AN INTRODUCTION TO PAC-BAYES LEARNING AND ITS LINKS TO FLAT MINIMA

SÉMINAIRE SO

Maxime Haddouche

Inria Paris

Mardi 4 Février 2025

SUMMARY

1. Introduction to PAC-Bayes Learning
2. PAC-Bayes with Weak Statistical Assumptions
3. Involving Flat Minima in PAC-Bayes

What is a learning theory problem?

A tuple $(\mathcal{Z}, \mathcal{H}, \ell)$: a data space $\mathcal{Z}$, a predictor space $h \in \mathcal{H}$, a mathematically well-defined problem $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$

What is a learning theory problem?

A tuple $(\mathcal{Z}, \mathcal{H}, \ell)$: a data space $\mathcal{Z}$, a predictor space $h \in \mathcal{H}$, a mathematically well-defined problem $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$

What is our goal?

We have access to a m-sized training set $\mathcal{S}_m = \{\mathbf{z}_1, \cdots, \mathbf{z}_m\}$. We aim to learn the best $h^\star \in \mathcal{H}$ to answer ℓ in a certain way

What is a learning theory problem?

A tuple $(\mathcal{Z}, \mathcal{H}, \ell)$: a data space $\mathcal{Z}$, a predictor space $h \in \mathcal{H}$, a mathematically well-defined problem $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$

What is our goal?

We have access to a m-sized training set $\mathcal{S}_m = \{\mathbf{z}_1, \cdots, \mathbf{z}_m\}$. We aim to learn the best $h^\star \in \mathcal{H}$ to answer ℓ in a certain way

- **Optimisation:** minimise the empirical risk $h^\star \in \operatorname{argmin}_{h \in \mathcal{H}} \hat{\mathsf{R}}_{\mathcal{S}_m}(h) := \frac{1}{m} \sum_{i=1}^{m} \ell(h, \mathbf{z}_i)$
- **Generalisation:** if $\mathcal{S}_m \sim \mathcal{D}^{\otimes m}$, minimise the theoretical risk $h^\star \in \operatorname{argmin}_{h \in \mathcal{H}} \mathsf{R}_{\mathcal{D}}(h) := \mathbb{E}_{\mathbf{z} \sim \mathcal{D}}[\ell(h, \mathbf{z})]$

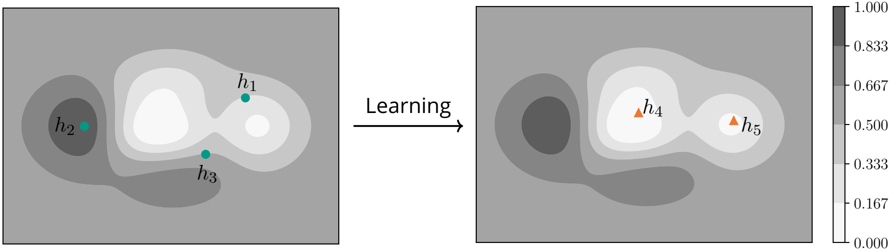
Supervised learning with linear classifiers:

- $\mathcal{Z} = \mathbb{R}^k \times \mathcal{Y}$ with $\mathcal{Y} = \{-1, 1\}$
- Loss $\ell(h, (x, y)) = \mathbb{1}\{h(x) \neq y\}$
- Linear classifiers: $\mathcal{H} := \{h_\theta(x) = sgn(\langle \theta, x \rangle)\}$, where $sgn(a)$ denotes the sign of a.

It may be hard to find directly the best h for complex predictor classes (eg neural nets). What could we do?

PAC-Bayes: Find the best distribution over $\mathcal{H}$!

