# Comparing Comparators in Generalisation Bounds

Benjamin Guedj

Workshop - Advances on post-Bayesian methods

15th May 2025, London



#### **Post-Bayes?**

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

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

Keywords: statistical learning theory, PAC-Bayes, generalisation bounds, computational statistics, theoretical analysis of deep learning, information theory Research at the crossroads of statistics, probability, machine learning, optimisation. *Mathematical foundations of machine learning* is a good tagline.

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

PAC-Bayes is all about understanding the generalisation abilities of hypotheses drawn from posterior distributions that are very rarely proper Bayesian posteriors.

#### Joint work with Fredrik Hellström



#### Featured at AISTATS 2024 in València

#### Take-home message

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

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

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



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 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  $h \colon \mathcal{X} \to \mathcal{Y}$ ,

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

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? 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  $\hat{L}$ , 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  $\hat{L}$ , let

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

denote the generalisation gap. We want

$$L(h) = \widehat{L}(h) + L(h) - \widehat{L}(h) = \widehat{L}(h) + \Gamma(h) \le \widehat{L}(h) + \text{Bound},$$

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

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

denote the generalisation gap. We want

$$L(h) = \widehat{L}(h) + L(h) - \widehat{L}(h) = \widehat{L}(h) + \Gamma(h) \le \widehat{L}(h) + \text{Bound},$$

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

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

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.

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

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

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

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

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

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 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,

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

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



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

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

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

- 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

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

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

### Variational definition of the $\operatorname{KL}\nolimits$ -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 $\operatorname{KL}\nolimits$ -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, A) be a measurable space.

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

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

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

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

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

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

$$-\operatorname{KL}(\boldsymbol{Q}\|\boldsymbol{G}) = -\int \log\left(\frac{\mathrm{d}\boldsymbol{Q}}{\mathrm{d}\boldsymbol{P}}\frac{\mathrm{d}\boldsymbol{P}}{\mathrm{d}\boldsymbol{G}}\right)\mathrm{d}\boldsymbol{Q}$$
$$\begin{split} -\operatorname{KL}(\boldsymbol{Q} \| \boldsymbol{G}) &= -\int \log \left( \frac{\mathrm{d}\boldsymbol{Q}}{\mathrm{d}\boldsymbol{P}} \frac{\mathrm{d}\boldsymbol{P}}{\mathrm{d}\boldsymbol{G}} \right) \mathrm{d}\boldsymbol{Q} \\ &= -\int \log \left( \frac{\mathrm{d}\boldsymbol{Q}}{\mathrm{d}\boldsymbol{P}} \right) \mathrm{d}\boldsymbol{Q} + \int \log \left( \frac{\mathrm{d}\boldsymbol{G}}{\mathrm{d}\boldsymbol{P}} \right) \mathrm{d}\boldsymbol{Q} \end{split}$$

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

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

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

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

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

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

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

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

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

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

$$Q_{\lambda} \propto \exp\left(-\lambda \widehat{\mathcal{L}}
ight) P = rging_{Q \ll P} \left\{ \widehat{\mathcal{L}}(Q) + rac{\operatorname{KL}(Q \| P)}{\lambda} 
ight\}.$$

#### "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)

## **Comparators in generalisation bounds**

- 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?

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.

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

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

where

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

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

If 
$$\widehat{\mathcal{L}} \leq \alpha$$
,  $\operatorname{KL}(Q_n || Q_0) \leq \beta$ , 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,

- $\alpha$  is the empirical loss  $\widehat{\mathcal{L}}(Q_n)$ ,
- $\beta$  is the KL divergence  $KL(Q_n || Q_0)$ ,
- *ι*(*n*) is a complexity term,
- $\delta$  is the confidence level,
- $\rho$  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  $\gamma \in \mathbb{R}$ 

$$\Delta_{\gamma}(p,q) = \gamma q - \log(1 - p + p e^{\gamma}),$$

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

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

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

# 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.

## Finding the optimal comparator

#### 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

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

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  $Q_n \ll Q_0$ :

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

where

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

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

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

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

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

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

$$\Delta_{\mathcal{P}}^{\Psi}(q,p) = \Psi_{\rho}^{*}(q) = \sup_{t\in\mathcal{T}_{\rho}} \left\{ tq - \Psi_{\rho}(t) \right\}.$$

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

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

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

- 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

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

#### **Examples of Cramér Functions**

- Bounded loss: binary KL 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:

$$egin{split} \Delta^{\Psi}_{ ext{Lap}}(q,p) &= rac{\sqrt{(q-p)^2+b^2}}{b} - 1 \ &+ \log\left(rac{2\left(b\sqrt{(q-p)^2+b^2}-b^2
ight)}{(q-p)^2}
ight) \end{split}$$

### Novel (tight) PAC-Bayes bounds

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  $Q_n \ll Q_0$ 

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

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

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

Furthermore, letting  $\bar{\Upsilon}(\mathcal{P}) := \Upsilon^{\mathcal{P}}_{\Delta^{\Psi}_{\mathcal{D}}}$ , we have:

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

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

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

- 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

- · 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

- · 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

#### **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

Foundations and Trends <sup>®</sup> in Machine Learning 18:1
Generalization Bounds
Perspectives from Information Theory and PAC-Bayes
Fredrik Hellström, Gluseppe 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

Foundo Mochin 18:1	itions and Trends <sup>®</sup> In le Learning
	Generalization Bounds
Pers	spectives from Information Theory and PAC-Bayes
	Fredrik Hellström, Giuseppe Durisi, Benjamin Guedj and Maxim Raginsky
	new
	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

Thank you!