

# Artificial intelligence from a statistical perspective

---

Benjamin Guedj

D3A Conference

27th August 2025, Nyborg, Denmark



Now: Professor of Machine Learning and Foundational Artificial Intelligence, UCL, UK  
& Research Director, Inria, France

*A while ago:* PhD in mathematical statistics, Sorbonne Université, France

*Even longer ago:* Research assistant, **DTU, Denmark**

<https://bguedj.github.io>

A person's hands are visible at the top, holding two white cards. The card on the left has the word "statistics" written on it, and the card on the right has the words "artificial intelligence" written on it. The background is slightly blurred, showing what appears to be a desk with a fan and a cup.

**statistics**

**artificial  
intelligence**

**Corporate needs you to find the differences  
between this picture and this picture.**



# Statistics vs. Machine Learning vs. Artificial Intelligence

## Statistics

- Inference under **uncertainty**
- **Estimation** & testing
- **Interpretability**

## Machine Learning

- Algorithms that **learn from data**
- **Prediction** & **generalisation**
- **Optimisation** focus

## Artificial Intelligence

- Broader goal: **intelligent behaviour**
- **Perception, reasoning, action, agents**
- May or may not rely on ML and statistics

## Common misuse

In everyday conversation, these terms are often used **interchangeably**; in reality, they are distinct — but overlapping — perspectives.

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, brittle 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.

Research at the crossroads of statistics, probability theory, machine learning, optimisation.

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

Generalisation theory is all about understanding how to design learning algorithm that learn well beyond training data. In machine learning, a model is only as good as its **out-of-sample** behaviour.

Generalisation in machine learning

Case study: PAC-Bayes-powered deep learning

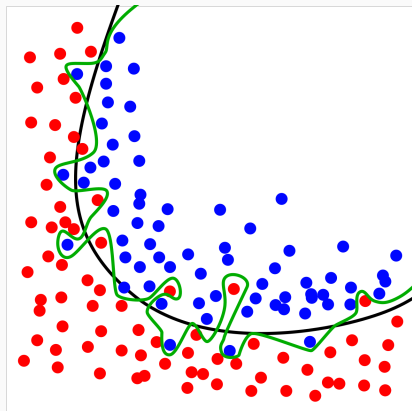
Comparators in generalisation bounds

# Generalisation in machine learning

---



# Learning is to be able to generalise



[Source: Wikipedia]

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

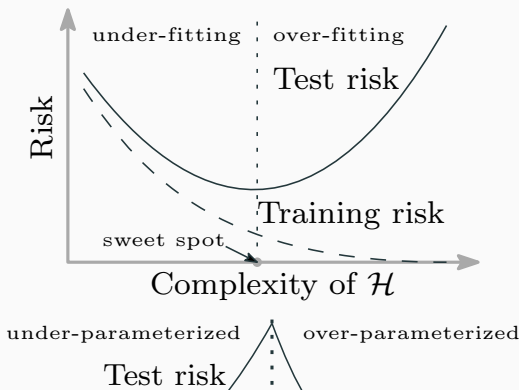
Memorising the already seen data is usually bad (overfitting)

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.



# Statistical Learning Theory is about high confidence

Fix an algorithm, hypothesis class, sample size. Generate random samples to study the distribution of test errors.

- Focusing on the mean of the error distribution?
  - ▷ can be (highly) misleading
- **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%.The same principles are at play in conformal prediction.

# Why generalisation matters in machine learning

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

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

- 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  $\hat{L}$ , let

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

denote the generalisation gap. We want

$$L(h) = \hat{L}(h) + L(h) - \hat{L}(h) = \hat{L}(h) + \Gamma(h) \leq \hat{L}(h) + \text{Bound},$$

This motivates *generalisation bounds*:  $\Gamma(h) \leq \text{Bound}$ , 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

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

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

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

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


Think of  $\mathcal{B}(n, \delta)$  as  $\text{Complexity} \times \frac{\log 1/\delta}{\sqrt{m}}$ . 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 is about PAC generalisation bounds for *distributions over hypotheses*. Let  $Q_n$  denote a posterior distribution that produces hypotheses,