Learning a posterior Q over models from m data and a prior distribution P

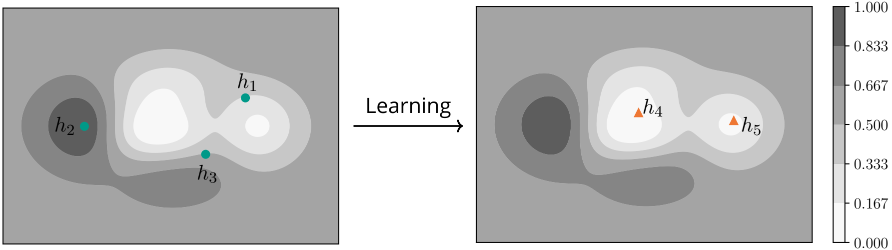


Prior distribution P Posterior distribution Q Loss

Image from Paul Viallard.

PAC-Bayes: Find the best distribution over $\mathcal{H}$!

Learning a posterior Q over models from m data and a prior distribution P



Prior distribution P | Posterior distribution Q | Loss

PAC-Bayesian generalisation bounds in a nutshell

With probability at least $1-\delta$

$$\text{performance gap}(\mathrm{Q}) \leq \text{bound}\left(\text{complexity}(\mathrm{Q}, \mathrm{P}), \frac{1}{m}, \ln\frac{1}{\delta}\right).$$

Image from Paul Viallard.

Notations:

- Predictor/hypothesis $h \in \mathcal{H}$, Data space $\mathcal{Z}$
- Loss $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$,
- Countable learning sample $\mathcal{S} = (\mathbf{z}_i)_{i \geq 1} \in \mathcal{Z}^{\mathbb{N}}$, with distribution $\mathcal{D}_{\mathcal{S}}$
- $\mathcal{S}_m$: Restriction of $\mathcal{S}$ to m first points with distribution $\mathcal{D}_m$
- Space of distributions over $\mathcal{H}$: $\mathcal{M}(\mathcal{H})$
- Posterior and prior distribution $\mathrm{Q}, \mathrm{P} \in \mathcal{M}(\mathcal{H})^2$
- If $\mathcal{S}_m \sim \mathcal{D}^m$ *i.i.d.*, Risks: $\mathrm{R}_{\mathcal{D}}(h) = \mathbb{E}_{\mathbf{z} \sim \mathcal{D}}\, \ell(h, \mathbf{z})$, $\hat{\mathrm{R}}_{\mathcal{S}_m}(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, \mathbf{z}_i)$
- Expected risks $\mathrm{R}_{\mathcal{D}}(\mathrm{Q}) = \mathop{\mathbb{E}}_{h \sim \mathrm{Q}} [\mathrm{R}_{\mathcal{D}}(h)]$, $\hat{\mathrm{R}}_{\mathcal{S}_m}(\mathrm{Q}) = \mathop{\mathbb{E}}_{h \sim \mathrm{Q}} [\hat{\mathrm{R}}_{\mathcal{S}_m}(h)]$

A FUNDAMENTAL RESULT: MCALLESTER'S BOUND

McAllester's bound (Maurer's improvement) Maurer (2004, Theorem 5) ($\ell \in [0,1]$)

For any $\mathrm{P} \in \mathcal{M}(\mathcal{H})$, with probability $1-\delta$ over $\mathcal{S}_m \sim \mathcal{D}^m$, for any $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$,

$$\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq \hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \sqrt{\frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \ln\frac{2\sqrt{m}}{\delta}}{2m}},$$

where $\mathrm{KL}(\mathrm{Q},\mathrm{P}) = \mathbb{E}_{h\sim\mathrm{Q}}\left[\frac{dQ}{dP}(h)\right]$.

No explicit dependency in the dimension of the problem (hidden in the KL term): positive phenomenon can be caught with the right priors (*e.g.* sparsity).

