On generalisation and learning

Benjamin Guedj

NYUAD

4th Frebruary 2026, Abu Dhabi

Professor of Machine Learning and Foundational Artificial Intelligence, UCL
Research Director, Inria
Scientific leader of the Inria-UCL joint research centre
`https://bguedj.github.io`

Foundations of AI

Practice is largely outpacing theory!

- Ubiquitous deployment of AI calls for **predictability**, **robustness**, **accountability**.
- Theory is a way to **separate signal from hype**, and to **design** better systems, rooted in foundational principles.

Human and artificial intelligence operate under uncertainty; **statistics** is the powerhouse of the quantification of uncertainty. Human intelligence is fundamentally statistical in nature.

A definition of learning: compressing past data into **generalisation-ready** systems.

On the mismatch between humans and machines

- Humans: sample-efficient, causal hints, priors from world knowledge.
- Machines: data-hungry, fragile under distribution shifts, compute-intensive.
- A more human-like or **frugal AI**: principled priors, compression, selective sensing, uncertainty-aware decisions.

Ultimately, we aim for similar or better performance with **less data**, **less compute**, and **predictable** and reproducible behaviour.

Mathematical foundations of intelligence

Research at the crossroads of statistics, probability theory, machine learning, optimisation. *Mathematical foundations of artificial intelligence* is a pretty good tagline.

Keywords: statistical learning theory, PAC-Bayes, generalisation bounds, concentration inequalities, computational statistics, theoretical analysis of deep learning and in particular generative models, information theory

Generalisation theory is all about understanding how to design learning algorithm that learn well beyond training data.

Think about the paradigms in machine vs. human intelligence, and how we think about data / experience.

Outline

Generalisation in machine learning

Generalisation-driven deep learning

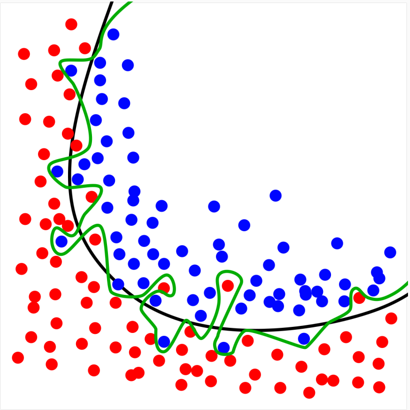
Wasserstein-based deviation bounds

Comparators in generalisation bounds

Information theory and PAC-Bayes united

Generalisation in machine learning

Learning is to be able to generalise



[Source: Wikipedia]

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

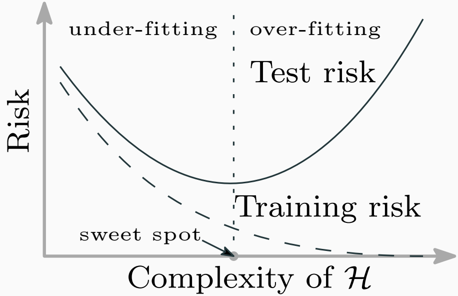
Memorising the already seen data is usually bad (overfitting)

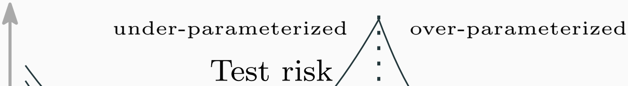
Generalisation is the ability to ’perform’ well on unseen data.

The deep learning era puts generalisation on the spot

Neural networks architectures trained on massive datasets achieve zero training error which strongly suggests to statisticians like me they may overfit.

However they often achieve remarkably low errors on test sets – hence the interest in generalisation bounds for deep networks.





Why generalisation matters in machine learning

Let $(X_i, Y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ be a sample drawn from some distribution $\mathcal{D}^{\otimes n}$, and let $\ell\colon \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ be a loss function. For any hypothesis $h\colon \mathcal{X} \to \mathcal{Y}$,

$$\widehat{L}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i), \quad L(h) = \mathbb{E}\ell(h(X), Y).$$

- How can we certify that a hypothesis with good performance on training data has similarly good performance on new, unseen data?
- When does a low training loss imply a low population loss?

