On generalisation and learning: some theoretical results for deep learning

Benjamin Guedj
https://bguedj.github.io
@bguedj

Université Laval
September 19th, 2024

The Alan Turing Institute

Inria

The Inria-London Programme

FRANCO-BRITISH COUNCIL

UCL

Express bio

Undergrad in pure mathematics, PhD in mathematical statistics [Sorbonne Université, 2011–2013].

Research Director at Inria (GENESIS project-team, Lille Nord Europe) and Professor of Machine Learning and Foundational Artificial Intelligence at University College London (Dept. of Computer Science and Centre for AI).

Turing Fellow of The Alan Turing Institute, Founder and Scientific director of Inria London.

Young Leader of the Franco-British Council, and Knight of the Order of the Academic Palms of the French Republic.

How it all started:

How it's going:

In a nutshell

Research at the crossroads of statistics, probability, machine learning, optimisation. ”Mathematical foundations of machine learning” says it all!

Statistical learning theory, PAC-Bayes, computational statistics, theoretical analysis of deep learning and representation learning, information theory...

In a nutshell

Research at the crossroads of statistics, probability, machine learning, optimisation. ”Mathematical foundations of machine learning” says it all!

Statistical learning theory, PAC-Bayes, computational statistics, theoretical analysis of deep learning and representation learning, information theory...

Personal obsession: **generalisation**.

In a nutshell

Research at the crossroads of statistics, probability, machine learning, optimisation. ”Mathematical foundations of machine learning” says it all!

Statistical learning theory, PAC-Bayes, computational statistics, theoretical analysis of deep learning and representation learning, information theory...

Personal obsession: **generalisation**.

Broad framework: foundational work on generalisation to contribute to frugal intelligent systems, in terms of data and/or compute.

In a nutshell

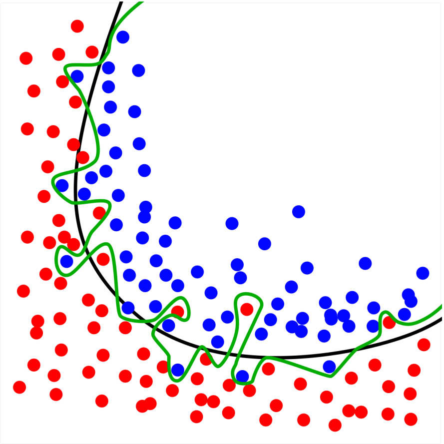
Research at the crossroads of statistics, probability, machine learning, optimisation. "Mathematical foundations of machine learning" says it all!

Statistical learning theory, PAC-Bayes, computational statistics, theoretical analysis of deep learning and representation learning, information theory...

Personal obsession: **generalisation**.

Broad framework: foundational work on generalisation to contribute to frugal intelligent systems, in terms of data and/or compute.
$\longrightarrow$ Project SHARP (PEPR IA) 2023-2029

Learning is to be able to generalise



[Credits: Wikipedia]

From examples, what can a system learn about the underlying phenomenon?

Memorising the already seen data is usually bad $\longrightarrow$ overfitting

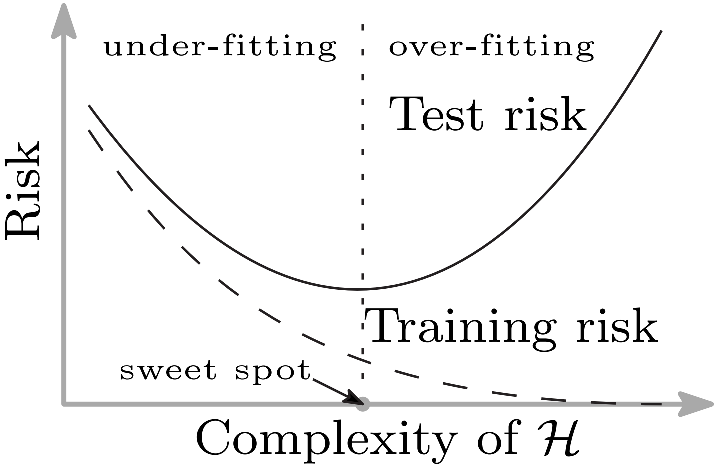
Generalisation is the ability to 'perform' well on unseen data.

Is deep learning breaking statistical learning theory?

Neural networks architectures trained on massive datasets achieve zero training error which does not bode well for their performance: this strongly suggests overfitting...

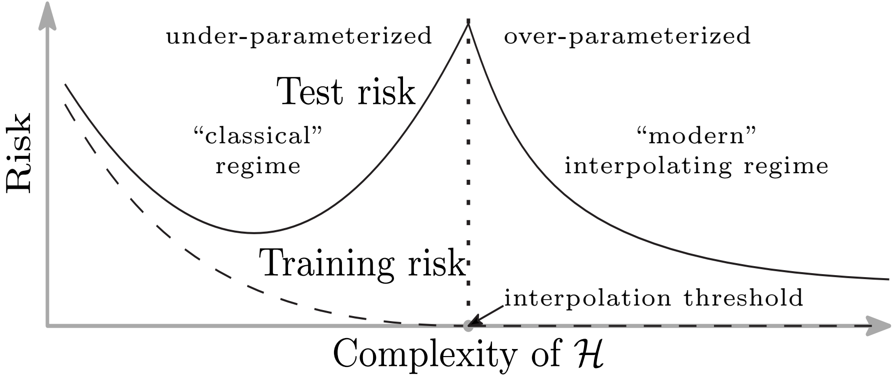
... yet they also achieve remarkably low errors on test sets!

A famous plot...



(Belkin et al., 2019)

... which might just be half of the picture



(Belkin et al., 2019)

Semantic representation to accelerate learning?

Fig. 1: What image representations do we learn by solving puzzles? Left: The image from which the tiles (marked with green lines) are extracted. Middle: A puzzle obtained by shuffling the tiles. Some tiles might be directly identifiable as object parts, but their identification is much more reliable once the correct ordering is found and the global figure emerges (Right).

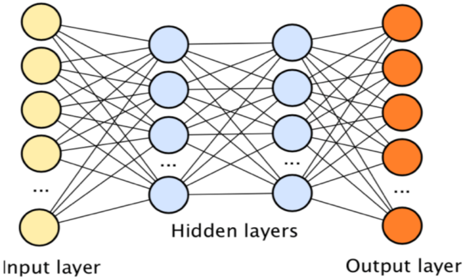
(Noroozi and Favaro, 2016)

Semantic content of data is key! $\longrightarrow$ MURI project (2018-2023)

A tale of two learners

A tale of two learners

First contender: a deep neural network



Typically identifies a specific item (say, a horse) in an image with accuracy $>$ 99%.
Training samples: millions of annotated images of horses – GPU-expensive training and significant environmental footprint.

A tale of two learners

Second contender: young children
(on this picture, aged 1 and 3)

A tale of two learners

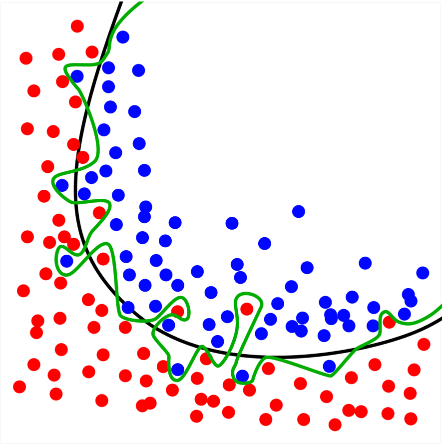
Second contender: young children
(on this picture, aged 1 and 3)

Identify horses with 100% accuracy. Also very good at transferring to *e.g.* zebras

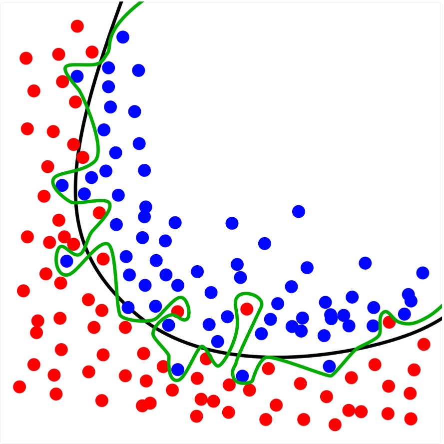
Training samples: a handful of children books, bedtime stories and (poorly executed) drawings.