A SIMPLE ROUTE OF PROOF

Step 1: A key ingredient: change of measure inequality

For any function f, any $Q \ll P$:

$$\mathbb{E}_{h\sim Q}[f(h)] - \ln\left(\mathbb{E}_{h\sim P}[\exp \circ f(h)]\right) \leq \mathrm{KL}(Q, P).$$

A SIMPLE ROUTE OF PROOF

Step 1: A key ingredient: change of measure inequality

For any function f, any $\mathrm{Q} \ll \mathrm{P}$:

$$\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}[f(h)] - \ln\left(\mathop{\mathbb{E}}_{h\sim\mathrm{P}}[\exp\circ f(h)]\right) \leq \mathrm{KL}(\mathrm{Q},\mathrm{P}).$$

Step 2: Markov's inequality

With probability at least $1-\delta$:

$$\begin{aligned}\mathop{\mathbb{E}}_{h\sim\mathrm{P}}[\exp\circ f(h)] &\leq \frac{1}{\delta}\mathop{\mathbb{E}}_{\mathcal{S}_m}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{P}}[\exp\circ f(h)]\right],\\ &= \frac{1}{\delta}\mathop{\mathbb{E}}_{h\sim\mathrm{P}}\left[\mathop{\mathbb{E}}_{\mathcal{S}_m}[\exp\circ f(h)]\right]. \qquad \text{(P data-free + Fubini)}\end{aligned}$$

A SIMPLE ROUTE OF PROOF (2)

Step 3: Choosing the right f.

Take $f(h) = m\,\mathrm{kl}\left(\mathsf{R}_{\mathcal{D}}(h), \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right)$ (kl= KL of Bernoullis).
Then Maurer (2004): for any h, loss in $[0,1]$:

$$\mathop{\mathbb{E}}_{\mathcal{S}_m}\left[\exp \circ f(h)\right] \leq 2\sqrt{m}$$

To conclude: $\mathrm{kl}(p,q) \geq 2(p-q)^2$.

A SECOND KEY RESULT: CATONI'S BOUND

Catoni's bound Alquier *et al.* (2016, Theorem 4.1) (ℓ σ-subgaussian)

For $\lambda > 0$, $\mathrm{P} \in \mathcal{M}(\mathcal{H})$, with probability $1{-}\delta$ over $\mathcal{S}_m \sim \mathcal{D}^m$, for any $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$,

$$\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq \hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{\mathrm{KL}(\mathrm{Q}, \mathrm{P}) + \ln \frac{1}{\delta}}{\lambda} + \frac{\lambda \sigma^2}{2m}.$$

Previous bounds: both fully empirical $\rightarrow$ optimisation in Q is feasible on $\mathcal{C} \subseteq \mathcal{M}(\mathcal{H})$!

$$\text{McAllester} \quad \mathrm{Q}_M := \underset{\mathrm{Q} \in \mathcal{C}}{\operatorname{argmin}}\ \hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \sqrt{\frac{\mathrm{KL}(\mathrm{Q}, \mathrm{P})}{2m}}.$$

For any $\lambda > 0$,

$$\text{Catoni} \quad \mathrm{Q}_C := \underset{\mathrm{Q} \in \mathcal{C}}{\operatorname{argmin}}\ \hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{\mathrm{KL}(\mathrm{Q}, \mathrm{P})}{\lambda}.$$

If $\mathcal{C} = \mathcal{M}(\mathcal{H})$, a *Gibbs posterior* $\mathrm{P}_{-\lambda \hat{\mathsf{R}}_{\mathcal{S}_m}}$ is the explicit minimiser of Catoni's bound:

$$d\mathrm{P}_{-\lambda \hat{\mathsf{R}}_{\mathcal{S}_m}}(h) = \frac{\exp(-\lambda \hat{\mathsf{R}}_{\mathcal{S}_m}(h))}{\mathbb{E}_{h \sim \mathrm{P}}[\exp(-\lambda \hat{\mathsf{R}}_{\mathcal{S}_m}(h))]} d\mathrm{P}(h).$$

Quick sum up

PAC-Bayes algorithms minimise theoretical bounds $\rightarrow$ sound theoretical guarantees comes with our posterior.

Drawbacks Often hard to optimise on $\mathcal{M}(\mathcal{H})$, and Gibbs posterior implementation is time-consuming.

PRACTICAL INSTANTIATION

Quick sum up

PAC-Bayes algorithms minimise theoretical bounds $\rightarrow$ sound theoretical guarantees comes with our posterior.

