On generalisation and learning

Benjamin Guedj

Erlangen AI Hub Conference 2025

9th June 2025, London

Professor of Machine Learning and Foundational Artificial Intelligence, UCL & Research Director, Inria
https://bguedj.github.io

Mathematical foundations of intelligence

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.

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, information theory, theoretical analysis of generative models

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, information theory, theoretical analysis of generative models

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

Outline

Generalisation in machine learning

Interlude: PAC-Bayes-powered Deep Learning

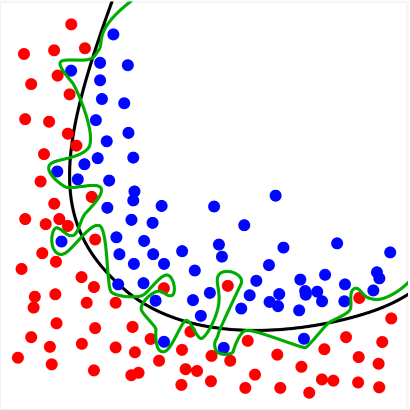
Comparators in generalisation bounds

- Finding the optimal comparator
- Novel (tight) PAC-Bayes bounds

Discussion

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.

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.

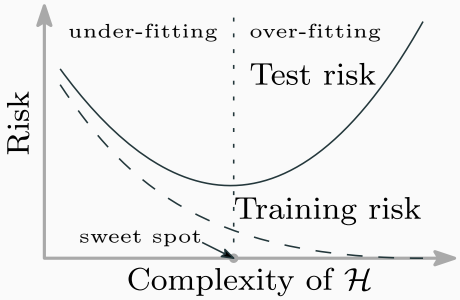
However they often achieve remarkably low errors on test sets – hence the interest in generalisation bounds for deep networks.

Belkin et al., Reconciling modern machine-learning practice and the classical bias-variance trade-off, PNAS, 2019

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.

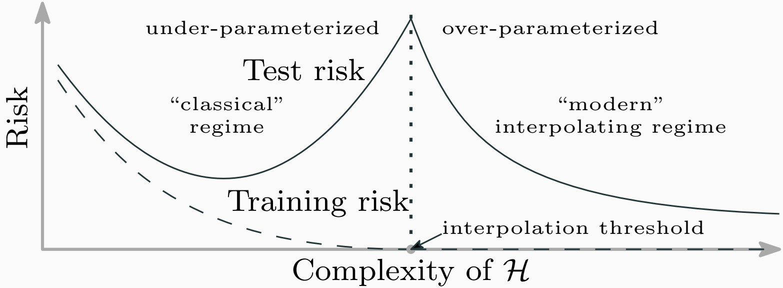
However they often achieve remarkably low errors on test sets – hence the interest in generalisation bounds for deep networks.



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.



Belkin et al., Reconciling modern machine-learning practice and the classical bias-variance trade-off, PNAS, 2019

Statistical Learning Theory is about high confidence

Statistical Learning Theory is about high confidence

Fix an algorithm, hypothesis class, sample size.

Statistical Learning Theory is about high confidence

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

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 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

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%

Why generalisation matters in machine learning

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 $\ell \colon \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ be a loss function. For any hypothesis $h \colon \mathcal{X} \to \mathcal{Y}$,

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

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 $\ell \colon \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ 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?

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 $\ell\colon \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ 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*.

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.

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

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

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

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.

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] $\geq 1 - \delta$.