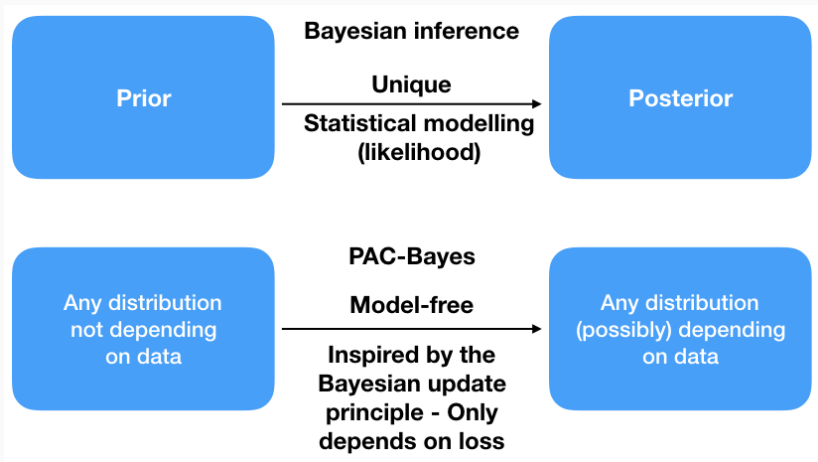
$$\hat{\mathcal{L}}(Q_n) = \mathbb{E}_{h \sim Q_n} \hat{L}(h) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{h \sim Q_n} \ell(h(X_i), Y_i),$$

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

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

 Alquier and Guedj, Simpler PAC-Bayesian bounds for hostile data, Machine Learning, 2018

 Viallard, Haddouche, Simsekli and Guedj, Learning via Wasserstein-Based High Probability Generalisation Bounds, NeurIPS, 2023





# 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

<https://postbayes.github.io/seminar/>

# 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

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

# What is this useful for?

From

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

- 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

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

(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  $\phi : A \rightarrow \mathbb{R}$  such that  $\int (\exp \circ \phi) dP < \infty$ ,

$$\log \int (\exp \circ \phi) dP = \sup_{Q \ll P} \left\{ \int \phi dQ - \text{KL}(Q \| P) \right\}.$$

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

$$\frac{dG}{dP}(a) = \frac{\exp \circ \phi(a)}{\int (\exp \circ \phi) dP}, \quad a \in A.$$

$$\log \int (\exp \circ \phi) dP = \sup_{Q \ll P} \left\{ \int \phi dQ - \text{KL}(Q \| P) \right\}, \quad \frac{dG}{dP} = \frac{\exp \circ \phi}{\int (\exp \circ \phi) dP}.$$

Proof: let  $Q \ll P$ .

$$\begin{aligned} -\text{KL}(Q \| G) &= - \int \log \left( \frac{dQ}{dP} \frac{dP}{dG} \right) dQ \\ &= - \int \log \left( \frac{dQ}{dP} \right) dQ + \int \log \left( \frac{dG}{dP} \right) dQ \\ &= -\text{KL}(Q \| P) + \int \phi dQ - \log \int (\exp \circ \phi) dP. \end{aligned}$$

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

$$0 = \sup_{Q \ll P} \left\{ \int \phi dQ - \text{KL}(Q \| P) \right\} - \log \int (\exp \circ \phi) dP.$$

Let  $\lambda > 0$  and take  $\phi = -\lambda \hat{\mathcal{L}}$ ,

$$Q_\lambda \propto \exp \left( -\lambda \hat{\mathcal{L}} \right) P = \arg \inf_{Q \ll P} \left\{ \hat{\mathcal{L}}(Q) + \frac{\text{KL}(Q \| P)}{\lambda} \right\}.$$

# "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 quantities
- PAC-Bayes bounds (focus of today)

## **Case study: PAC-Bayes-powered 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
- non-vacuous when evaluated numerically