Drawbacks Often hard to optimise on $\mathcal{M}(\mathcal{H})$, and Gibbs posterior implementation is time-consuming.

Questions:

- **How are those algorithms instantiated in practice?**
- **Are these algorithms efficient and do they come with non-vacuous theoretical guarantees?**

COMMON PRACTICES

Instantiation

- Use of multiple data-free priors (grid + union bounds)
- Sacrifice some part of the data to train the prior.
- $\mathcal{C}$ is often a set of Gaussians (closed form of the KL)

Efficiency

- Non-vacuous generalisation guarantees attainable for small deep nets (Dziugaite *et al.*, 2017 and following works)
- Faster convergence rates via small variance (Tolstikhin *et al.*, 2013)
- When vacuous, use of PAC-Bayes bounds as correlation measures for generalisation (Neyshabur *et al.*, 2017)

TAKE HOME MESSAGES

In 20+ years of development:

- Inspiration from the Bayesian paradigm
- Little attention on statistical assumptions (*i.i.d.* data, subgaussian losses) except few works *e.g.* (Seldin *et al.*, 2012; Kuzborskij *et al.*, 2019).
- Priors and posteriors are designed *w.r.t.* to the KL divergence: either Gibbs (closed form) or Gaussian (computation)

Can we weaken the statistical assumption on the loss?

TAKE HOME MESSAGES

In 20+ years of development:

- Inspiration from the Bayesian paradigm
- Little attention on statistical assumptions (*i.i.d.* data, subgaussian losses) except few works *e.g.* (Seldin *et al.*, 2012; Kuzborskij *et al.*, 2019).
- Priors and posteriors are designed *w.r.t.* to the KL divergence: either Gibbs (closed form) or Gaussian (computation)

Can we weaken the statistical assumption on the loss?

Yes: we can extend Catoni's bound for any countable dataset $\mathcal{S}$ and finite variance assumption

Theorem

For any data-free prior $\mathrm{P} \in \mathcal{M}(\mathcal{H})$, any $\lambda > 0$, any collection of martingales $(M_m(h))_{m\geq 1}$ indexed by $h \in \mathcal{H}$, the following holds with probability $1-\delta$ over the sample $\mathcal{S} = (\mathbf{z}_i)_{i\in\mathbb{N}}$, for all $m \in \mathbb{N}/\{0\}$, $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$:

$$|M_m(\mathrm{Q})| \leq \frac{\mathrm{KL}(\mathrm{Q}, \mathrm{P}) + \log(2/\delta)}{\lambda} + \frac{\lambda}{2}\left([M]_m(\mathrm{Q}) + \langle M \rangle_m(\mathrm{Q})\right).$$

Theorem

For any data-free prior $\mathrm{P} \in \mathcal{M}(\mathcal{H})$, any $\lambda > 0$, any collection of martingales $(M_m(h))_{m\geq 1}$ indexed by $h \in \mathcal{H}$, the following holds with probability $1-\delta$ over the sample $\mathcal{S} = (\mathbf{z}_i)_{i\in\mathbb{N}}$, for all $m \in \mathbb{N}/\{0\}$, $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$:

$$|M_m(\mathrm{Q})| \leq \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(2/\delta)}{\lambda} + \frac{\lambda}{2}\left([M]_m(\mathrm{Q}) + \langle M\rangle_m(\mathrm{Q})\right).$$

Required: finiteness of $([M_m], \langle M\rangle_m)_{m\geq 1}$ (variance terms)