Typical approach: bound the *generalisation gap*. For a hypothesis h, population loss L and training loss $\widehat{L}$, let

$$\Gamma(h) := L(h) - \widehat{L}(h)$$

denote the generalisation gap. We want

$$L(h) = \widehat{L}(h) + L(h) - \widehat{L}(h) = \widehat{L}(h) + \Gamma(h) \leq \widehat{L}(h) + \mathrm{Bound},$$

This motivates *generalisation bounds*: $\Gamma(h) \leq \mathrm{Bound}$, with several flavours

- hypothesis-dependent vs. hypothesis-free
- (data generating) distribution-dependent vs. distribution-free
- in expectation
- with (arbitrarily) high probability

The PAC (Probably Approximately Correct) framework

📕 Valiant, A theory of the learnable, Communications of the ACM, 1984

$\mathbb{P}$[large error] $\leq \delta$. The 'confidence parameter' δ can be thought of as the probability of being misled by the training set.

Hence high confidence: $\mathbb{P}$[approximately correct] $\geq 1 - \delta$.

With high probability, the generalisation gap of an hypothesis h is at most something we can control and even compute. For any $\delta > 0$,

$$\mathbb{P}\left[L(h) \leq \widehat{L}(h) + \mathcal{B}(n, \delta)\right] \geq 1 - \delta.$$

Think of $\mathcal{B}(n, \delta)$ as $\mathrm{Complexity} \times \frac{\log 1/\delta}{\sqrt{n}}$. PAC bounds are high confidence statements on the tail of the distribution of population losses (think of a statistical test at level $1 - \delta$).

PAC-Bayes

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

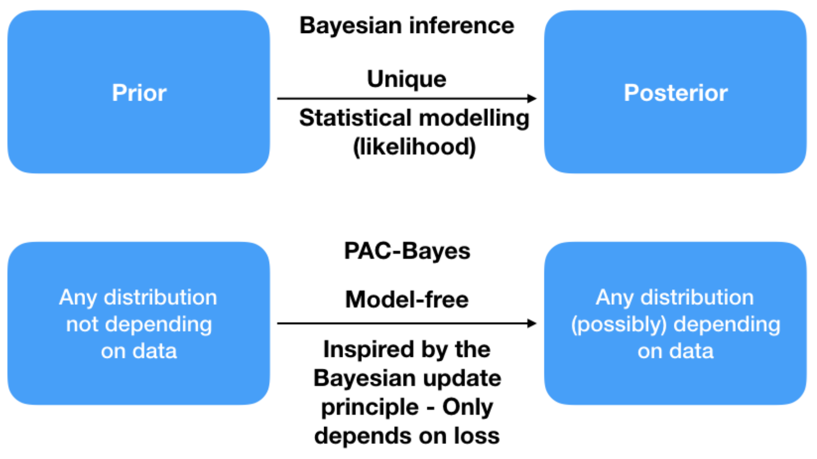
Notation

Let Q_n denote a posterior distribution that produces hypotheses,

$$\widehat{\mathcal{L}}(Q_n) = \mathbb{E}_{h \sim Q_n} \widehat{L}(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{h \sim Q_n} \ell(h(X_i), Y_i),$$

$$\mathcal{L}(Q_n) = \mathbb{E}_{h \sim Q_n} L(h) = \mathbb{E}_{h \sim Q_n} \mathbb{E} \ell(h(X), Y).$$

We compare Q_n to a prior Q_0, typically through the KL divergence $\mathrm{KL}(Q_n || Q_0) = \mathbb{E}_{h \sim Q_n} \log \frac{Q_n(h)}{Q_0(h)}$.


Prior
Bayesian inference
Unique
Statistical modelling (likelihood)
Posterior
Any distribution not depending on data
PAC-Bayes
Model-free
Inspired by the Bayesian update principle - Only depends on loss
Any distribution (possibly) depending on data

What makes PAC-Bayes a post-Bayes approach?