# Binary Activated Networks (NeurIPS 2019)

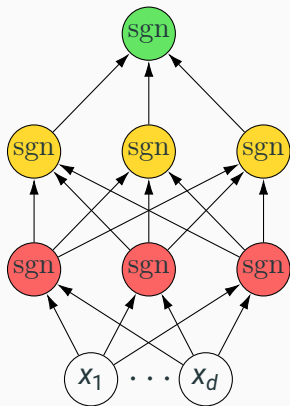


$\mathbf{x} \in \mathbb{R}^{d_0}, y \in \{-1, 1\}$ . Architecture:

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

Parameters:

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



**Prediction**

$$f_{\theta}(\mathbf{x}) = \text{sgn}(\mathbf{w}_L \text{sgn}(\mathbf{w}_{L-1} \text{sgn}(\dots \text{sgn}(\mathbf{w}_1 \mathbf{x})))) ,$$

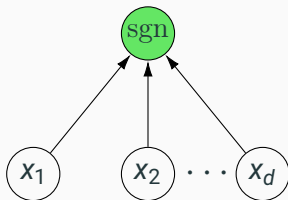
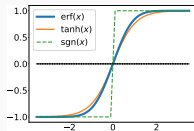
# Building block: one layer (aka linear predictor)

Model  $f_{\mathbf{w}}(\mathbf{x}) \stackrel{\text{def}}{=} \text{sgn}(\mathbf{w} \cdot \mathbf{x})$ , 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

$$F_{\mathbf{w}}(\mathbf{x}) \stackrel{\text{def}}{=} \mathbf{E}_{\mathbf{v} \sim Q_{\mathbf{w}}} f_{\mathbf{v}}(\mathbf{x}) = \text{erf}\left(\frac{\mathbf{w} \cdot \mathbf{x}}{\sqrt{d} \|\mathbf{x}\|}\right)$$

- Sampling + closed form of the KL + a few other tricks + extension to an arbitrary number of layers



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

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

# 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 <sub>ℓ</sub>	linear loss, L2 regularized	80%	20%	valid linear loss	random init
<b>PBGNet</b>	<b>PAC-Bayes bound</b>	<b>100 %</b>	<b>-</b>	<b>PAC-Bayes bound</b>	<b>random init</b>
PBGNet <sub>pre</sub>					
– pretrain	linear loss (20 epochs)	50%	-	-	random init
– final	PAC-Bayes bound	50%	-	<b>PAC-Bayes bound</b>	pretrain

Dataset	MLP-tanh		PBGNet <sub>ℓ</sub>		PBGNet			PBGNet <sub>pre</sub>		
	$\mathcal{L}$	$\hat{\mathcal{L}}$	$\hat{\mathcal{L}}$	$\hat{\mathcal{L}}$	$\mathcal{L}$	$\hat{\mathcal{L}}$	Bound	$\mathcal{L}$	$\hat{\mathcal{L}}$	Bound
ads	0.021	0.037	0.018	<b>0.032</b>	0.024	0.038	<b>0.283</b>	0.034	0.033	<b>0.058</b>
adult	0.128	0.149	0.136	<b>0.148</b>	0.158	0.154	<b>0.227</b>	0.153	0.151	<b>0.165</b>
mnist17	0.003	<b>0.004</b>	0.008	0.005	0.007	0.009	<b>0.067</b>	0.003	0.005	<b>0.009</b>
mnist49	0.002	<b>0.013</b>	0.003	0.018	0.034	0.039	<b>0.153</b>	0.018	0.021	<b>0.030</b>
mnist56	0.002	0.009	0.002	0.009	0.022	0.026	<b>0.103</b>	0.008	<b>0.008</b>	<b>0.017</b>
mnistLH	0.004	<b>0.017</b>	0.005	0.019	0.071	0.073	<b>0.186</b>	0.026	0.026	<b>0.033</b>

## **Comparators in generalisation bounds**

---

---

# Comparing Comparators in Generalization Bounds

---

**Fredrik Hellström**  
University College London

**Benjamin Guedj**  
Inria and University College London



# The typical approach

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

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

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



# Generic PAC-Bayes Bound with a comparator

## Theorem

Assume the loss  $\ell$  is bounded by 1. For any comparator  $\Delta$ ,

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

where

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

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

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

- 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

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

- Binary KL divergence


$$\begin{aligned}\Delta(p, q) &= \text{kl}(q, p) = \text{KL}(\text{Bern}(q) \parallel \text{Bern}(p)) \\ &= q \log \frac{q}{p} + (1 - q) \log \frac{1 - q}{1 - p},\end{aligned}$$

and we get the Maurer-Langford-Seeger bound

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

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 relax the boundedness assumption.