Toolbox: Ville's inequality and supermartingales

Corollary

For any data-free prior $P \in \mathcal{M}(\mathcal{H})$, any $\lambda > 0$, the following holds with probability $1-\delta$ over the sample $\mathcal{S} = (\mathbf{z}_i)_{i\in\mathbb{N}}$, for all $m \in \mathbb{N}/\{0\}$, $Q \in \mathcal{M}(\mathcal{H})$

$$\mathbb{E}_{h\sim \mathrm{Q}}[\mathsf{R}(h)] \leq \mathbb{E}_{h\sim \mathrm{Q}}\left[\hat{\mathsf{R}}_{\mathcal{S}_m}(h) + \frac{\lambda}{2m}\sum_{i=1}^{m}\left(\ell(h,z_i) - \mathsf{R}_{\mathcal{D}}(h)\right)^2\right] + \frac{\mathrm{KL}(\mathrm{Q},\mathrm{P}) + \log(2/\delta)}{\lambda m} + \frac{\lambda}{2}\mathbb{E}_{h\sim \mathrm{Q}}\left[Var_{\mathcal{D}}(h)\right],$$

where $Var_{\mathcal{D}}(h)$ is the variance of $\ell(h,\cdot)$.

Interesting property: time-uniform bound.

A TIGHTER BOUND FOR NON-NEGATIVE LOSSES

In Chugg *et al.* (2023): tighter bound for nonnegative loss:

Corollary

For $\ell \geq 0$, any data-free prior $P \in \mathcal{M}(\mathcal{H})$, any $\lambda > 0$, the following holds with probability $1-\delta$ over the sample $\mathcal{S} = (\mathbf{z}_i)_{i\in\mathbb{N}}$, for all $m \in \mathbb{N}/\{0\}$, $Q \in \mathcal{M}(\mathcal{H})$

$$\mathbb{E}_{h\sim Q}[\mathsf{R}(h)] \leq \mathbb{E}_{h\sim Q}\left[\hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right] + \frac{\mathrm{KL}(Q,P)+\log(2/\delta)}{\lambda m} + \frac{\lambda}{2}\mathbb{E}_{h\sim Q}\left[\ell(h,\mathbf{z})^2\right],$$

TAKE-HOME MESSAGE

We can recover Catoni's bound at the sole price of uniformly bounded variance. We also reached time-uniform generalisation bounds

We can recover Catoni's bound at the sole price of uniformly bounded variance. We also reached time-uniform generalisation bounds

Drawback: those bounds holds for any $\mathbb{Q}$ and $\mathcal{D}$ simultaneously. Gastpar *et al.* (2023) showed such bounds were limited in the overparametrised setting.

We can recover Catoni's bound at the sole price of uniformly bounded variance. We also reached time-uniform generalisation bounds

Drawback: those bounds holds for any $\mathbb{Q}$ and $\mathcal{D}$ simultaneously. Gastpar *et al.* (2023) showed such bounds were limited in the overparametrised setting.

Question: can we incorporate some benefits of a successful learning process in such bounds?

Yes for flat minima !!

What is a flat minimum?

A minimum such that its neighbourhood nearly minimises the loss.

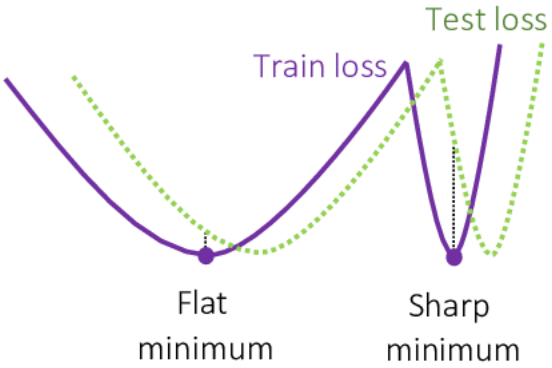


Image from **liebenwein2021sparse**.

FLAT MINIMA AND GENERALISATION ARE CORRELATED!

Correlations with generalisation recently emerged:

- Flat minima of $\hat{\mathrm{R}}_{\mathcal{S}}$.
  PAC-Bayes based correlation measure : works for many datasets (Neyshabur *et al.*, 2017; Dziugaite *et al.*, 2020; Jiang *et al.*, 2020)

- Flat minima of the adversarial loss in the context of adversarially robust learning. (Stutz *et al.*, 2021)

- Flat minima implies generalisation for 2-layers nets (Wen *et al.*, 2023).