- 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 PAC-Bayesian bound

- Shawe-Taylor and Williamson, A PAC analysis of a Bayes estimator, COLT, 1997
- McAllester, Some PAC-Bayesian theorems, COLT, 1998
- McAllester, PAC-Bayesian model averaging, COLT, 1999

Prototypical bound

For any prior Q_0, any $\delta \in (0, 1]$, we have

$$\mathbb{P}\left(\forall\, Q_n:\ \mathcal{L}(Q_n) \leq \widehat{\mathcal{L}}(Q_n) + \sqrt{\frac{\mathrm{KL}(Q_n \| Q_0) + \log(2\sqrt{n}/\delta)}{2n}}\right) \geq 1 - \delta.$$

What is this useful for?

From

$$\mathbb{P}\left[\forall\, Q_n : \mathcal{L}(Q_n) \leq \widehat{\mathcal{L}}(Q_n) + \mathcal{B}(n, \delta, Q_n)\right] \geq 1 - \delta,$$

- We can compute the numerical value of the bound $\mathcal{B}(n, \delta, Q_n)$,
- We can train new algorithms and derive new hypotheses, with

$$Q^\star \in \underset{Q_n \ll Q_0}{\arg\inf} \left\{ \widehat{\mathcal{L}}(Q_n) + \mathcal{B}(n, \delta, Q_n) \right\}$$

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

Variational definition of the KL-divergence

📕 Csiszár., I-divergence geometry of probability distributions and minimization problems, Annals of Probability, 1975

📕 Donsker and Varadhan, Asymptotic evaluation of certain Markov process expectations for large time,

Communications on Pure and Applied Mathematics, 1975

📕 Catoni, Statistical Learning Theory and Stochastic Optimization, Springer, 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 $Q \ll P$.

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

$\mathrm{KL}(\cdot \| \cdot)$ 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 $\phi = -\lambda \widehat{\mathcal{L}}$,

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

"Why should I care about generalisation?"

Generalisation bounds are both a safety check (theoretical and possibly numerical guarantee on the performance of hypotheses on unseen data) and an original training objective.

Formalisms for generalisation

- Concentration inequalities
- Rademacher complexities
- VC-dimension
- Information-theoretic
- PAC-Bayes bounds

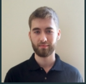
Generalisation-driven deep learning

- Letarte, Germain, Guedj and Laviolette, Dichotomize and generalize: PAC-Bayesian binary activated deep neural networks, NeurIPS, 2019
- Biggs and Guedj, Differentiable PAC-Bayes Objectives with Partially Aggregated Neural Networks, Entropy, 2021
- Biggs and Guedj, On Margins and Derandomisation in PAC-Bayes, AISTATS, 2022
- Cherief-Abdellatif, Shi, Doucet and Guedj, On PAC-Bayesian reconstruction guarantees for VAEs, AISTATS, 2022
- Biggs and Guedj, Non-Vacuous Generalisation Bounds for Shallow Neural Networks, ICML, 2022

Common trait of these works: for specific architectures of deep neural networks, we obtain PAC-Bayes generalisation bounds which are

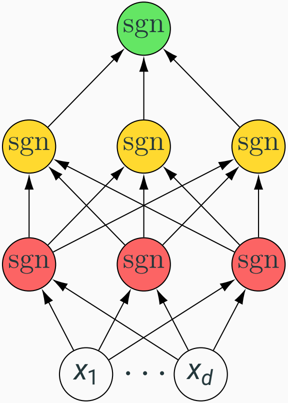
- used as a training objective – delivering networks which achieve the best generalisation performance
- evaluated numerically: all are non-vacuous

Binary Activated 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.
- $\theta = \mathrm{vec}\left(\{\mathbf{W}_k\}_{k=1}^{L}\right) \in \mathbb{R}^D$

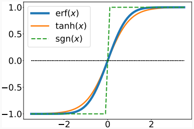
Prediction

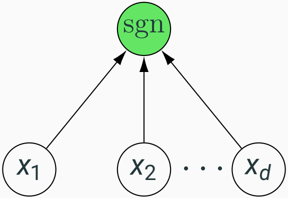
$$f_\theta(\mathbf{x}) = \mathrm{sgn}\left(\mathbf{w}_L \mathrm{sgn}\left(\mathbf{W}_{L-1} \mathrm{sgn}\left(\ldots \mathrm{sgn}\left(\mathbf{W}_1 \mathbf{x}\right)\right)\right)\right),$$

