

Rethinking Generalisation: Beyond KL with Geometry and Comparators

Benjamin Guedj

Selected Lectures on Statistical Machine Learning
25th February 2026, Centre Inria de l'Université Côte d'Azur



<https://bguedj.github.io>

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

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.

This talk: recent advances beyond classical generalisation bounds, replacing KL divergences with Wasserstein distances, using comparators to make bounds tighter, deep connection between information theory and PAC-Bayes.

Generalisation in machine learning

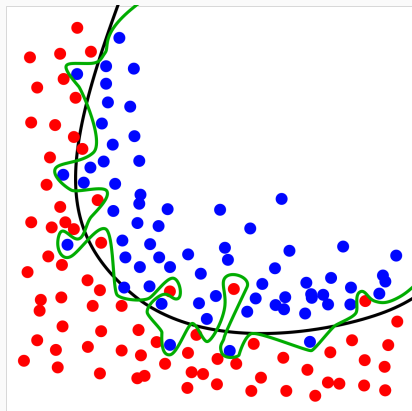
Wasserstein-based deviation bounds

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.

The deep learning era puts generalisation on the spot

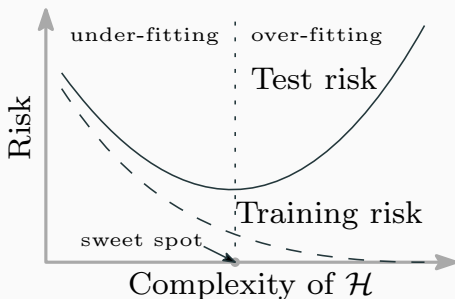
Neural networks architectures trained on massive datasets achieve zero training error which strongly suggests to statisticians like me they may overfit.

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.

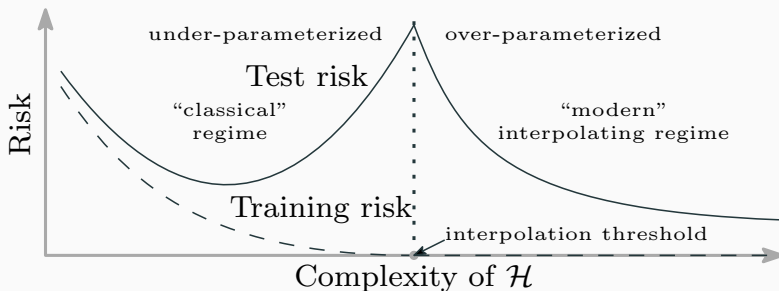
However they often achieve remarkably low errors on test sets – hence the interest in generalisation bounds for deep networks.



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

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 $h: \mathcal{X} \rightarrow \mathcal{Y}$,

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

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*.

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.

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

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

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

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’ 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

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

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

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’ δ 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 $\delta > 0$,

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

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’ δ 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 $\delta > 0$,

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

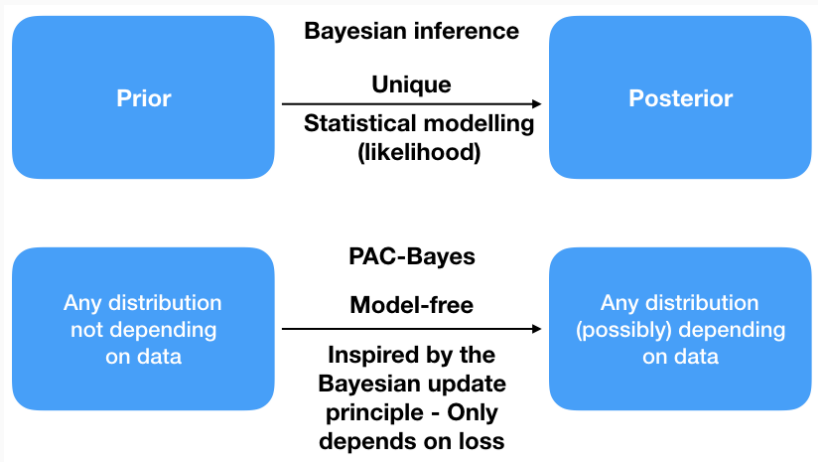
Think of $\mathcal{B}(n, \delta)$ as $\text{Complexity} \times \frac{\log 1/\delta}{\sqrt{n}}$. PAC bounds are high confidence statements on the tail of the distribution of population losses (think of a statistical test at level $1 - \delta$).

PAC-Bayes 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 $\text{KL}(Q_n || Q_0) = \mathbb{E}_{h \sim Q_n} \log \frac{Q_n(h)}{Q_0(h)}$.



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

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

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

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

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

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

Proof: let $Q \ll P$.

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

Proof: let $Q \ll P$.

$$- \text{KL}(Q \| G) = - \int \log \left(\frac{dQ}{dP} \frac{dP}{dG} \right) dQ$$

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

Proof: let $Q \ll P$.

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

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

Proof: let $Q \ll P$.

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

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

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 $Q = G$:

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

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 $Q = G$:

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

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

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?"

"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

When classical PAC-Bayes bounds fall short

- **Geometry mismatch.** The usual **KL divergence** ignores the geometry of the data space.
 - KL blows up when $\rho \not\ll \pi$,
 - offers no notion of distance or curvature.