Also expensive training.

Learning is to be able to generalise...



Learning is to be able to generalise...



... but not from scratch! Tackling each learning task as a fresh draw unlikely to be efficient – must not be blind to context.

Need to incorporate structure / semantic information / implicit representations of the ”sensible” world.

Should lead to better algorithms design (more ”intelligent”, frugal / resources-efficient, etc.)

Part I

A Primer on PAC-Bayesian Learning
(embarrassingly short version of our ICML 2019 tutorial
with John Shawe-Taylor)

https://bguedj.github.io/icml2019/index.html

The simplest setting

Learning algorithm $A : \mathcal{Z}^m \to \mathcal{H}$

- $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
- $\mathcal{H}$ = hypothesis class

Training set (aka sample): $S_m = ((X_1, Y_1), \ldots, (X_m, Y_m))$
a finite sequence of input-output examples.

- Data-generating distribution $\mathbb{P}$ over $\mathcal{Z}$.
- Learner doesn't know $\mathbb{P}$, only sees the training set.
- The training set examples are *i.i.d.* from $\mathbb{P}$: $S_m \sim \mathbb{P}^m$

Statistical Learning Theory is about high confidence

For a fixed algorithm, function class and sample size, generating random samples $\longrightarrow$ distribution of test errors

Statistical Learning Theory is about high confidence

For a fixed algorithm, function class and sample size, generating random samples $\longrightarrow$ distribution of test errors

- Focusing on the mean of the error distribution?
  - ▷ can be misleading: learner only has one sample

Statistical Learning Theory is about high confidence

For a fixed algorithm, function class and sample size, generating random samples $\longrightarrow$ distribution of test errors

- Focusing on the mean of the error distribution?
  - ▷ can be misleading: learner only has one sample
- Statistical Learning Theory: tail of the distribution
  - ▷ finding bounds which hold with high probability over random samples of size m

Statistical Learning Theory is about high confidence

For a fixed algorithm, function class and sample size, generating random samples $\longrightarrow$ distribution of test errors

- Focusing on the mean of the error distribution?
  - ▷ can be misleading: learner only has one sample
- Statistical Learning Theory: tail of the distribution
  - ▷ finding bounds which hold with high probability over random samples of size m
- Compare to a statistical test – at 99% confidence level
  - ▷ chances of the conclusion not being true are less than 1%

Statistical Learning Theory is about high confidence

For a fixed algorithm, function class and sample size, generating random samples $\longrightarrow$ distribution of test errors

- Focusing on the mean of the error distribution?

  ▷ can be misleading: learner only has one sample

- Statistical Learning Theory: tail of the distribution

  ▷ finding bounds which hold with high probability over random samples of size m

- Compare to a statistical test – at 99% confidence level

  ▷ chances of the conclusion not being true are less than 1%

- PAC: probably approximately correct (Valiant, 1984)
  Use a 'confidence parameter' δ: $\mathbb{P}^m[\text{large error}] \leqslant \delta$
  δ is the probability of being misled by the training set

Statistical Learning Theory is about high confidence

For a fixed algorithm, function class and sample size, generating random samples $\longrightarrow$ distribution of test errors

- Focusing on the mean of the error distribution?
  - ▷ can be misleading: learner only has one sample
- Statistical Learning Theory: tail of the distribution
  - ▷ finding bounds which hold with high probability over random samples of size m
- Compare to a statistical test – at 99% confidence level
  - ▷ chances of the conclusion not being true are less than 1%
- PAC: probably approximately correct (Valiant, 1984)
  Use a 'confidence parameter' δ: $\mathbb{P}^m[\text{large error}] \leqslant \delta$
  δ is the probability of being misled by the training set
- Hence high confidence: $\mathbb{P}^m[\text{approximately correct}] \geqslant 1 - \delta$

What to achieve from the sample?

What to achieve from the sample?

Use the available sample to:

1. learn a predictor
2. certify the predictor's performance

What to achieve from the sample?

Use the available sample to:

1. learn a predictor
2. certify the predictor's performance

Learning a predictor:

- algorithm driven by some learning principle
- informed by prior knowledge resulting in inductive bias

What to achieve from the sample?

Use the available sample to:

1. learn a predictor
2. certify the predictor's performance

Learning a predictor:

- algorithm driven by some learning principle
- informed by prior knowledge resulting in inductive bias

Certifying performance:

- what happens beyond the training set
- generalisation bounds

What to achieve from the sample?

Use the available sample to:

1. learn a predictor
2. certify the predictor's performance

Learning a predictor:

- algorithm driven by some learning principle
- informed by prior knowledge resulting in inductive bias

Certifying performance:

- what happens beyond the training set
- generalisation bounds

Actually these two goals interact with each other!

Generalisation

Loss function $\ell(h(X), Y)$ to measure the discrepancy between a predicted output $h(X)$ and the true output Y.

Empirical risk: $R_{\text{in}}(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h(X_i), Y_i)$
(in-sample)

Theoretical risk: $R_{\text{out}}(h) = \mathbb{E}\big[\ell(h(X), Y)\big]$
(out-of-sample)

If predictor h does well on the in-sample (X, Y) pairs...

...will it still do well on out-of-sample pairs?

Generalisation gap: $\Delta(h) = R_{\text{out}}(h) - R_{\text{in}}(h)$

Upper bounds: with high probability $\Delta(h) \leqslant \epsilon(m, \delta)$

▶ $R_{\text{out}}(h) \leqslant R_{\text{in}}(h) + \epsilon(m, \delta)$

Flavours:

- distribution-free
- algorithm-free
- distribution-dependent
- algorithm-dependent

The PAC (Probably Approximately Correct) framework

In a nutshell: with high probability, the generalisation error of an hypothesis h is at most something we can control and even compute. For any $\delta > 0$,

$$\mathbb{P}\left[R_{\mathrm{out}}(h) \leqslant R_{\mathrm{in}}(h) + \epsilon(m, \delta)\right] \geqslant 1 - \delta.$$

Think of $\epsilon(m, \delta)$ as $\mathrm{Complexity} \times \frac{\log \frac{1}{\delta}}{\sqrt{m}}$.

This is about high confidence statements on the tail of the distribution of test errors (compare to a statistical test at level $1 - \delta$).

PAC-Bayes is about PAC generalisation bounds for ***distributions over hypotheses***.

"Why should I care about generalisation?"

Generalisation bounds are a safety check: they give a theoretical guarantee on the performance of a learning algorithm on any unseen data.

Generalisation bounds:

- provide a computable control on the error on any unseen data with prespecified confidence
- explain why some specific learning algorithms actually work
- and even lead to designing new algorithms which scale to more complex settings

Take-home message

PAC-Bayes is a generic framework to efficiently rethink generalisation for numerous statistical learning algorithms. It leverages the flexibility of Bayesian inference and allows to derive new learning algorithms.