Building block: one layer (aka linear predictor)

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

- Linear classifiers $\mathcal{F}_d \overset{\text{def}}{=} \{f_{\mathbf{v}} | \mathbf{v} \in \mathbb{R}^d\}$
- Predictor
  $F_{\mathbf{w}}(\mathbf{x}) \overset{\text{def}}{=} \mathbf{E}_{\mathbf{v} \sim Q_{\mathbf{w}}} f_{\mathbf{v}}(\mathbf{x}) \;=\; \mathrm{erf}\left(\frac{\mathbf{w} \cdot \mathbf{x}}{\sqrt{d}\|\mathbf{x}\|}\right)$
- 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 $\theta \in \mathbb{R}^D$

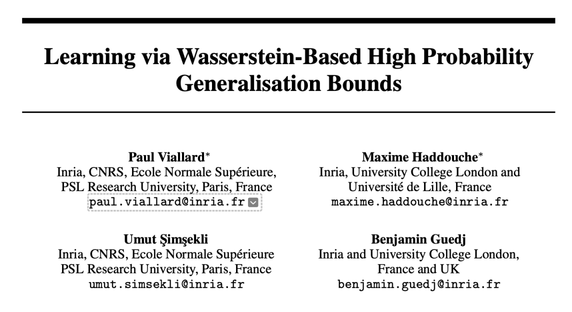
$$\mathcal{L}(F_\theta) \leq \inf_{C>0} \left\{ \frac{1}{1-e^{-C}} \left(1 - \exp\left(-C\widehat{\mathcal{L}}(F_\theta) - \frac{\mathrm{KL}(\theta,\theta_0) + \log \frac{2\sqrt{m}}{\delta}}{m} \right) \right) \right\}.$$

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$_{\text{pre}}$					
– pretrain	linear loss (20 epochs)	50%	-	-	random init
– final	PAC-Bayes bound	50%	-	PAC-Bayes bound	pretrain

Dataset	MLP–tanh		PBGNet$_\ell$		PBGNet			PBGNet$_{\text{pre}}$		
	$\mathcal{L}$	$\widehat{\mathcal{L}}$	$\widehat{\mathcal{L}}$	$\widehat{\mathcal{L}}$	$\mathcal{L}$	$\widehat{\mathcal{L}}$	Bound	$\mathcal{L}$	$\widehat{\mathcal{L}}$	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

Wasserstein-based deviation bounds



Learning via Wasserstein-Based High Probability Generalisation Bounds

Paul Viallard*
Inria, CNRS, Ecole Normale Supérieure,
PSL Research University, Paris, France
paul.viallard@inria.fr

Maxime Haddouche*
Inria, University College London and
Université de Lille, France
maxime.haddouche@inria.fr

Umut Şimşekli
Inria, CNRS, Ecole Normale Supérieure
PSL Research University, Paris, France
umut.simsekli@inria.fr

Benjamin Guedj
Inria and University College London,
France and UK
benjamin.guedj@inria.fr

Viallard, Haddouche, Simsekli and Guedj, *Learning via Wasserstein-based high probability generalisation bounds*, NeurIPS 2023.

Why Wasserstein instead of KL?

- Classical PAC-Bayes bounds use $\mathrm{KL}(\rho\|\pi)$, which can:
  - ignore geometry of $\mathcal{H}$ or $\mathcal{Z}$;
  - break when $\rho \not\ll \pi$;
  - be vacuous with heavy-tailed losses.
- The Wasserstein distance

  $$W(\rho,\pi) = \inf_{\gamma\in\Gamma(\rho,\pi)} \mathbb{E}_{(h,h')\sim\gamma}\left[d(h,h')\right]$$

  encodes geometry and does not require absolute continuity.
- We provide high-probability PAC-Bayes bounds with W, valid under weak moment assumptions and even non-i.i.d. data.

Setup: priors and multiple priors

- Hypothesis space $\mathcal{H}$ with metric d; data $S = (z_1, \ldots, z_m) \sim \mu^m$.
- Prior $\pi \in \mathcal{M}(\mathcal{H})$, posterior $\rho \in \mathcal{M}(\mathcal{H})$.
- Split S into K disjoint subsets $S_1, \ldots, S_K$.
- Each prior $\pi_{i,S}$ is built from data disjoint from S_i (independence for the bound).