We let

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

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

Let  $X$  be a real-valued random variable. The **cumulant generating function (CGF)** of  $X$  is

$$\psi_X(t) = \log \mathbb{E} \left[ e^{tX} \right].$$

## Theorem — Average Case Generalisation Bound

Let  $\mathcal{P}$  be a set of distributions such that for all  $r \in [0, \infty)$ , there exists  $P_r \in \mathcal{P}$  with mean  $r$ . Let  $\mathcal{C}$  be the set of proper, convex, lower semicontinuous functions  $\mathbb{R}^2 \rightarrow \mathbb{R}$ , and let  $\mathcal{F} \subset \mathcal{C}$  be the set of  $f$  satisfying:

$$\mathbb{E} \left[ e^{f(\hat{\mathcal{L}}(h), \mathcal{L}(h))} \right] \leq \mathbb{E}_{x \sim P_{\mathcal{L}(h)}} \left[ e^{f(\bar{x}, \mathcal{L}(h))} \right].$$

Then for all  $\Delta \in \mathcal{F}$  and all  $Q_n \ll Q_0$ :

$$\Delta(\hat{\mathcal{L}}(Q_n), \mathcal{L}(Q_n)) \leq \frac{\text{KL}(Q_n D^n \| Q_0 D^n) + \log \Upsilon_{\Delta}^{\mathcal{P}}(n)}{n},$$

where

$$\Upsilon_{\Delta}^{\mathcal{P}}(n) = \sup_{r \in [0, \infty)} \mathbb{E}_{x \sim P_r} [\exp(n \Delta(\bar{x}, r))].$$

## How do we make this relevant beyond bounded losses?

Recall that  $\sigma$ -sub-Gaussian random variables are characterized by having a CGF that is dominated by the CGF of some Gaussian distribution with variance  $\sigma^2$ , with similar notions for, e.g., sub-gamma and sub-exponential random variables.

The convex conjugate of a function  $f$  is given by

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

## Definition of Sub- $\mathcal{P}$ Losses

Let  $\mathcal{P}$  be a set of distributions such that, for all  $r \in [0, \infty)$ , there exists  $P_r \in \mathcal{P}$  with first moment  $r$ .

For all  $r \in [0, \infty)$ , let  $\mathcal{T}_r \subset \mathbb{R}$  and  $\mathcal{T} = \{\mathcal{T}_r : r \in [0, \infty)\}$ . We say that the loss is *sub*-( $\mathcal{P}, \mathcal{T}$ ) if, for all  $h$  and  $t \in \mathcal{T}_{\mathcal{L}(h)}$ , we have

$$\mathbb{E} [\exp(t \ell(h(X), Y))] \leq \mathbb{E}_{x \sim P_{\mathcal{L}(h)}} [\exp(tx)].$$

If  $\mathcal{T}_r = \mathbb{R}$  for all  $r \in [0, \infty)$ , we say that the loss is *sub*- $\mathcal{P}$ .

A sub- $\mathcal{P}$  loss never has heavier tails than those of  $\mathcal{P}$ .



