main assumptions made for McAllester and Catoni's bounds are bounded loss, iid data and data-free prior. Recently some works tried to overcome those restrictive assumptions, tailored for supervised classification, to reach other learning problems.

PAC-Bayes beyond the usual assumptions

The three main assumptions made for McAllester and Catoni's bounds are bounded loss, iid data and data-free prior. Recently some works tried to overcome those restrictive assumptions, tailored for supervised classification, to reach other learning problems.

- Oneto et al. 2016 exploit Catoni's methodology on localised priors to propose results for bounded losses and data-dependent priors.

PAC-Bayes beyond the usual assumptions (2)

- Alquier and Guedj, 2017 propose loose PAC-Bayes bound for hostile data, i.e. no assumption on the data distribution was made and losses are possibly unbounded.
- The recent work of Rivasplata et al. 2020 proposes a generic PAC-Bayes theorem to unbounded losses, dependent data and data-dependent priors
- The work of Haddouche et al. 2021 tackles the problem of bounded losses in PAC-Bayes learning by proposing to re-explorate classical PAC-Bayes routes with a specific type of unbounded loss.

Towards Online Learning

The PAC-Bayesian literature has mainly been conceived in a batch learning setting. However a short line of work has emerged, building links between PAC-Bayes and online learning:

Towards Online Learning

The PAC-Bayesian literature has mainly been conceived in a batch learning setting. However a short line of work has emerged, building links between PAC-Bayes and online learning:

- Gerchinowitz, 2011 proposes PAC-Bayesian adaptive regret bounds for online linear regression.
- Seldin et al. 2012, proposed PAC-Bayesian inequalities for martingales which forms a more dynamic structure than iid data.
- Haddouche and Guedj, 2022 propose a theoretical bridge between PAC-Bayes and online learning through a PAC-Bayesian theoretical result fully adapted to the online learning philosophy.

Towards Bandits and RL

Some authors exploited the flexibility of PAC-Bayes theory to deal with the question of bandits and reinforcement learning (RL)

- Fard and Pineau, 2010 introduces the idea of PAC- Bayesian model-selection in reinforcement learning (RL).
- Fard and Pineau, 2012, develop their methodology to create PAC-Bayesian policy evaluation.
- Seldin et al. 2012 proposed an application of a PAC-Bayes bound to multi-armed bandits
- Flynn et al. 2022 studies multi-armed bandits in the framework of lifelong learning with a PAC-Bayesian analysis.

How to obtain a PAC-Bayesian algorithm?

The main idea: generate a posterior Q from the prior P and the data by optimising PAC-Bayesian upper bound.

How to obtain a PAC-Bayesian algorithm?

The main idea: generate a posterior Q from the prior P and the data by optimising PAC-Bayesian upper bound.
For McAllester's bound with a given sample $S = (z_1, \ldots, z_m)$ and prior P we have the output $\hat{Q}$:

$$\hat{Q} := \operatorname{argmin}_Q \mathbb{E}_{h \sim Q} [R_m(h)] + \sqrt{\frac{KL(Q, P) + \log(2\sqrt{m}/\delta)}{2m}}$$

How to obtain a PAC-Bayesian algorithm?

The main idea: generate a posterior Q from the prior P and the data by optimising PAC-Bayesian upper bound.
For McAllester's bound with a given sample $S = (z_1, \ldots, z_m)$ and prior P we have the output $\hat{Q}$:

$$\hat{Q} := \operatorname{argmin}_Q \mathbb{E}_{h \sim Q} \left[R_m(h) \right] + \sqrt{\frac{KL(Q, P) + \log(2\sqrt{m}/\delta)}{2m}}$$

This problem may be too complicated in itself given the wide range of available distributions!

PAC-Bayesian algorithms (2)

Usually practitionners restrict themselves to a specific (parametric) class of measures such as Gaussian distributions to simplify the optimisation and deal with the KL divergence term.
PAC-Bayes algorithm remained mainly theoretical until the groundbreaking work of Dziugaite et al. 2017 which computed a PAC-Bayesian training of neural network with non-vacuous theoretical guarantees. This attracted a lot of attention.

PAC-Bayes and Neural Networks

Since Dziugaite et al. 2017, many works tackled the question of a theoretical understanding of neural networks:

- The work of Letarte et al. 2019 focused on PAC-Bayesian binary activated deep NNs,
- Rivasplata et al. 2019 proposed a study of a PAC-Bayesian version of backpropagation,
- Perez-Ortiz et al. 2021 thought about the optimal amount of data required to tune the prior distribution in order to obtain quickly tight risk certificates,
- Another work of Perez-Ortiz et al. 2021 enhanced the theoretical certifications for porbabilistic neural networks.