$\rightarrow$ Data-dependent priors remain valid via sample splitting.

Theorem 2: High-probability Wasserstein PAC-Bayes bound

Assume ℓ is L-Lipschitz in h and non-negative. For any $\delta \in (0, 1]$, with probability at least $1 - \delta$ over $S \sim \mu^m$, the following holds for the distributions $\pi_{i,S} := \pi_i(S, \cdot)$ and for any $\rho \in \mathcal{M}(\mathcal{H})$:

$$\boxed{\mathbb{E}_{h\sim\rho}\big[R_\mu(h) - \widehat{R}_S(h)\big] \;\leq\; \sum_{i=1}^{K} \frac{2|S_i|L}{m}\, W\big(\rho, \pi_{i,S}\big) \;+\; \sum_{i=1}^{K} \sqrt{\frac{2\,|S_i|\,\ln\big(K/\delta\big)}{m^2}}\;.}$$

From bound to learning objective

Minimising the RHS of Theorem 2 gives:

$$\rho^\star \in \arg \min_{\rho \in \mathcal{M}(\mathcal{H})} \left[\mathbb{E}_{h\sim\rho} \widehat{R}_S(h) + \sum_{i=1}^{K} \frac{2|S_i|L}{m} W\big(\rho, \pi_{i,S}\big) \right].$$

For deterministic predictors ($\rho = \delta_{h_w}$):

$$h_w^\star \in \arg\min_{w} \; \widehat{R}_S(h_w) + \varepsilon \sum_{i=1}^{K} \frac{|S_i|}{m} d\big(h_w, h_{w_i}\big).$$

$\rightarrow$ Wasserstein acts as a geometry-aware regulariser.

Theorem 4: Online Wasserstein PAC-Bayes bound (statement)

Assume the loss $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$ is L-Lipschitz in h, and that priors $\pi_i(S, \cdot)$ satisfy bounded conditional second moments:

$$\forall i, S : \quad \mathbb{E}_{h \sim \pi_i(S,\cdot)}\left[\mathbb{E}_{i-1}\left[\ell(h, z_i)^2\right]\right] \leq 1.$$

Then for any $\delta \in (0, 1]$, with probability at least $1 - \delta$ over $S \sim \mu^m$, for data-dependent priors $\pi_{i,S} = \pi_i(S, \cdot)$ and any posterior sequence $(\rho_i)_{i=1}^m$,

$$\boxed{\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{h \sim \rho_i}\Big[\mathbb{E}[\ell(h, z_i) \mid \mathcal{F}_{i-1}] - \ell(h, z_i)\Big] \leq \frac{2L}{m}\sum_{i=1}^{m} W(\rho_i, \pi_{i,S}) + \sqrt{\frac{2\ln(1/\delta)}{m}}\ .}$$

Theorem 4: interpretation and learning rule

- This is the first online PAC-Bayes bound using Wasserstein regularisation.
- Controls the expected regret of the online learner:

$$\text{Regret} = \frac{1}{m}\sum_{i=1}^{m}\Big(\mathbb{E}_{h\sim\rho_i}[\ell(h,z_i)] - \mathbb{E}_{h\sim\pi_{i,S}}[\ell(h,z_i)]\Big).$$

- The additional term $\frac{2L}{m}\sum_i W(\rho_i, \pi_{i,S})$ penalises geometric deviation from the prior sequence.
- The corresponding online update rule:

$$\rho_i \in \arg\min_{\rho}\ \mathbb{E}_{h\sim\rho}[\ell(h,z_i)] + 2L\, W(\rho, \pi_{i,S}), \qquad i = 1,\ldots,m.$$

- For deterministic learners:

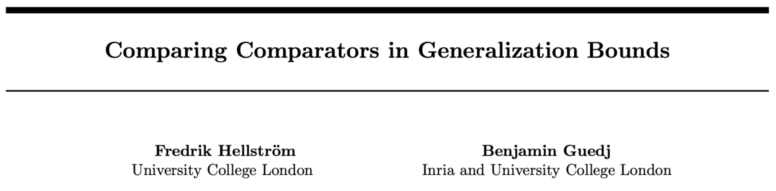
$$h_i \in \arg\min_{h}\ \ell(h,z_i) + d(h,h_{i-1}), \quad d(h,h_{i-1}) \leq 1.$$

$\rightarrow$ Geometry-aware online learning with transport regularisation.