FLAT MINIMA AND GENERALISATION ARE CORRELATED!

Correlations with generalisation recently emerged:

- Flat minima of $\hat{\mathrm{R}}_{\mathcal{S}}$.
  PAC-Bayes based correlation measure : works for many datasets (Neyshabur *et al.*, 2017; Dziugaite *et al.*, 2020; Jiang *et al.*, 2020)

- Flat minima of the adversarial loss in the context of adversarially robust learning. (Stutz *et al.*, 2021)

- Flat minima implies generalisation for 2-layers nets (Wen *et al.*, 2023).

Can we go beyond correlation or 2-layers net and obtain sound generalisation bounds involving directly flat minima?

Notation: for any Q, $\mathrm{H}^1(\mathrm{Q}) := \{f \in \mathrm{L}^2(\mathrm{Q}) \cap \mathrm{D}_1(\mathbb{R}^d) \mid \|\nabla f\| \in \mathrm{L}^2(\mathrm{Q})\}$

Poincaré

Q is $\mathtt{Poinc}(c_P)$ if for all $f \in \mathrm{H}^1(\mathrm{Q})$:

$$\mathrm{Var}(f) \leq c_P(\mathrm{Q}) \mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[\|\nabla f(h)\|^2\right],$$

Log-Sobolev

Q is $\mathtt{L\text{-}Sob}(c_{LS})$ if for all function $f \in \mathrm{H}^1(\mathrm{Q})$:

$$\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[f^2(h)\log\left(\frac{f^2(h)}{\mathbb{E}_{h\sim\mathrm{Q}}\left[f^2(h)\right]}\right)\right] \leq c_{LS}(\mathrm{Q}) \mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left[\|\nabla f(h)\|^2\right],$$

ESSENTIAL TOOLS: POINCARÉ AND LOG-SOBOLEV INEQUALITIES

Notation: for any Q, $\mathrm{H}^1(\mathrm{Q}) := \{f \in \mathrm{L}^2(\mathrm{Q}) \cap \mathrm{D}_1(\mathbb{R}^d) \mid \|\nabla f\| \in \mathrm{L}^2(\mathrm{Q})\}$

Poincaré

Q is $\mathtt{Poinc}(c_P)$ if for all $f \in \mathrm{H}^1(\mathrm{Q})$:

$$\mathrm{Var}(f) \leq c_P(\mathrm{Q}) \underset{h\sim\mathrm{Q}}{\mathbb{E}} \left[\|\nabla f(h)\|^2\right],$$

Log-Sobolev

Q is $\mathtt{L\text{-}Sob}(c_{LS})$ if for all function $f \in \mathrm{H}^1(\mathrm{Q})$:

$$\underset{h\sim\mathrm{Q}}{\mathbb{E}} \left[f^2(h) \log\left(\frac{f^2(h)}{\mathbb{E}_{h\sim\mathrm{Q}}\left[f^2(h)\right]}\right)\right] \leq c_{LS}(\mathrm{Q}) \underset{h\sim\mathrm{Q}}{\mathbb{E}} \left[\|\nabla f(h)\|^2\right],$$

Gaussian distributions and Gibbs posteriors are `Poinc` and `L-Sob`!

Notation: $\mathrm{Err}(\ell, \mathrm{Q}, \mathbf{z}) := \mathbb{E}_{h\sim \mathrm{Q}}[\ell(h, \mathbf{z})]$