Other routes

- Mhameddi et al. 2020 proposed to exploit PAC-Bayes flexibility in the context of the Conditional Value at Risk.
- Cantelobre et al. 2020 proposed a PAC-Bayesian view on structured prediction with implicit loss embeddings.
- Dziugaite and Roy, 2018 mixed up PAC-Bayes with differential privacy
- Haddouche et al. 2020 proposed generalisation bounds for the kernel PCA algorithm based on PAC-Bayesian results.

Other routes

- Mhameddi et al. 2020 proposed to exploit PAC-Bayes flexibility in the context of the Conditional Value at Risk.
- Cantelobre et al. 2020 proposed a PAC-Bayesian view on structured prediction with implicit loss embeddings.
- Dziugaite and Roy, 2018 mixed up PAC-Bayes with differential privacy
- Haddouche et al. 2020 proposed generalisation bounds for the kernel PCA algorithm based on PAC-Bayesian results.

Thus, PAC-Bayes theory is both theoretical and practical, is adaptible to many learning situations and seems to be one of the few to propose non-vacuous theoretical guarantees for neural networks.

PAC-Bayes learning with unbounded losses

We now present a modern PAC-Bayes extension which overcomes (at a certain point) the classical assumption of a bounded loss.

PAC-Bayes learning with unbounded losses

We now present a modern PAC-Bayes extension which overcomes (at a certain point) the classical assumption of a bounded loss.
Our alternative assumption: the HYPE condition:

Definition (Hypothesis-dependent range (HYPE))

A loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}^+$ is said to satisfy the **hypothesis-dependent range** ($\texttt{HYPE}$) condition if there exists a function $K : \mathcal{H} \to \mathbb{R}^+ \backslash \{0\}$ such that $\sup_{z \in \mathcal{Z}} \ell(h, z) \leqslant K(h)$ for every predictor h. We then say that ℓ is $\texttt{HYPE}(K)$ compliant.

A preliminary result (adapted from Germain et al., 2009)

Theorem

For any $P \in \mathcal{M}_1^+(\mathcal{H})$ *with no dependency on data, for any convex function* $F : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$*, for any* $\alpha \in \mathbb{R}$ *and for any* $\delta \in [0 : 1]$*, we have with probability at least* $1 - \delta$ *over size-m samples* $\mathcal{S}$*, for any* $Q \in \mathcal{M}_1^+(\mathcal{H})$ *such that* $Q \ll P$ *and* $P \ll Q$*:*

$$F\left(\mathbb{E}_{h\sim Q}\left[R_m(h)\right], \mathbb{E}_{h\sim Q}\left[R(h)\right]\right) \leqslant \frac{1}{m^\alpha}\left(KL(Q||P) + \log\left(\frac{1}{\delta}\mathbb{E}_{h\sim P}\mathbb{E}_{\mathcal{S}}\ e^{m^\alpha F(R_m(h),R(h))}\right)\right)$$

Proof

Since $\mathbb{E}_{h\sim P}\left[e^{F(R_S(h),R(h))}\right]$ is a nonnegative random variable, we know that, by Markov's inequality, for any $h \in \mathcal{H}$:

$$\mathbb{P}\left(\mathbb{E}_{h\sim P}\left[e^{F(R_S(h),R(h))}\right] > \frac{1}{\delta}\mathbb{E}_S\,\mathbb{E}_{h\sim P}\left[e^{F(R_S(h),R(h))}\right]\right) \leqslant \delta.$$

Proof

Since $\mathbb{E}_{h\sim P}\left[e^{F(R_S(h),R(h))}\right]$ is a nonnegative random variable, we know that, by Markov's inequality, for any $h \in \mathcal{H}$:

$$\mathbb{P}\left(\mathbb{E}_{h\sim P}\left[e^{F(R_S(h),R(h))}\right] > \frac{1}{\delta}\mathbb{E}_S\,\mathbb{E}_{h\sim P}\left[e^{F(R_S(h),R(h))}\right]\right) \leqslant \delta.$$

So with probability of at least $1-\delta$, we have:

$$\mathbb{E}_{h\sim P}\left[e^{F(R_S(h),R(h))}\right] \leqslant \frac{1}{\delta}\mathbb{E}_S\,\mathbb{E}_{h\sim P}\left[e^{F(R_S(h),R(h))}\right] = \frac{\chi}{\delta}.$$

Proof

Applying the log function gives us:

$$\log\left(\mathbb{E}_{h\sim P}\left[e^{F(R_S(h),R(h))}\right]\right) \leqslant \log\left(\frac{\chi}{\delta}\right).$$

Proof

Applying the log function gives us:

$$\log\left(\mathbb{E}_{h\sim P}\left[e^{F(R_S(h),R(h))}\right]\right) \leqslant \log\left(\frac{\chi}{\delta}\right).$$