Take-home message

- High-probability Wasserstein PAC-Bayes bounds for batch and online settings.
- Linear W-terms $\Rightarrow$ optimisable objectives and deterministic predictors.
- Especially robust under heavy tails and geometry-sensitive $\mathcal{H}$.

Comparators in generalisation bounds



Comparing Comparators in Generalization Bounds

Fredrik Hellström
University College London

Benjamin Guedj
Inria and University College London

Hellström and Guedj, Comparing comparators in generalization bounds, AISTATS, 2024

The typical approach

- Most generalisation bounds are about bounding the difference $\mathcal{L} - \widehat{\mathcal{L}}$
- Simple, and easy to interpret, but not always tight!
- Can we do better?

Generalising with Comparator Functions

We define the comparator function as $\Delta\colon [0,\infty)^2 \to [0,\infty)$ convex.

A comparator function computes a discrepancy between the training and population loss.

Generic PAC-Bayes Bound with a comparator

Theorem
Assume the loss ℓ is bounded by 1. For any comparator Δ,

$$\mathbb{P}\left[\Delta(\widehat{\mathcal{L}}, \mathcal{L}) \leq \frac{\mathrm{KL}(Q_n \| Q_0) + \log \frac{\Upsilon_\Delta(n)}{\delta}}{n}\right] \geq 1 - \delta,$$

where

$$\Upsilon_\Delta(n) = \sup_{r \in [0,1]} \sum_{k=0}^{n} \binom{n}{k} r^k (1-r)^{n-k} e^{n\Delta(k/n, r)}.$$

Bégin et al., PAC-Bayesian bounds based on the Rényi divergence, AISTATS, 2016

Many known bounds arise as instances of the bound from Bégin et al. (2016). Examples:

- Difference: $\Delta(p, q) = p - q$, we obtain McAllester's bound

$$\mathbb{P}\left(\mathcal{L}(Q_n) \leq \widehat{\mathcal{L}}(Q_n) + \sqrt{\frac{\mathrm{KL}(Q_n \| Q_0) + \log(2\sqrt{n}/\delta)}{2n}}\right) \geq 1 - \delta.$$

- Catoni's family, for any $\gamma \in \mathbb{R}$

$$\Delta_\gamma(p, q) = \gamma q - \log(1 - p + pe^\gamma),$$

  and we get the bound

$$\mathbb{P}\left(\Delta_\gamma(\widehat{\mathcal{L}}(Q_n), \mathcal{L}(Q_n)) \leq \frac{\mathrm{KL}(Q_n \| Q_0) + \log \frac{1}{\delta}}{n}\right) \geq 1 - \delta,$$

- Binary KL divergence

$$\Delta(p, q) = \mathsf{kl}(q, p) = \mathrm{KL}(\mathrm{Bern}(q) \,\|\, \mathrm{Bern}(p)) \\ = q \log \frac{q}{p} + (1 - q) \log \frac{1 - q}{1 - p},$$

and we get the Maurer-Langford-Seeger bound

$$\mathbb{P}\left(\mathsf{kl}(\widehat{\mathcal{L}}(Q_n), \mathcal{L}(Q_n)) \leq \frac{\mathrm{KL}(Q_n \| Q_0) + \log \frac{2\sqrt{n}}{\delta}}{n} \right) \geq 1 - \delta.$$

So which comparator gives the best bound?

When the loss is bounded, the kl is the optimal comparator (up to a log term), as established by Foong et al. (2021).

Foong et al., How Tight Can PAC-Bayes be in the Small Data Regime?, NeurIPS, 2021

In this work we prove a bound which is valid even when the loss is unbounded, and we establish the optimal comparator is the Cramér function (the convex conjugate of the cumulant generating function).