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 $\delta > 0$,

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

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 $\mathrm{Complexity} \times \frac{\log 1/\delta}{\sqrt{m}}$. 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

PAC-Bayes is about PAC generalisation bounds for *distributions over hypotheses*. Let Q_n denote a posterior distribution that produces hypotheses,

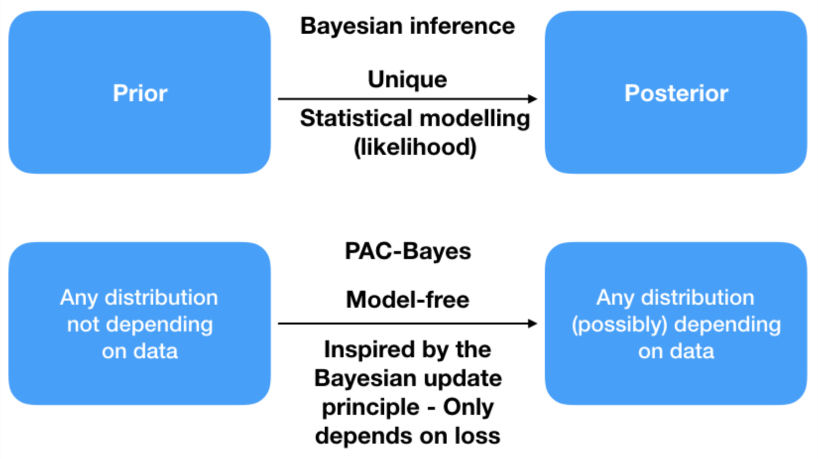
$$\widehat{\mathcal{L}}(Q_n) = \mathbb{E}_{h\sim Q_n}\widehat{L}(h) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{h\sim Q_n}\ell(h(X_i), Y_i),$$

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

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

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


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

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

- 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

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

(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

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 $Q \ll P$.

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

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 $Q = G$:

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

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 $\phi = -\lambda \widehat{\mathcal{L}}$,

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

"Why should I care about generalisation?"

"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)

Interlude: 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
- evaluated numerically: all are non-vacuous

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

Corollary

If $\widehat{\mathcal{L}} \leq \alpha$, $\mathrm{KL}(Q_n \| Q_0) \leq \beta$, and $\Upsilon_\Delta(n) \leq \iota(n)$, we obtain the bound

$$\mathbb{P}\left(\mathcal{L}(Q_n) \leq B_n^\Delta(\alpha, \beta, \iota)\right) \geq 1 - \delta,$$

where

$$B_n^\Delta(\alpha, \beta, \iota) = \sup_{\rho \in [0,1]} \left\{ \rho : \Delta(\alpha, \rho) \leq \frac{\beta + \log \frac{\iota(n)}{\delta}}{n} \right\}.$$

In the previous bound,

- α is the empirical loss $\widehat{\mathcal{L}}(Q_n)$,
- β is the KL divergence $\mathrm{KL}(Q_n \| Q_0)$,
- $\iota(n)$ is a complexity term,
- δ is the confidence level,
- ρ is the variable representing the population loss $\mathcal{L}(Q_n)$.

Given that the comparator between training and population loss is bounded, what is the largest population loss still compatible with the bound?

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

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

  and we get the Maurer-Langford-Seeger bound

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

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

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.

Bounds in expectation

We let

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

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 $\mathbb{R}^2 \to \mathbb{R}$, and let $\mathcal{F} \subset \mathcal{C}$ be the set of f satisfying:

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

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

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

where

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

How do we make this relevant beyond bounded losses?

Recall that σ-sub-Gaussian random variables are characterized by having a CGF that is dominated by the CGF of some Gaussian distribution with variance σ^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

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

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

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

$$\widehat{\mathcal{L}}(Q_n) \leq \widehat{B}_n^{\Delta_{\mathcal{P}}^{\Psi}} \left(\widehat{\mathcal{L}}(Q_n), \mathrm{KL}(Q_n D^n \| Q_0 D^n), 1 \right)$$
$$\leq \widehat{B}_n^{\Delta} \left(\widehat{\mathcal{L}}(Q_n), \mathrm{KL}(Q_n D^n \| Q_0 D^n), \Upsilon_{\mathcal{P}}^{\Delta}(n) \right).$$

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 generalization bounds is somewhat natural.