We now rename $A := \log\left(\mathbb{E}_{h\sim P}\left[e^{F(R_S(h),R(h))}\right]\right)$.
Furthermore, if we denote by $\frac{dQ}{dP}$ the Radon-Nikodym derivative of Q with respect to P when $Q \ll P$, we then have, for all Q such that $Q \sim P$:

Proof

$$A = \log\left(\mathbb{E}_{h\sim Q}\left[\frac{dP}{dQ}e^{F(R_S(h),R(h))}\right]\right)$$

$$= \log\left(\mathbb{E}_{h\sim Q}\left[\left(\frac{dQ}{dP}\right)^{-1}e^{F(R_S(h),R(h))}\right]\right) \qquad \left(\tfrac{dP}{dQ} = \left(\tfrac{dQ}{dP}\right)^{-1}\right)$$

Proof

$$A = \log\left(\mathbb{E}_{h\sim Q}\left[\frac{dP}{dQ}e^{F(R_S(h),R(h))}\right]\right)$$
$$= \log\left(\mathbb{E}_{h\sim Q}\left[\left(\frac{dQ}{dP}\right)^{-1}e^{F(R_S(h),R(h))}\right]\right) \qquad \left(\tfrac{dP}{dQ} = \left(\tfrac{dQ}{dP}\right)^{-1}\right)$$

and by concavity of log and Jensen's inequality,

$$\geqslant -\mathbb{E}_{h\sim Q}\left[\log\left(\frac{dQ}{dP}\right)\right] + \mathbb{E}_{h\sim Q}\left[F(R_S(h), R(h))\right]$$
$$= -\,\mathrm{KL}(Q\|P) + \mathbb{E}_{h\sim Q}\left[F(R_S(h), R(h))\right]$$

Proof

while by convexity of F with Jensen's inequality,

$$\geqslant -\mathrm{KL}(Q\|P) + F\left(\mathbb{E}_{h\sim Q}\left[R_S(h)\right], \mathbb{E}_{h\sim Q}\left[R(h)\right]\right).$$

Proof

while by convexity of F with Jensen's inequality,

$$\geqslant -\mathrm{KL}(Q||P) + F\left(\mathbb{E}_{h\sim Q}\left[R_S(h)\right], \mathbb{E}_{h\sim Q}\left[R(h)\right]\right).$$

Hence, for Q such that $Q \sim P$,

$$F\left(\mathbb{E}_{h\sim Q}\left[R_S(h)\right], \mathbb{E}_{h\sim Q}\left[R(h)\right]\right) \leqslant \mathrm{KL}(Q||P) + A.$$

Proof

So with probability $1 - \delta$, for Q such that $Q \sim P$,

$$F\left(\mathbb{E}_{h\sim Q}\left[R_S(h)\right], \mathbb{E}_{h\sim Q}\left[R(h)\right]\right) \leqslant \mathrm{KL}(Q||P) + \log\left(\frac{\chi}{\delta}\right).$$

Proof

So with probability $1 - \delta$, for Q such that $Q \sim P$,

$$F\left(\mathbb{E}_{h\sim Q}\left[R_S(h)\right], \mathbb{E}_{h\sim Q}\left[R(h)\right]\right) \leqslant \mathrm{KL}(Q||P) + \log\left(\frac{\chi}{\delta}\right).$$

This concludes the proof.

A first general PAC-Bayes bound

Now if one exploits our self-boundedness condition, one has:

Theorem

Let the loss ℓ being $\mathtt{HYPE}(K)$ compliant. For any $P \in \mathcal{M}(\mathcal{H})$ with no data dependency, for any $\alpha \in \mathbb{R}$ and for any $\delta \in [0:1]$, we have with probability at least $1-\delta$ over size-m samples $\mathcal{S}$, for any Q such that $Q \ll P$ and $P \ll Q$:

$$\mathbb{E}_{h\sim Q}\left[R(h)\right] \leqslant \mathbb{E}_{h\sim Q}\left[R_m(h)\right] + \frac{KL(Q\|P) + \log\left(\frac{1}{\delta}\right)}{m^\alpha} + \frac{1}{m^\alpha}\log\left(\mathbb{E}_{h\sim P}\left[\exp\left(\frac{K(h)^2}{2m^{1-2\alpha}}\right)\right]\right).$$