Generalisation bounds and the optimal comparator

We control the gap between empirical and true risk:

$$\widehat{\mathcal{L}}(Q_n) \quad \text{vs.} \quad \mathcal{L}(Q_n)$$

General PAC-Bayesian bounds take the form:

$$\Delta\left(\widehat{\mathcal{L}}(Q_n), \mathcal{L}(Q_n)\right) \lesssim \frac{\mathrm{KL}(Q_n \| Q_0) + \text{complexity}}{n}$$

Key result: among all valid comparator functions Δ,

$$\Delta^\star(q, p) = \Psi_p^*(q)$$

where $\Psi_p(t) = \log \mathbb{E}_{x \sim p}[e^{tx}]$ is the cumulant generating function.

The convex conjugate of a function f is given by

$$f^*(y) = \sup_x \left\{ \langle x, y \rangle - f(x) \right\}.$$

Where does the Cramér function come from?

For independent random variables, rare deviations of empirical averages satisfy

$$\mathbb{P}(\bar{X} \approx q) \asymp \exp\big(-n\,\Psi_p^*(q)\big)$$

(decays exponentially in *n* with rate given by the Cramér function)

📘 Cramér, On a new limit theorem of the theory of probability, Uspekhi Mathematicheskikh Nauk, 1944

📘 Boucheron et al., Concentration inequalities, A nonasymptotic theory of independence, Oxford University Press, 2013

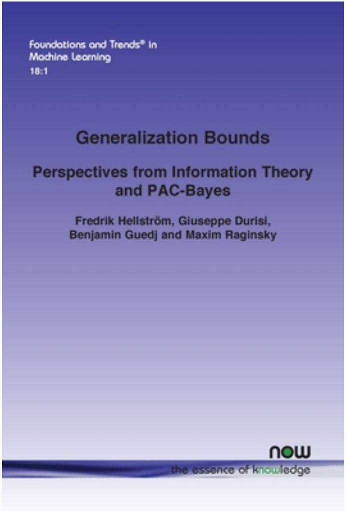
The Cramér function governs concentration of measure and large deviations. Hence

- Generalisation is controlling rare deviations from expectation
- PAC-Bayes bounds inherit the same rate function
- The tight comparator is naturally Ψ^*

Generalisation theory is a large deviation phenomenon.

Information theory and PAC-Bayes united

Everything you’ve ever wanted to know about generalisation



Foundations and Trends® in
Machine Learning
18:1

Generalization Bounds

Perspectives from Information Theory
and PAC-Bayes

Fredrik Hellström, Giuseppe Durisi,
Benjamin Guedj and Maxim Raginsky

now
the essence of knowledge

Hellström, Durisi, Guedj, and Raginsky, Generalization Bounds: Perspectives from Information Theory and PAC-Bayes, Foundations and Trends in Machine Learning, 2025

What the book is about

- Offers a unified view of generalisation through two complementary theories:
    - PAC-Bayes bounds: relate predictors to priors and posteriors;
    - Information-theoretic bounds: relate data to algorithms.
- Both rely on the same three-step reasoning:
    1. control exponential moments of the loss;
    2. perform a change of measure;
    3. derive a concentration inequality.
- The book presents this pattern in a modular way, with examples from algorithmic stability and deep learning.

$\rightarrow$ One common foundation for modern generalisation theory.

Bridging two ways of reasoning

- PAC-Bayes view: compares the learner's average performance under a posterior and a prior.
- Information-theoretic view: quantifies how much the algorithm reveals about its training data.
- These perspectives are mathematically equivalent: a PAC-Bayes bound can be written as an information-theoretic bound with a matched reference distribution.
- PAC-Bayes is *constructive* — it suggests training objectives. Information theory is *diagnostic* — it measures complexity and stability.

$\rightarrow$ Two complementary lenses on generalisation.

Wrapping up

- **AI is everywhere** — but often misunderstood and overhyped.
- **Theory is central** to making AI reliable, interpretable, and trustworthy.
- From **uncertainty quantification** to **generalisation** and **frugal AI**, foundational principles guide how we design, analyse, and deploy intelligent systems.

Thank you!