Some (shameless self-)pointers

◇ New monograph Hellström, Durisi, Guedj and Raginsky (2024), ”Generalization Bounds: Perspectives from Information Theory and PAC-Bayes” https://arxiv.org/abs/2309.04381

◇ New ICML 2023 workshop ”PAC-Bayes meets interactive learning” https://bguedj.github.io/icml2023-workshop/

◇ ICML 2019 tutorial ”A Primer on PAC-Bayesian Learning” https://bguedj.github.io/icml2019/

◇ Survey in the Journal of the French Mathematical Society: Guedj (2019) https://arxiv.org/abs/1901.05353

◇ NeurIPS 2017 workshop ”(Almost) 50 Shades of Bayesian Learning: PAC-Bayesian trends and insights” https://bguedj.github.io/nips2017/

Before PAC-Bayes

- Single hypothesis h (building block):
  with probability $\geqslant 1 - \delta, \quad R_{\mathrm{out}}(h) \leqslant R_{\mathrm{in}}(h) + \sqrt{\frac{1}{2m} \log\left(\frac{1}{\delta}\right)}$.

Before PAC-Bayes

- Single hypothesis h (building block):

  with probability $\geqslant 1-\delta, \quad R_{\text{out}}(h) \leqslant R_{\text{in}}(h) + \sqrt{\frac{1}{2m}\log\left(\frac{1}{\delta}\right)}$.

- Finite function class $\mathcal{H}$ (worst-case approach):

  w.p. $\geqslant 1-\delta, \quad \forall h \in \mathcal{H}, \quad R_{\text{out}}(h) \leqslant R_{\text{in}}(h) + \sqrt{\frac{1}{2m}\log\left(\frac{|\mathcal{H}|}{\delta}\right)}$

- Structural risk minimisation: data-dependent hypotheses h_i associated with prior weight p_i

  w.p. $\geqslant 1-\delta, \quad \forall h_i \in \mathcal{H}, \quad R_{\text{out}}(h_i) \leqslant R_{\text{in}}(h_i) + \sqrt{\frac{1}{2m}\log\left(\frac{1}{p_i\delta}\right)}$

- Uncountably infinite function class: VC dimension, Rademacher complexity...

Before PAC-Bayes

- Single hypothesis h (building block):

  with probability $\geqslant 1 - \delta, \quad R_{\text{out}}(h) \leqslant R_{\text{in}}(h) + \sqrt{\frac{1}{2m} \log\left(\frac{1}{\delta}\right)}.$

- Finite function class $\mathcal{H}$ (worst-case approach):

  w.p. $\geqslant 1 - \delta, \quad \forall h \in \mathcal{H}, \quad R_{\text{out}}(h) \leqslant R_{\text{in}}(h) + \sqrt{\frac{1}{2m} \log\left(\frac{|\mathcal{H}|}{\delta}\right)}$

- Structural risk minimisation: data-dependent hypotheses h_i associated with prior weight p_i

  w.p. $\geqslant 1 - \delta, \quad \forall h_i \in \mathcal{H}, \quad R_{\text{out}}(h_i) \leqslant R_{\text{in}}(h_i) + \sqrt{\frac{1}{2m} \log\left(\frac{1}{p_i \delta}\right)}$

- Uncountably infinite function class: VC dimension, Rademacher complexity...

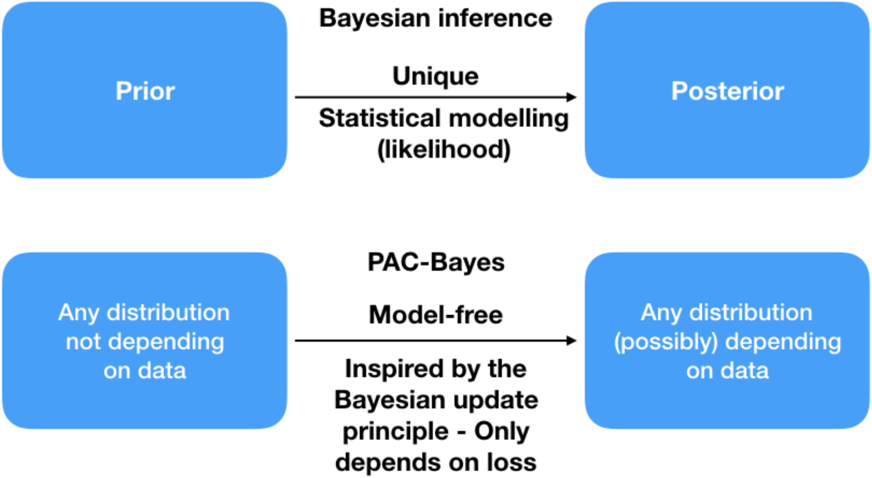
These approaches are suited to analyse the performance of individual functions, and take some account of correlations.

Before PAC-Bayes

- Single hypothesis h (building block):
  with probability $\geqslant 1-\delta, \quad R_{\text{out}}(h) \leqslant R_{\text{in}}(h) + \sqrt{\frac{1}{2m}\log\left(\frac{1}{\delta}\right)}$.
- Finite function class $\mathcal{H}$ (worst-case approach):
  w.p. $\geqslant 1-\delta, \quad \forall h \in \mathcal{H}, \quad R_{\text{out}}(h) \leqslant R_{\text{in}}(h) + \sqrt{\frac{1}{2m}\log\left(\frac{|\mathcal{H}|}{\delta}\right)}$
- Structural risk minimisation: data-dependent hypotheses h_i associated with prior weight p_i
  w.p. $\geqslant 1-\delta, \quad \forall h_i \in \mathcal{H}, \quad R_{\text{out}}(h_i) \leqslant R_{\text{in}}(h_i) + \sqrt{\frac{1}{2m}\log\left(\frac{1}{p_i\delta}\right)}$
- Uncountably infinite function class: VC dimension, Rademacher complexity...

These approaches are suited to analyse the performance of individual functions, and take some account of correlations.
$\longrightarrow$ Extension: PAC-Bayes allows to consider *distributions* over hypotheses.

PAC-Bayes



"Prior": exploration mechanism of $\mathcal{H}$

"Posterior" is the twisted prior after confronting with data

PAC-Bayes bounds vs. Bayesian inference

Prior P, posterior $Q \ll P$. Define the risk of a distribution:

$$R_{\text{in}}(Q) \equiv \int_{\mathcal{H}} R_{\text{in}}(h)\, dQ(h) \qquad R_{\text{out}}(Q) \equiv \int_{\mathcal{H}} R_{\text{out}}(h)\, dQ(h)$$

Kullback-Leibler divergence $\mathrm{KL}(Q\|P) = \mathop{\mathbf{E}}_{h\sim Q} \ln \frac{Q(h)}{P(h)}$.

- **Prior**
  - PAC-Bayes: bounds hold for any distribution
  - Bayes: prior choice impacts inference
- **Posterior**
  - PAC-Bayes: bounds hold for any distribution
  - Bayes: posterior uniquely defined by prior and statistical model
- **Data distribution**
  - PAC-Bayes: bounds hold for any distribution
  - Bayes: statistical modelling choices impact inference

A classical PAC-Bayesian bound

Pre-history: PAC analysis of Bayesian estimators
(Shawe-Taylor and Williamson, 1997)

Birth: PAC-Bayesian bound
(McAllester, 1998, 1999)

Prototypical bound

For any prior P, any $\delta \in (0, 1]$, we have

$$\mathbb{P}^m\left(\forall\, Q \text{ on } \mathcal{H}\colon\ R_{\mathrm{out}}(Q) \leqslant R_{\mathrm{in}}(Q) + \sqrt{\frac{\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{m}}{\delta}}{2m}}\right) \geqslant 1 - \delta\,.$$

PAC-Bayes-driven learning algorithms

With an arbitrarily high probability and for any posterior distribution Q,

$$\text{Error on unseen data} \leqslant \text{Error on sample} + \text{complexity term}$$

$$R_{\text{out}}(Q) \leqslant R_{\text{in}}(Q) + F(Q, \cdot).$$

This defines a principled strategy to obtain new learning algorithms:

$$h \sim Q^\star$$

$$Q^\star \in \operatorname*{arg\,inf}_{Q \ll P} \left\{ R_{\text{in}}(Q) + F(Q, \cdot) \right\}$$

(optimisation problem which can be solved or approximated by [stochastic] gradient descent-flavoured methods, Monte Carlo Markov Chain, variational inference...)