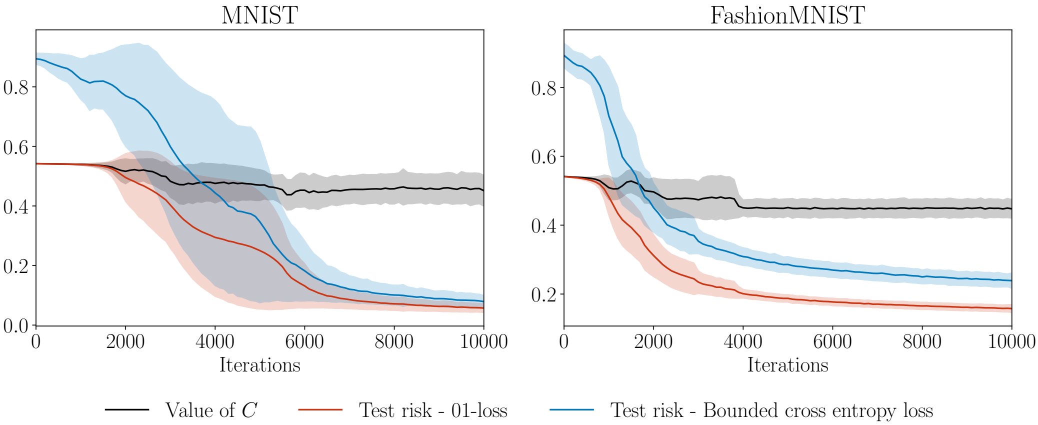
Assumption

$\mathrm{Q} \in \mathcal{M}(\mathcal{H})$ is *quadratically self-bounded w.r.t.* ℓ and $C > 0$ (namely $\mathtt{QSB}(\ell, C)$) if

$$\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[\mathrm{Err}(\ell, \mathrm{Q}, \mathbf{z})^2\right] \leq C\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \; (= C\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\left[\mathrm{Err}(\ell, \mathrm{Q}, \mathbf{z})\right])$$

- QSB intricates $\mathcal{D} \in \mathcal{M}(\mathcal{Z})$ with $\mathrm{Q} \in \mathcal{M}(\mathcal{H})$
- Satisfied if $\ell \in [0, K]$ with $C = K$.
- Also satisfied for unbounded lipschitz losses in a certain setting.

QSB holds for 3-layer neural nets trained on MNIST (black curve)!



GENERALISATION BOUNDS VIA FLAT MINIMA (2)

Theorem

For any $C > 0$, data-free prior P, with probability at least $1 - \delta$ for any $m > 0$, and Q being $\texttt{Poinc}(c_P)$, $\texttt{QSB}(\ell, C)$,

$$\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq 2\hat{\mathsf{R}}_{\mathcal{S}}(\mathrm{Q}) + 2C\frac{KL(\mathrm{Q},\mathrm{P}) + \log(1/\delta)}{m} + \frac{1}{C}c_P(\mathrm{Q}) \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right].$$

GENERALISATION BOUNDS VIA FLAT MINIMA (2)

Theorem

For any $C > 0$, data-free prior P, with probability at least $1-\delta$ for any $m > 0$, and Q being $\texttt{Poinc}(c_P)$, $\texttt{QSB}(\ell, C)$,

$$\mathrm{R}_{\mathcal{D}}(\mathrm{Q}) \leq 2\hat{\mathrm{R}}_{\mathcal{S}}(\mathrm{Q}) + 2C\frac{KL(\mathrm{Q},\mathrm{P}) + \log(1/\delta)}{m} + \frac{1}{C}c_P(\mathrm{Q}) \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right].$$

If $\mathcal{D}$ is also $\texttt{Poinc}$:

With more minor technical assumptions, for any Q being $\texttt{Poinc}(c_P)$ with $\mathrm{R}_{\mathcal{D}}(\mathrm{Q}) \leq C$:

$$\mathrm{R}_{\mathcal{D}}(\mathrm{Q}) \leq 2\hat{\mathrm{R}}_{\mathcal{S}}(\mathrm{Q}) + 2C\frac{KL(\mathrm{Q},\mathrm{P}) + \log(1/\delta)}{m} + \frac{1}{C}\left(c_P(\mathrm{Q}) \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}\left(\|\nabla_h \ell(h,\mathbf{z})\|^2\right)\right] + c_P(\mathcal{D}) \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}}\left(\left\|\mathop{\mathbb{E}}_{h\sim\mathrm{Q}}[\nabla_z \ell(h,\mathbf{z})]\right\|^2\right)\right).$$

FULLY EMPIRICAL FAST RATE

Drawback: bounds are not empirical.