 Bégin, Germain, Laviolette & Roy, *PAC-Bayesian bounds based on the Rényi divergence*, AISTATS, 2016.

 Alquier & Guedj, *Simpler PAC-Bayesian bounds for hostile data*, Machine Learning, 2018.

- **Not all generalisation gaps are equal.** Standard PAC-Bayes bounds control a **single scalar gap**, but cannot adapt to the structure of the prediction problem.

When classical PAC-Bayes bounds fall short

- **Geometry mismatch.** The usual **KL divergence** ignores the geometry of the data space.
 - KL blows up when $\rho \not\ll \pi$,
 - offers no notion of distance or curvature.

 Bégin, Germain, Laviolette & Roy, *PAC-Bayesian bounds based on the Rényi divergence*, AISTATS, 2016.

 Alquier & Guedj, *Simpler PAC-Bayesian bounds for hostile data*, Machine Learning, 2018.

- **Not all generalisation gaps are equal.** Standard PAC-Bayes bounds control a **single scalar gap**, but cannot adapt to the structure of the prediction problem.

Two contributions

- (1) *geometric reformulation via Wasserstein distances,*
- (2) *rethinking the notion of generalisation through comparators.*

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 $\text{KL}(\rho \parallel \pi)$, which can:
 - ignore geometry of $\mathcal{H}$ or $\mathcal{Z}$;
 - break when $\rho \not\ll \pi$;
 - be vacuous with heavy-tailed losses.

Why Wasserstein instead of KL?

- Classical PAC-Bayes bounds use $\text{KL}(\rho \parallel \pi)$, which can:
 - ignore geometry of $\mathcal{H}$ or $\mathcal{Z}$;
 - break when $\rho \not\ll \pi$;
 - be vacuous with heavy-tailed losses.
- The (order 1) **Wasserstein distance**

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

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 $S = (z_1, \dots, z_m) \sim \mu^m$.
 - **Prior** $\pi \in \mathcal{M}(\mathcal{H})$, **posterior** $\rho \in \mathcal{M}(\mathcal{H})$.
 - Split S into K disjoint subsets $S_1, \dots, S_K$.
 - Each prior $\pi_{i,S}$ is built from data disjoint from S_i (independence for the bound).
- Data-dependent priors remain valid via sample splitting.

Theorem: 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 $\rho \in \mathcal{M}(\mathcal{H})$:

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

1. Prove a **Wasserstein deviation inequality** using the Kantorovich–Rubinstein dual for W .
2. Prove **Catoni-type** high-probability control.
3. Uniformise over all ρ via $W(\rho, \pi_{i,S})$ terms.
4. Use **sample splitting** to construct independent $\pi_{i,S}$ and take a union bound over $i = 1, \dots, K$.

→ **Geometry-aware, linear in W , high-probability bound.**

From bound to learning objective

Minimising the RHS of the Theorem gives:

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

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

$$h_w^* \in \arg \min_w \hat{R}_S(h_w) + \varepsilon \sum_{i=1}^K \frac{|S_i|}{m} d(h_w, h_{w_i}).$$

→ Wasserstein acts as a geometry-aware regulariser.

Interpreting the parameter ε

- In the deterministic case, $W(\rho, \pi_{i,S}) = d(h_w, h_{w_i})$.
- The theoretical weight $2L$ becomes a tunable ε :

$$h_w^* = \arg \min_w \hat{R}_S(h_w) + \varepsilon \sum_{i=1}^K \frac{|S_i|}{m} d(h_w, h_{w_i}).$$

- ε controls the trade-off between:
 - empirical risk minimisation (fit), and
 - geometric regularisation (proximity to priors).
- Analogous to the inverse temperature $1/\lambda$ in Gibbs posteriors.

Theorem: Online Wasserstein PAC-Bayes bound

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

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

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 $(\rho_i)_{i=1}^m$,

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

Theorem: interpretation and learning rule

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

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

- 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:

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

- For deterministic learners:

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

→ **Geometry-aware online learning with transport regularisation.**

- **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



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 ℓ is bounded by 1. For any comparator Δ ,

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

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

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.

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

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

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

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

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

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 — 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 $Q_n \ll Q_0$

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

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:

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

Main takeaways

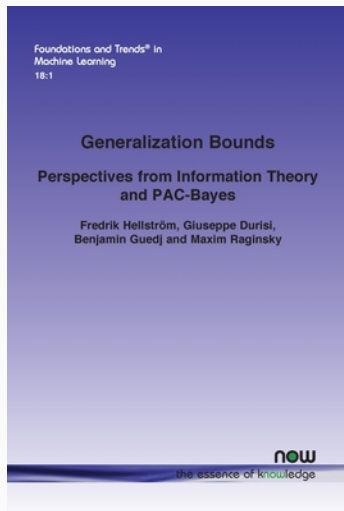
- Comparator choice is crucial in generalisation
- The optimal choice for unbounded losses: Cramér function derived from CGF
- For NEFs, this is equivalent to using the KL divergence

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.

Information theory and PAC-Bayes united

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

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.

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.

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**.
- **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.

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.

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.

→ Two complementary lenses on generalisation.

Thank you!