SVMs, KL-regularized Adaboost, exponential weights are all minimisers of PAC-Bayes bounds.

Recap

What we’ve seen so far

Recap

What we've seen so far

- Statistical learning theory is about high confidence control of generalisation

Recap

What we've seen so far

- Statistical learning theory is about high confidence control of generalisation
- PAC-Bayes is a generic, powerful tool to derive generalisation bounds...

Recap

What we've seen so far

- Statistical learning theory is about high confidence control of generalisation
- PAC-Bayes is a generic, powerful tool to derive generalisation bounds...
- ... and invent new learning algorithms with a Bayesian flavour

Recap

What we've seen so far

- Statistical learning theory is about high confidence control of generalisation
- PAC-Bayes is a generic, powerful tool to derive generalisation bounds...
- ... and invent new learning algorithms with a Bayesian flavour
- PAC-Bayes mixes tools from statistics, probability theory, optimisation, and is now quickly re-emerging as a key theory and practical framework in machine learning (and in particular deep learning)

Recap

What we've seen so far

- Statistical learning theory is about high confidence control of generalisation
- PAC-Bayes is a generic, powerful tool to derive generalisation bounds...
- ... and invent new learning algorithms with a Bayesian flavour
- PAC-Bayes mixes tools from statistics, probability theory, optimisation, and is now quickly re-emerging as a key theory and practical framework in machine learning (and in particular deep learning)

What is coming next

Recap

What we've seen so far

- Statistical learning theory is about high confidence control of generalisation
- PAC-Bayes is a generic, powerful tool to derive generalisation bounds...
- ... and invent new learning algorithms with a Bayesian flavour
- PAC-Bayes mixes tools from statistics, probability theory, optimisation, and is now quickly re-emerging as a key theory and practical framework in machine learning (and in particular deep learning)

What is coming next

- A biased sample of our many contributions to PAC-Bayes!

Part II

The PAC-Bayes frontline

(focus on deep learning)

- Journal of Statistical Planning and Inference, Guedj and Robbiano (2018). PAC-Bayesian high dimensional bipartite ranking
- Machine Learning, Alquier and Guedj (2018). Simpler PAC-Bayesian bounds for hostile data
- NeurIPS 2019, Mhammedi, Grünwald and Guedj (2019). PAC-Bayes Un-Expected Bernstein Inequality
- NeurIPS 2019, Letarte, Germain, Guedj and Laviolette (2019). **Dichotomize and generalize: PAC-Bayesian binary activated deep neural networks**
- UAI 2020, Nozawa, Germain and Guedj (2020). PAC-Bayesian contrastive unsupervised representation learning
- preprint, Cantelobre, Guedj, Perez-Ortiz and Shawe-Taylor (2020). A PAC-Bayesian Perspective on Structured Prediction with Implicit Loss Embeddings
- NeurIPS 2020 (spotlight), Mhammedi, Guedj and Williamson (2020). PAC-Bayesian Bound for the Conditional Value at Risk
- Entropy, Haddouche, Guedj, Rivasplata and Shawe-Taylor (2021). PAC-Bayes unleashed: generalisation bounds with unbounded losses
- Entropy, Guedj and Pujol (2021). Still no free lunches: the price to pay for tighter PAC-Bayes bounds
- Entropy, Biggs and Guedj (2021). Differentiable PAC-Bayes Objectives with Partially Aggregated Neural Networks
- NeurIPS 2021, Zantedeschi, Viallard, Morvant, Emonet, Habrard, Germain and Guedj (2021). Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Bound
- preprint, Perez-Ortiz, Rivasplata, Guedj, Gleeson, Zhang, Shawe-Taylor, Bober and Kittler (2021). Learning PAC-Bayes Priors for Probabilistic Neural Networks
- AISTATS 2022, Biggs and Guedj (2022). **On Margins and Derandomisation in PAC-Bayes**
- AISTATS 2022, Cherief-Abdellatif, Shi, Doucet and Guedj (2022). On PAC-Bayesian reconstruction guarantees for VAEs
- ICML 2022, Biggs and Guedj (2022). **Non-Vacuous Generalisation Bounds for Shallow Neural Networks**
- preprint, Adams, Shawe-Taylor and Guedj (2022). Controlling Confusion via Generalisation Bounds
- preprint, Picard-Weibel and Guedj (2022). On change of measure inequalities for f-divergences
- NeurIPS 2022, Biggs, Zandeteschi and Guedj (2022). On Margins and Generalisation for Voting Classifiers
- NeurIPS 2022, Haddouche and Guedj (2022). Online PAC-Bayesian Learning
- preprint, Clerico, Deligiannidis, Guedj and Doucet (2022). A PAC-Bayes bound for deterministic classifiers
- TMLR, Haddouche and Guedj (2023). PAC-Bayes with Unbounded Losses through Supermartingales
- AISTATS 2023, Biggs and Guedj (2023). Tighter PAC-Bayes Generalisation Bounds by Leveraging Example Difficulty
- preprint, Haddouche and Guedj (2023). Wasserstein PAC-Bayes Learning: Exploiting Optimisation Guarantees to Explain Generalisation
- NeurIPS 2023, Viallard, Haddouche, Şimşekli and Guedj. Learning via Wasserstein-Based High Probability Generalisation Bounds
- Foundations and Trends in Machine Learning, Hellström, Durisi, Guedj and Raginsky (2023). Generalization Bounds: Perspectives from Information Theory and PAC-Bayes
- AISTATS 2024, Hellström and Guedj (2023). Comparing Comparators in Generalization Bounds
- preprint, Jobic, Haddouche and Guedj (2023). Federated Learning with Nonvacuous Generalisation Bounds

Some of my partners in crime

Binary Activated Neural Networks (NeurIPS 2019)

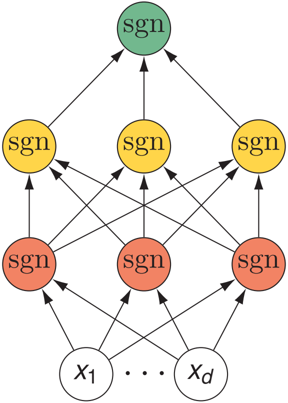
  

$\mathbf{x} \in \mathbb{R}^{d_0}$, $y \in \{-1, 1\}$. Architecture:

- L *fully connected* layers, d_k denotes the number of neurons of the k^{th} layer
- $\mathrm{sgn}(a) = 1$ if $a > 0$ and $\mathrm{sgn}(a) = -1$ otherwise

Parameters:

- $\mathbf{W}_k \in \mathbb{R}^{d_k \times d_{k-1}}$ denotes the weight matrices.
- $\theta = \mathrm{vec}\left(\{\mathbf{W}_k\}_{k=1}^{L}\right) \in \mathbb{R}^D$



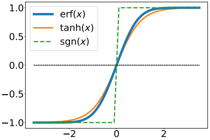
Prediction

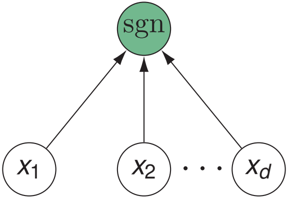
$$f_\theta(\mathbf{x}) = \mathrm{sgn}\left(\mathbf{w}_L \mathrm{sgn}\left(\mathbf{W}_{L-1} \mathrm{sgn}\left(\ldots \mathbf{sgn}\left(\mathbf{W}_1 \mathbf{x}\right)\right)\right)\right),$$

Building block: one layer (aka linear predictor)

Model $f_{\mathrm{w}}(\mathrm{x}) \stackrel{\text{def}}{=} \mathrm{sgn}(\mathbf{w} \cdot \mathbf{x})$, with $\mathbf{w} \in \mathbb{R}^d$.