A first theorem for gaussian priors and posteriors

Theorem

Let $\alpha \in \mathbb{R}$ and $N \geqslant 6$. If the loss ℓ is $\mathtt{HYPE}(K)$ compliant with $K(h) = B\|h\| + C$, with $B > 0$, $C \geqslant 0$, then we have, for any Gaussian prior $P = \mathcal{N}(0, \sigma^2 \boldsymbol{I}_N)$ with $\sigma^2 = t\frac{m^{1-2\alpha}}{B^2}$, $0 < t < 1$. We have with probability $1 - \delta$ over size-m samples $\mathcal{S}$, for any $Q \in \mathcal{M}_1^+(\mathcal{H})$ such that $Q \ll P$ and $P \ll Q$, with $f(t) = \frac{1-t}{t}$:

$$\mathbb{E}_{h\sim Q}[R(h)] \leqslant \mathbb{E}_{h\sim Q}[R_m(h)] + \frac{\mathrm{KL}(Q\|P) + \log(2/\delta)}{m^\alpha} + \frac{C^2}{2m^{1-\alpha}}\left(1 + f(t)^{-1}\right) + \frac{N}{m^\alpha}\left(\log\left(1 + \left(\frac{C}{\sqrt{2f(t)m^{1-2\alpha}}}\right)\right) + \log\left(\frac{1}{\sqrt{1-t}}\right)\right)$$

ψ-risks

Idea: modify the estimator R_m to attenuate the influence of the empirical losses $(\ell(h, z_i))_{i=1..m}$ that exceed the threshold s.

Definition ($\psi - risks$)

For every $s > 0$, $\psi : \mathbb{R}^+ \to \mathbb{R}^+$, for any $h \in \mathcal{H}$, we define the *empirical* ψ*-risk* $R_{m,\psi,s}$ and the *theoretical* ψ*-risk* $R_{\psi,s}$ as follows:

$$R_{m,\psi,s}(h) := \frac{s}{m} \sum_{i=1}^{m} \psi\left(\frac{\ell(h, z_i)}{s}\right)$$

and

$$R_{\psi,s}(h) := \mathbb{E}_{\mathcal{S}}\left[R_{m,\psi,s}(h)\right] = \mathbb{E}_{\mu}\left[s\,\psi\left(\frac{\ell(h, z)}{s}\right)\right]$$

where $z \sim \mu$.

Softening functions

We now focus on what we call *softening functions*, *i.e.* functions that will temperate high values of the loss function ℓ.

Definition (Softening function)

We say that $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ is a softening function if:

- $\forall x \in [0; 1], \psi(x) = x$,
- ψ is non-decreasing,
- $\forall x \geqslant 1, \psi(x) \leqslant x$.

We let $\mathcal{F}$ denote the set of all softening functions.

The main theorem

Theorem

Let ℓ being $\mathtt{HYPE}(K)$ compliant. Then for any $P \in \mathcal{M}_1^+(\mathcal{H})$ with no data dependency, for any $\alpha \in \mathbb{R}$, for any $\psi \in \mathcal{F}$ and for any $\delta \in [0:1]$, we have with probability at least $1-\delta$ over size-m samples $\mathcal{S}$, for any Q such that $Q \ll P$ and $P \ll Q$:

$$\begin{aligned}\mathbb{E}_{h\sim Q}\left[R(h)\right] \leqslant \mathbb{E}_{h\sim Q}\left[R_{m,\psi,s}(h)\right] + \mathbb{E}_{h\sim Q}\left[K(h)\mathbb{1}\{K(h) \geqslant s\}\right] \\ + \frac{\mathrm{KL}(Q\|P) + \log\left(\frac{1}{\delta}\right)}{m^\alpha} \\ + \frac{1}{m^\alpha}\log\left(\mathbb{E}_{h\sim P}\left[\exp\left(\frac{s^2}{2m^{1-2\alpha}}\psi\left(\frac{K(h)}{s}\right)^2\right)\right]\right).\end{aligned}$$

Remarks

The previous result is valid under the following hypothesis:

$$\forall Q \in \mathcal{M}_1^+(\mathcal{H}), \quad \mathbb{E}_{h\sim Q}[K(h)] < +\infty. \tag{1}$$

Remarks

The previous result is valid under the following hypothesis:

$$\forall Q \in \mathcal{M}_1^+(\mathcal{H}), \quad \mathbb{E}_{h\sim Q}[K(h)] < +\infty. \tag{1}$$

Notice that for every posterior Q, the function $\psi : x \mapsto x\mathbb{1}\{x \leqslant 1\} + \mathbb{1}\{x > 1\}$ is such that $\mathbb{E}_{h\sim P}\left[\exp\left(\frac{s^2}{2m^{1-2\alpha}}\psi\left(\frac{K(h)}{s}\right)^2\right)\right] < +\infty$. Thus, one strength of our theorem is to provide a PAC-Bayesian bound valid for any measure verifying a single hypothesis on K. The choice of ψ minimising the bound is still an open problem.

What this talk could have also been about

1. Online PAC-Bayes learning (`https://arxiv.org/pdf/2206.00024.pdf`)
2. PAC-Bayes for kernel PCA (https://arxiv.org/abs/2012.10369, to be updated)
3. Optimistic adaptation of classical online algorithms (online soon!)

Thank you for listening !