📚 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 case of natural exponential families

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

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

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

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

Examples of Cramér Functions

- Bounded loss: binary KL $\mathrm{kl}(q, p)$,
- 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:

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

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 $Q_n \ll Q_0$

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

Theorem — Near-Optimality of the Cramér Comparator i

Assume that the loss is sub-$(\mathcal{P}, \mathcal{T})$. Then, for any $\Delta \in \mathcal{F}$, the following holds:

$$B_n^{\Delta_{\mathcal{P}}^{\Psi}}(\widehat{\mathcal{L}}(Q_n), \mathrm{KL}(Q_n \| Q_0), 1) \leq B_n^{\Delta}\left(\widehat{\mathcal{L}}(Q_n), \mathrm{KL}(Q_n \| Q_0), \Upsilon_{\Delta}^{\mathcal{P}}(n)\right).$$

Furthermore, letting $\bar{\Upsilon}(\mathcal{P}) := \Upsilon_{\Delta_{\mathcal{P}}^{\Psi}}^{\mathcal{P}}$, we have:

$$\mathcal{L}(Q_n) \leq B_n^{\Delta_{\mathcal{P}}^{\Psi}}\left(\widehat{\mathcal{L}}(Q_n), \mathrm{KL}(Q_n \| Q_0), \bar{\Upsilon}(\mathcal{P})\right).$$

Finally, for any fixed $t \in \mathcal{T}_p$, define $\Delta_{\mathcal{P}}^{t}(q, p) = tq - \Psi_p(t)$. Then:

$$\mathcal{L}(Q_n) \leq B_n^{\Delta_{\mathcal{P}}^{t}}\left(\widehat{\mathcal{L}}(Q_n), \mathrm{KL}(Q_n \| Q_0), 1\right).$$

Theorem — Near-Optimality of the Cramér Comparator ii

The first inequality shows that the Cramér comparator gives the smallest possible bound up to the normalisation factor.

The second inequality is a valid PAC-Bayesian generalisation bound using $\Delta_{\mathcal{P}}^{\Psi}$.

The third provides a parametric bound for fixed t, useful for optimisation.

Discussion

Main takeaways

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

Main takeaways

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

In a nutshell

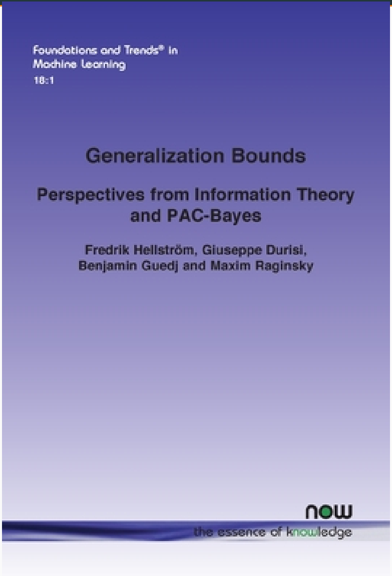
The tightest (up to log terms) generalisation bounds with controllable moment-generating functions are obtained with the Cramér function as the comparator function.

Open Questions

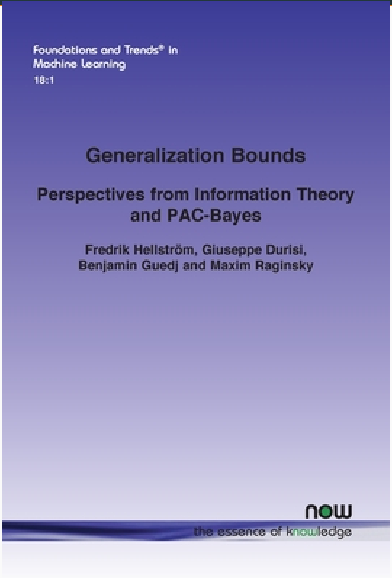
- Can we extend beyond CGF-controlled losses?
- Can we eliminate the log slack?
- Does this strategy apply to heavy-tailed losses?
- Can we derive conditional mutual information bounds?
- Empirical calibration of CGFs in practice

Everything you've ever wanted to know about generalisation

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



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

Thank you!