- Linear classifiers $\mathcal{F}_d \stackrel{\text{def}}{=} \{f_{\mathbf{v}} | \mathbf{v} \in \mathbb{R}^d\}$
- Predictor $F_{\mathrm{w}}(\mathrm{x}) \stackrel{\text{def}}{=} \mathbf{E}_{\mathbf{v} \sim Q_{\mathbf{w}}} f_{\mathbf{v}}(\mathbf{x}) = \mathrm{erf}\left(\frac{\mathbf{w} \cdot \mathbf{x}}{\sqrt{d}\|\mathbf{x}\|}\right)$
- Sampling + closed form of the KL + a few other tricks + extension to an arbitrary number of layers





Generalisation bound

Let F_θ denote the network with parameter θ. With probability at least $1-\delta$, for any $\theta \in \mathbb{R}^D$

$$R_{\mathrm{out}}(F_\theta) \leqslant \inf_{C>0} \left\{ \frac{1}{1-e^{-C}} \left(1 - \exp\left(-C R_{\mathrm{in}}(F_\theta) - \frac{\mathrm{KL}(\theta, \theta_0) + \log \frac{2\sqrt{m}}{\delta}}{m} \right) \right) \right\}.$$

Numerical experiments

Model name	Cost function	Train split	Valid split	Model selection	Prior
MLP–tanh	linear loss, L2 regularized	80%	20%	valid linear loss	-
PBGNet$_\ell$	linear loss, L2 regularized	80%	20%	valid linear loss	random init
PBGNet	**PAC-Bayes bound**	**100 %**	**-**	**PAC-Bayes bound**	**random init**
PBGNet$_{pre}$					
– pretrain	linear loss (20 epochs)	50%	-	-	random init
– final	PAC-Bayes bound	50%	-	PAC-Bayes bound	pretrain

Dataset	MLP–tanh		PBGNet$_\ell$		PBGNet			PBGNet$_{pre}$		
	R_{in}	R_{out}	R_{in}	R_{out}	R_{in}	R_{out}	Bound	R_{in}	R_{out}	Bound
ads	0.021	0.037	0.018	**0.032**	0.024	0.038	0.283	0.034	0.033	0.058
adult	0.128	0.149	0.136	**0.148**	0.158	0.154	0.227	0.153	0.151	0.165
mnist17	0.003	**0.004**	0.008	0.005	0.007	0.009	0.067	0.003	0.005	0.009
mnist49	0.002	**0.013**	0.003	0.018	0.034	0.039	0.153	0.018	0.021	0.030
mnist56	0.002	0.009	0.002	0.009	0.022	0.026	0.103	0.008	**0.008**	0.017
mnistLH	0.004	**0.017**	0.005	0.019	0.071	0.073	0.186	0.026	0.026	0.033

On Margins and Derandomisation in PAC-Bayes

AISTATS 2022

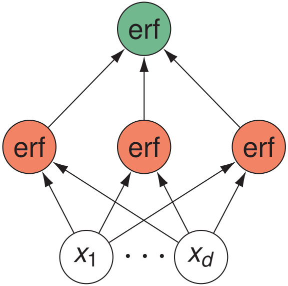
We provide a unified framework for derandomising PAC-Bayes bounds with margins, leading to new bounds or greatly simplified proofs for

- L_2 and L_1 normed linear predictors,
- Linear predictors with a learned randomised feature space,
- One-hidden-layer neural networks with erf activations,
- Deep ReLU networks.

Key idea: PAC-Bayes bounds are (mostly) SOTA, but apply for non-deterministic randomised predictions. Large margin deterministic predictors give similar predictive performance to their randomised counterparts.

SHEL: An unusual neural architecture

Binary $\mathcal{Y} = \{\pm 1\}$ or multiclass classification $\mathcal{Y} = \{1, \ldots, c\}$. Predictors f are score-valued: $f(x) \in \mathbb{R}^c$ (multiclass) or $f(x) \in \mathbb{R}$ (binary). We define the binary margin $M_{\text{bin}}(f, (x, y)) = yf(x)$ and multiclass margin $M_{\text{multi}}(f, (x, y)) = f(x)[y] - \max_{k \neq y} f(x)[k]$.



$$R_{\text{out}}(f) = \Pr\{(x, y) : M(f, (x, y)) \leqslant 0\},$$
$$R_{\text{in},\gamma}(f) = m^{-1}|\{(x, y) \in S : M(f, (x, y)) \leqslant \gamma\}|.$$

Single Hidden Erf Layer (SHEL) network: elementwise error function activations $F_{U,V}(x) = V\text{erf}\left(\frac{Ux}{\sqrt{2}||x||_2}\right)$.

Some of our results

Theorem. For the SHEL network with K hidden units, and any margin γ,

$$R_{\text{out}}(F_{U,V}) \leqslant R_{\text{in},\gamma}(F_{U,V}) + \widetilde{\mathcal{O}}\left(\frac{\sqrt{K}}{\gamma\sqrt{m}}(\|V\|_{\max}\|U - U^0\|_F + \|V\|_F)\right).$$

Theorem. Let F be a fully-connected feed-forward ReLU neural network with d layers and no more than h neurons per layer. We assume bounded inputs and bounded spectral norm of the weight matrices, then for any margin γ, with probability at least $1 - \delta$,

$$R_{\text{out}}(F) \leqslant R_{\text{in},\gamma}(F) + \widetilde{\mathcal{O}}\left(\sqrt{\frac{hC\log(mdh)}{\gamma^2 m}\sum_{i=1}^{d}\frac{\|W_i - W_i^0\|_F^2}{\|W_i\|_2^2} + \frac{\log 1/\delta + d\log\log C}{m}}\right).$$

Non-vacuous Generalisation bounds for Shallow Neural Networks

ICML 2022

Goal: Non-vacuous bounds for neural networks which are trained via standard SGD, with no randomisation of outputs.

We manage to do this for single hidden layer networks with erf or GELU activations ($\mathrm{GELU}(x) = \frac{x}{2}(1 + \mathrm{erf}(x/\sqrt{2}))$)

Key idea: express the network as a PAC-Bayesian majority vote!

	Data	Test Err	Bound	Data-dependent prior
ERF	Binary-MNIST	0.038	0.837	0.286
ERF	Binary-Fashion	0.085	0.426	0.297
ERF	MNIST	0.046	0.772	0.490
ERF	Fashion	0.150	0.984	0.727
GELU	MNIST	0.043	0.693	0.293
GELU	Fashion	0.153	0.976	0.568

An attempt at summarising my research

Quest for generalisation guarantees (about half *via* PAC-Bayes)

Directions:

- Generic bounds (relaxing assumptions such as iid or boundedness, new concentration inequalities, . . .)
- Tight bounds for self-certifying specific algorithms (deep neural networks, NMF, . . .)
- Towards new measures of performance (CVaR, ranking, contrastive losses, . . .)
- Coupling theory and implemented algorithms: bound-driven algorithms
- Impact beyond learning theory (providing guidelines to machine learning users, sustainable / frugal machine learning)

Thanks!