## Theorem — Optimal Comparator and Bound

Assume that the loss is sub- $(\mathcal{P}, \mathcal{T})$ . Let  $\Psi_p(t) = \log \mathbb{E}_{x \sim P_p}[e^{tx}]$  be the CGF of the distribution  $P_p$ , and let the **Cramér function** be defined as

$$\Delta_{\mathcal{P}}^{\Psi}(q, p) = \Psi_p^*(q) = \sup_{t \in \mathcal{T}_p} \{tq - \Psi_p(t)\}.$$

Define the bound functional


$$\widehat{B}_n^{\Delta}(\alpha, \beta, \iota) = \sup_{\rho \in \mathcal{L}} \left\{ \rho : \Delta(\alpha, \rho) \leq \frac{\beta + \log \iota(n)}{n} \right\}.$$

Then, for any  $\Delta \in \mathcal{F}$ , we have

$$\begin{aligned} \widehat{\mathcal{L}}(Q_n) &\leq \widehat{B}_n^{\Delta_{\mathcal{P}}^{\Psi}} \left( \widehat{\mathcal{L}}(Q_n), \text{KL}(Q_n D^n \| Q_0 D^n), 1 \right) \\ &\leq \widehat{B}_n^{\Delta} \left( \widehat{\mathcal{L}}(Q_n), \text{KL}(Q_n D^n \| Q_0 D^n), \gamma_{\mathcal{P}}^{\Delta}(n) \right). \end{aligned}$$

In other words, the optimal average generalisation bound is obtained with the Cramér function as comparator.

For independent and identically distributed random variables, the Cramér function characterises the probability of rare events. Thus, the connection to generalisation bounds is somewhat natural.

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

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

# The case of natural exponential families

- If  $\mathcal{P}$  is a NEF, the Cramér function is a KL

$$\Delta_{\mathcal{P}}^{\Psi}(q, p) = \Psi_p^*(q) = \text{KL}(P_q \| P_p).$$

- For the case of Gaussian distributions with known variance, the optimal comparator is given by

$$\text{KL} \left( \mathcal{N}(q, \sigma^2) \| \mathcal{N}(p, \sigma^2) \right) = \frac{(q - p)^2}{2\sigma^2}.$$

## Examples of Cramér Functions

- Bounded loss: binary KL  $\text{kl}(q, p)$ ,
- Sub-Gaussian:  $\frac{(q-p)^2}{2\sigma^2}$ ,
- Sub-Poisson:  $p - q + q \log(q/p)$ ,
- Sub-Gamma:  $k(\frac{q}{p} - 1 - \log \frac{q}{p})$ ,
- Sub-Laplacian:

$$\Delta_{\text{Lap}}^{\Psi}(q, p) = \frac{\sqrt{(q-p)^2 + b^2}}{b} - 1 + \log \left( \frac{2 \left( b \sqrt{(q-p)^2 + b^2} - b^2 \right)}{(q-p)^2} \right).$$

## Theorem — Generic PAC-Bayesian Bound for Sub- $\mathcal{P}$ losses

Assume the loss is Sub- $\mathcal{P}$ . Then for any  $\Delta \in \mathcal{F}$ , with probability at least  $1 - \delta$ , the following holds simultaneously for all posteriors  $Q_n \ll Q_0$

$$\Delta \left( \hat{\mathcal{L}}(Q_n), \mathcal{L}(Q_n) \right) \leq \frac{\text{KL}(Q_n \| Q_0) + \log \frac{\gamma_{\Delta}^{\mathcal{P}}(n)}{\delta}}{n}.$$

# Main takeaways

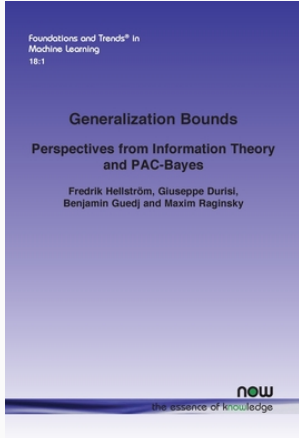
- Comparator choice is crucial in generalisation
- The optimal choice for unbounded losses: Cramér function derived from CGF
- For NEFs, this is equivalent to using the KL divergence

## In a nutshell

The tightest (up to log terms) generalisation bounds with controllable moment-generating functions are obtained with the Cramér function as the comparator function.

- **AI is everywhere** — but often misunderstood.
- **Statistics is central** to making AI reliable, interpretable, and trustworthy.
- From **uncertainty quantification** to **frugal AI**, statistical thinking guides how we design and assess intelligent systems.

# Everything you've ever wanted to know about generalisation



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

Tusind tak!