What this talk could have been about...

- Tighter PAC-Bayes bounds (Mhammedi et al., 2019)
- PAC-Bayes for conditional value at risk (Mhammedi et al., 2020)
- PAC-Bayes-driven deep neural networks (Biggs and Guedj, 2021, 2022; Pérez-Ortiz et al., 2021a,b)
- PAC-Bayes and robust learning (Guedj and Pujol, 2021)
- PAC-Bayes for unbounded losses (Haddouche et al., 2021)
- PAC-Bayesian online clustering (Li et al., 2018)
- PAC-Bayesian bipartite ranking (Guedj and Robbiano, 2018)
- Online k-means clustering (Cohen-Addad et al., 2021)
- Sequential learning of principal curves (Li and Guedj, 2021)
- PAC-Bayes for heavy-tailed, dependent data (Alquier and Guedj, 2018)
- Stability and generalisation (Celisse and Guedj, 2016)
- Additive regression (Guedj and Alquier, 2013)
- Stochastic majority votes (Zantedeschi et al., 2021)
- Dynamic regret bounds (Haddouche et al., 2023)

+ a few more in the pipe, soon on arXiv

- Contrastive unsupervised learning (Nozawa et al., 2020)
- Generalisation bounds for structured prediction (Cantelobre et al., 2020)
- MMD aggregated two sample tests (Schrab et al., 2023, 2022a,b)
- Image denoising (Guedj and Rengot, 2020)
- Data augmentation (Wei et al., 2022), invariance principles (Cantelobre et al., 2022)
- Matrix factorisation (Alquier and Guedj, 2017; Chrétien and Guedj, 2020)
- Preventing model overfitting (Zhang et al., 2023)
- Decentralised learning with aggregation (Klein et al., 2020)
- Ensemble learning and nonlinear aggregation (Biau et al., 2016) in Python (Guedj and Srinivasa Desikan, 2018, 2020)
- Identifying subcommunities in social networks and application to forecasting elections (Vendeville et al., 2021, 2022)
- Upper and lower bounds for kernel PCA (Haddouche et al., 2020)
- Prediction with multi-task Gaussian processes (Leroy et al., 2022, 2023)

NOW HIRING

`https://bguedj.github.io`

`@bguedj`

References I

P. Alquier and B. Guedj. An oracle inequality for quasi-Bayesian nonnegative matrix factorization. *Mathematical Methods of Statistics*, 26(1):55–67, 2017. ISSN 1934-8045. doi: 10.3103/S1066530717010045. URL https://link.springer.com/article/10.3103%2FS1066530717010045.

P. Alquier and B. Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018. ISSN 1573-0565. doi: 10.1007/s10994-017-5690-0. URL https://doi.org/10.1007/s10994-017-5690-0.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1903070116. URL https://www.pnas.org/content/116/32/15849.

G. Biau, A. Fischer, B. Guedj, and J. D. Malley. COBRA: A combined regression strategy. *Journal of Multivariate Analysis*, 146: 18–28, 2016. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2015.04.007. URL http://www.sciencedirect.com/science/article/pii/S0047259X15000950. Special Issue on Statistical Models and Methods for High or Infinite Dimensional Spaces.

F. Biggs and B. Guedj. Differentiable PAC-Bayes objectives with partially aggregated neural networks. *Entropy*, 23(10), 2021. ISSN 1099-4300. doi: 10.3390/e23101280. URL https://www.mdpi.com/1099-4300/23/10/1280.

F. Biggs and B. Guedj. On margins and derandomisation in PAC-Bayes. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics [AISTATS]*, volume 151 of *Proceedings of Machine Learning Research*, pages 3709–3731. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/biggs22a.html.

T. Cantelobre, B. Guedj, M. Pérez-Ortiz, and J. Shawe-Taylor. A PAC-Bayesian perspective on structured prediction with implicit loss embeddings. Submitted., 2020. URL https://arxiv.org/abs/2012.03780.

T. Cantelobre, C. Ciliberto, B. Guedj, and A. Rudi. Measuring dissimilarity with diffeomorphism invariance. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning [ICML]*, volume 162 of *Proceedings of Machine Learning Research*, pages 2572–2596. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/cantelobre22a.html.

O. Catoni. *Statistical Learning Theory and Stochastic Optimization*. École d'Été de Probabilités de Saint-Flour 2001. Springer, 2004.

A. Celisse and B. Guedj. Stability revisited: new generalisation bounds for the Leave-one-Out. Preprint., 2016. URL https://arxiv.org/abs/1608.06412.

References II

S. Chrétien and B. Guedj. Revisiting clustering as matrix factorisation on the Stiefel manifold. In G. Nicosia, V. Ojha, E. La Malfa, G. Jansen, V. Sciacca, P. Pardalos, G. Giuffrida, and R. Umeton, editors, *LOD – The Sixth International Conference on Machine Learning, Optimization, and Data Science*, pages 1–12. Springer International Publishing, 2020. ISBN 978-3-030-64583-0. doi: 10.1007/978-3-030-64583-0_1. URL https://link.springer.com/chapter/10.1007%2F978-3-030-64583-0_1.

V. Cohen-Addad, B. Guedj, V. Kanade, and G. Rom. Online k-means clustering. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics [AISTATS]*, volume 130 of *Proceedings of Machine Learning Research*, pages 1126–1134. PMLR, April 2021. URL http://proceedings.mlr.press/v130/cohen-addad21a.html.

I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3:146–158, 1975.

M. D. Donsker and S. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28, 1975.

B. Guedj. A primer on PAC-Bayesian learning. In *Proceedings of the second congress of the French Mathematical Society*, volume 33, 2019. URL https://arxiv.org/abs/1901.05353.

B. Guedj and P. Alquier. PAC-Bayesian estimation and prediction in sparse additive models. *Electron. J. Statist.*, 7:264–291, 2013. doi: 10.1214/13-EJS771. URL https://doi.org/10.1214/13-EJS771.

B. Guedj and L. Pujol. Still no free lunches: the price to pay for tighter PAC-Bayes bounds. *Entropy*, 23(11), 2021. ISSN 1099-4300. doi: 10.3390/e23111529. URL https://www.mdpi.com/1099-4300/23/11/1529.

B. Guedj and J. Rengot. Non-linear aggregation of filters to improve image denoising. In K. Arai, S. Kapoor, and R. Bhatia, editors, *SAI: Intelligent Computing*, pages 314–327. Springer International Publishing, 2020. ISBN 978-3-030-52246-9. doi: 10.1007/978-3-030-52246-9_22. URL https://link.springer.com/chapter/10.1007%2F978-3-030-52246-9_22.

B. Guedj and S. Robbiano. PAC-Bayesian high dimensional bipartite ranking. *Journal of Statistical Planning and Inference*, 196:70 – 86, 2018. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2017.10.010. URL http://www.sciencedirect.com/science/article/pii/S0378375817301945.

B. Guedj and B. Srinivasa Desikan. Pycobra: A Python toolbox for ensemble learning and visualisation. *Journal of Machine Learning Research*, 18(190):1–5, 2018. URL http://jmlr.org/beta/papers/v18/17-228.html.

References III

B. Guedj and B. Srinivasa Desikan. Kernel-based ensemble learning in Python. *Information*, 11(2):63, Jan 2020. ISSN 2078-2489. doi: 10.3390/info11020063. URL http://dx.doi.org/10.3390/info11020063.

M. Haddouche, B. Guedj, and J. Shawe-Taylor. Upper and lower bounds on the performance of Kernel PCA. Submitted., 2020. URL https://arxiv.org/abs/2012.10369.

M. Haddouche, B. Guedj, O. Rivasplata, and J. Shawe-Taylor. PAC-Bayes unleashed: generalisation bounds with unbounded losses. *Entropy*, 23(10), 2021. ISSN 1099-4300. doi: 10.3390/e23101330. URL https://www.mdpi.com/1099-4300/23/10/1330.

M. Haddouche, B. Guedj, and O. Wintenberger. Optimistically tempered online learning. Submitted., 2023. URL https://arxiv.org/abs/2301.07530.

J. Klein, M. Albardan, B. Guedj, and O. Colot. Decentralized learning with budgeted network load using Gaussian copulas and classifier ensembles. In P. Cellier and K. Driessens, editors, *ECML-PKDD 2019: Machine Learning and Knowledge Discovery in Databases*, pages 301–316. Springer International Publishing, 2020. ISBN 978-3-030-43823-4. doi: 10.1007/978-3-030-43823-4_26. URL https://link.springer.com/chapter/10.1007%2F978-3-030-43823-4_26.

A. Leroy, P. Latouche, B. Guedj, and S. Gey. MAGMA: Inference and prediction with multi-task Gaussian processes. *Machine Learning*, 2022. doi: 10.1007/s10994-022-06172-1. URL https://arxiv.org/abs/2007.10731.

A. Leroy, P. Latouche, B. Guedj, and S. Gey. Cluster-specific predictions with multi-task Gaussian processes. *Journal of Machine Learning Research [JMLR]*, 24(5):1–49, 2023. URL https://jmlr.org/papers/v24/20-1321.html.

L. Li and B. Guedj. Sequential learning of principal curves: Summarizing data streams on the fly. *Entropy*, 23(11), 2021. ISSN 1099-4300. doi: 10.3390/e23111534. URL https://www.mdpi.com/1099-4300/23/11/1534.

L. Li, B. Guedj, and S. Loustau. A quasi-Bayesian perspective to online clustering. *Electron. J. Statist.*, 12(2):3071–3113, 2018. doi: 10.1214/18-EJS1479. URL https://doi.org/10.1214/18-EJS1479.

D. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 1998.

D. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37, 1999.

References IV

Z. Mhammedi, P. Grünwald, and B. Guedj. PAC-Bayes un-expected Bernstein inequality. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems [NeurIPS] 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 12180–12191, 2019. URL `http://papers.nips.cc/paper/9387-pac-bayes-un-expected-bernstein-inequality`.

Z. Mhammedi, B. Guedj, and R. C. Williamson. PAC-Bayesian bound for the conditional value at risk. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems [NeurIPS] 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/d02e9bdc27a894e882fa0c9055c99722-Abstract.html`.

M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

K. Nozawa, P. Germain, and B. Guedj. PAC-Bayesian contrastive unsupervised representation learning. In *Conference on Uncertainty in Artificial Intelligence [UAI]*, 2020. URL `https://proceedings.mlr.press/v124/nozawa20a.html`.

M. Pérez-Ortiz, O. Rivasplata, B. Guedj, M. Gleeson, J. Zhang, J. Shawe-Taylor, M. Bober, and J. Kittler. Learning PAC-Bayes priors for probabilistic neural networks. Submitted., 2021a. URL `https://arxiv.org/abs/2109.10304`.

M. Pérez-Ortiz, O. Rivasplata, E. Parrado-Hernandez, B. Guedj, and J. Shawe-Taylor. Progress in self-certified neural networks. In *NeurIPS 2021 workshop Bayesian Deep Learning [BDL]*, 2021b. URL `http://bayesiandeeplearning.org/2021/papers/38.pdf`.

A. Schrab, B. Guedj, and A. Gretton. KSD aggregated goodness-of-fit test. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems [NeurIPS]*, volume 35, pages 32624–32638. Curran Associates, Inc., 2022a. URL `https://papers.nips.cc/paper_files/paper/2022/hash/d241a7b1499cee1bf40769ceade2444d-Abstract-Conference.html`.

A. Schrab, I. Kim, B. Guedj, and A. Gretton. Efficient aggregated kernel tests using incomplete *u*-statistics. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems [NeurIPS]*, volume 35, pages 18793–18807. Curran Associates, Inc., 2022b. URL `https://papers.nips.cc/paper_files/paper/2022/hash/774164b966cc277c82a960934445140d-Abstract-Conference.html`.

References V

A. Schrab, I. Kim, M. Albert, B. Laurent, B. Guedj, and A. Gretton. MMD aggregated two-sample test. *Journal of Machine Learning Research [JMLR]*, 24(194):1–81, 2023. URL https://jmlr.org/papers/v24/21-1289.html.

J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a Bayes estimator. In *Proceedings of the 10th annual conference on Computational Learning Theory*, pages 2–9. ACM, 1997. doi: 10.1145/267460.267466.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

A. Vendeville, B. Guedj, and S. Zhou. Forecasting elections results via the voter model with stubborn nodes. *Applied Network Science*, 6, 2021. doi: 10.1007/s41109-020-00342-7. URL https://appliednetsci.springeropen.com/articles/10.1007/s41109-020-00342-7.

A. Vendeville, B. Guedj, and S. Zhou. Towards control of opinion diversity by introducing zealots into a polarised social group. In R. M. Benito, C. Cherifi, H. Cherifi, E. Moro, L. M. Rocha, and M. Sales-Pardo, editors, *Complex Networks & Their Applications X*, pages 341–352. Springer International Publishing, 2022. ISBN 978-3-030-93413-2. doi: 10.1007/978-3-030-93413-2_29. URL https://link.springer.com/chapter/10.1007%2F978-3-030-93413-2_29.

J. Wei, Q. Chen, P. Peng, B. Guedj, and L. Li. Reprint: a randomized extrapolation based on principal components for data augmentation. Submitted., 2022. URL https://arxiv.org/abs/2204.12024.

V. Zantedeschi, P. Viallard, E. Morvant, R. Emonet, A. Habrard, P. Germain, and B. Guedj. Learning stochastic majority votes by minimizing a PAC-Bayes generalization bound. In A. Beygelzimer, P. Liang, J. W. Vaughan, and Y. Dauphin, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems [NeurIPS] 2021*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0415740eaa4d9decbc8da001d3fd805f-Abstract.html.

J. M. Zhang, M. Harman, B. Guedj, E. T. Barr, and J. Shawe-Taylor. Model validation using mutated training labels: An exploratory study. *Neurocomputing*, 539, 2023. doi: 10.1016/j.neucom.2023.02.042. URL https://www.sciencedirect.com/science/article/abs/pii/S0925231223001911.

Variational definition of KL-divergence (Csiszár, 1975; Donsker and Varadhan, 1975; Catoni, 2004).

Variational definition of KL-divergence (Csiszár, 1975; Donsker and Varadhan, 1975; Catoni, 2004).

Let $(A, \mathcal{A})$ be a measurable space.

(i) For any probability P on $(A, \mathcal{A})$ and any measurable function $\phi : A \to \mathbb{R}$ such that $\int (\exp \circ \phi) \mathrm{d}P < \infty$,

$$\log \int (\exp \circ \phi) \mathrm{d}P = \sup_{Q \ll P} \left\{ \int \phi \mathrm{d}Q - \mathrm{KL}(Q, P) \right\}.$$

(ii) If ϕ is upper-bounded on the support of P, the supremum is reached for the Gibbs distribution G given by

$$\frac{\mathrm{d}G}{\mathrm{d}P}(a) = \frac{\exp \circ \phi(a)}{\int (\exp \circ \phi) \mathrm{d}P}, \quad a \in A.$$

$$\log \int (\exp \circ \phi) \mathrm{d}P = \sup_{Q \ll P} \left\{ \int \phi \mathrm{d}Q - \mathrm{KL}(Q, P) \right\}, \quad \frac{\mathrm{d}G}{\mathrm{d}P} = \frac{\exp \circ \phi}{\int (\exp \circ \phi) \mathrm{d}P}.$$

Proof: let $Q \ll P$.

$$\log \int (\exp \circ \phi) \mathrm{d}P = \sup_{Q \ll P} \left\{ \int \phi \mathrm{d}Q - \mathrm{KL}(Q, P) \right\}, \quad \frac{\mathrm{d}G}{\mathrm{d}P} = \frac{\exp \circ \phi}{\int (\exp \circ \phi) \mathrm{d}P}.$$

Proof: let $Q \ll P$.

$$-\mathrm{KL}(Q, G) = -\int \log \left(\frac{\mathrm{d}Q}{\mathrm{d}P} \frac{\mathrm{d}P}{\mathrm{d}G} \right) \mathrm{d}Q$$

$$\log \int (\exp\circ\phi)\mathrm{d}P = \sup_{Q\ll P}\left\{\int \phi \mathrm{d}Q - \mathrm{KL}(Q,P)\right\}, \quad \frac{\mathrm{d}G}{\mathrm{d}P} = \frac{\exp\circ\phi}{\int(\exp\circ\phi)\mathrm{d}P}.$$

Proof: let $Q \ll P$.

$$\begin{aligned} -\mathrm{KL}(Q,G) &= -\int \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\frac{\mathrm{d}P}{\mathrm{d}G}\right)\mathrm{d}Q \\ &= -\int \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)\mathrm{d}Q + \int \log\left(\frac{\mathrm{d}G}{\mathrm{d}P}\right)\mathrm{d}Q \end{aligned}$$

$$\log \int (\exp \circ \phi) \mathrm{d}P = \sup_{Q \ll P} \left\{ \int \phi \mathrm{d}Q - \mathrm{KL}(Q, P) \right\}, \quad \frac{\mathrm{d}G}{\mathrm{d}P} = \frac{\exp \circ \phi}{\int (\exp \circ \phi) \mathrm{d}P}.$$

Proof: let $Q \ll P$.

$$\begin{aligned}
-\mathrm{KL}(Q, G) &= -\int \log \left(\frac{\mathrm{d}Q}{\mathrm{d}P} \frac{\mathrm{d}P}{\mathrm{d}G} \right) \mathrm{d}Q \\
&= -\int \log \left(\frac{\mathrm{d}Q}{\mathrm{d}P} \right) \mathrm{d}Q + \int \log \left(\frac{\mathrm{d}G}{\mathrm{d}P} \right) \mathrm{d}Q \\
&= -\mathrm{KL}(Q, P) + \int \phi \mathrm{d}Q - \log \int (\exp \circ \phi) \, \mathrm{d}P.
\end{aligned}$$

$$\log \int (\exp \circ \phi) \mathrm{d}P = \sup_{Q \ll P} \left\{ \int \phi \mathrm{d}Q - \mathrm{KL}(Q, P) \right\}, \quad \frac{\mathrm{d}G}{\mathrm{d}P} = \frac{\exp \circ \phi}{\int (\exp \circ \phi) \mathrm{d}P}.$$

Proof: let $Q \ll P$.

$$\begin{aligned}
-\mathrm{KL}(Q, G) &= -\int \log \left(\frac{\mathrm{d}Q}{\mathrm{d}P} \frac{\mathrm{d}P}{\mathrm{d}G} \right) \mathrm{d}Q \\
&= -\int \log \left(\frac{\mathrm{d}Q}{\mathrm{d}P} \right) \mathrm{d}Q + \int \log \left(\frac{\mathrm{d}G}{\mathrm{d}P} \right) \mathrm{d}Q \\
&= -\mathrm{KL}(Q, P) + \int \phi \mathrm{d}Q - \log \int (\exp \circ \phi) \, \mathrm{d}P.
\end{aligned}$$

$\mathrm{KL}(\cdot, \cdot)$ is non-negative, $Q \mapsto -\mathrm{KL}(Q, G)$ reaches its max. in $Q = G$:

$$\log \int (\exp \circ \phi) \mathrm{d}P = \sup_{Q \ll P} \left\{ \int \phi \mathrm{d}Q - \mathrm{KL}(Q, P) \right\}, \quad \frac{\mathrm{d}G}{\mathrm{d}P} = \frac{\exp \circ \phi}{\int (\exp \circ \phi) \mathrm{d}P}.$$

Proof: let $Q \ll P$.

$$\begin{aligned}
-\mathrm{KL}(Q, G) &= -\int \log \left(\frac{\mathrm{d}Q}{\mathrm{d}P} \frac{\mathrm{d}P}{\mathrm{d}G} \right) \mathrm{d}Q \\
&= -\int \log \left(\frac{\mathrm{d}Q}{\mathrm{d}P} \right) \mathrm{d}Q + \int \log \left(\frac{\mathrm{d}G}{\mathrm{d}P} \right) \mathrm{d}Q \\
&= -\mathrm{KL}(Q, P) + \int \phi \mathrm{d}Q - \log \int (\exp \circ \phi) \, \mathrm{d}P.
\end{aligned}$$

$\mathrm{KL}(\cdot, \cdot)$ is non-negative, $Q \mapsto -\mathrm{KL}(Q, G)$ reaches its max. in $Q = G$:

$$0 = \sup_{Q \ll P} \left\{ \int \phi \mathrm{d}Q - \mathrm{KL}(Q, P) \right\} - \log \int (\exp \circ \phi) \, \mathrm{d}P.$$

$$\log \int (\exp \circ \phi) \mathrm{d}P = \sup_{Q \ll P} \left\{ \int \phi \mathrm{d}Q - \mathrm{KL}(Q, P) \right\}, \quad \frac{\mathrm{d}G}{\mathrm{d}P} = \frac{\exp \circ \phi}{\int (\exp \circ \phi) \mathrm{d}P}.$$

Proof: let $Q \ll P$.

$$\begin{aligned} -\mathrm{KL}(Q, G) &= -\int \log \left(\frac{\mathrm{d}Q}{\mathrm{d}P} \frac{\mathrm{d}P}{\mathrm{d}G} \right) \mathrm{d}Q \\ &= -\int \log \left(\frac{\mathrm{d}Q}{\mathrm{d}P} \right) \mathrm{d}Q + \int \log \left(\frac{\mathrm{d}G}{\mathrm{d}P} \right) \mathrm{d}Q \\ &= -\mathrm{KL}(Q, P) + \int \phi \mathrm{d}Q - \log \int (\exp \circ \phi) \, \mathrm{d}P. \end{aligned}$$

$\mathrm{KL}(\cdot, \cdot)$ is non-negative, $Q \mapsto -\mathrm{KL}(Q, G)$ reaches its max. in $Q = G$:

$$0 = \sup_{Q \ll P} \left\{ \int \phi \mathrm{d}Q - \mathrm{KL}(Q, P) \right\} - \log \int (\exp \circ \phi) \, \mathrm{d}P.$$

Let $\lambda > 0$ and take $\phi = -\lambda R_{\mathrm{in}}$,

$$Q_\lambda \propto \exp(-\lambda R_{\mathrm{in}}) \, P = \operatorname*{arg\,inf}_{Q \ll P} \left\{ R_{\mathrm{in}}(Q) + \frac{\mathrm{KL}(Q, P)}{\lambda} \right\}.$$