FULLY EMPIRICAL FAST RATE

Drawback: bounds are not empirical.

Solution: $\mathcal{C}^2$ gradient-lipschitz losses!

Theorem

For any $C_1, C_2, c > 0$, with probability at least $1-\delta$, for any $m > 0$, Q being $\mathtt{Poinc}(c_P)$ with constant c, $\mathtt{QSB}(\ell, C_1)$, $\mathtt{QSB}\left(\|\nabla_h \ell\|^2, C_2\right)$,

$$\mathsf{R}_{\mathcal{D}}(\mathrm{Q}) \leq 2\hat{\mathsf{R}}_{\mathcal{S}}(\mathrm{Q}) + \mathcal{O}\left(\mathop{\mathbb{E}}_{h\sim \mathrm{Q}} \left[\frac{1}{m} \sum_{i=1}^{m} \|\nabla_h \ell(h, \mathbf{z}_i)\|^2 \right] + \frac{\mathrm{KL}(\mathrm{Q}, \mathrm{P}) + \log(1/\delta)}{m} \right).$$

If Q satisfies either

1. Flat minima for $\hat{R}_{\mathcal{S}}$ and $R_{\mathcal{D}}$,
2. if ℓ gradient-lipschitz, flat minima for $\hat{R}_{\mathcal{S}}$ and small gradient norms on each training data,

then Q generalises well!

Drawback: with Poincaré posteriors, KL is uncontrolled.

Drawback: with Poincaré posteriors, KL is uncontrolled.

Solution: Gibbs posterior with log-Sobolev priors!

Definition

$\mathrm{P}_{-\gamma \hat{\mathsf{R}}_{\mathcal{S}}}$ is the Gibbs posterior *w.r.t.* prior P with *inverse temperature* $\gamma > 0$ if

$$d\mathrm{P}_{-\gamma \hat{\mathsf{R}}_{\mathcal{S}}}(h) \propto \exp\left(-\gamma \hat{\mathsf{R}}_{\mathcal{S}}(h)\right) dP(h)$$

.

Why focus on those?

- Minimise Catoni's bound
- if P `L-Sob`(+ technical assumptions) and $\ell = \ell_1 + \ell_2$ (ℓ_1 convex, twice differentiable, ℓ_2 bounded) then $\mathrm{P}_{-\gamma \hat{\mathsf{R}}_{\mathcal{S}}}$ is `L-Sob`.

Theorem

For any $C > 0$, any $\gamma > 0$, any prior P $\texttt{L-Sob}(c_{LS})$ (+ technical assumptions), if $\ell = \ell_1 + \ell_2$ (as above), then with probability at least $1 - \delta$, for any $m > 0$, Q being $\texttt{QSB}(\ell, C)$:

$$\mathsf{R}_{\mathcal{D}}(\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}}}) \leq 2\hat{\mathsf{R}}_{\mathcal{S}}(\mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}}}) + \mathcal{O}\left(C \frac{\gamma^2 \, \mathbb{E}_{h\sim \mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}}}} \left[\|\nabla_h \hat{\mathsf{R}}_{\mathcal{S}}(h)\|^2 \right] + \log(1/\delta)}{m} + \frac{1}{C} \mathop{\mathbb{E}}_{\mathbf{z}\sim\mathcal{D}} \left[\mathop{\mathbb{E}}_{h\sim \mathrm{P}_{-\gamma\hat{\mathsf{R}}_{\mathcal{S}}}} \left(\|\nabla_h \ell(h,\mathbf{z})\|^2 \right) \right] \right).$$

Theorem

For any $C > 0$, any $\gamma > 0$, any prior P $\texttt{L-Sob}(c_{LS})$ (+ technical assumptions), if $\ell = \ell_1 + \ell_2$ (as above), then with probability at least $1 - \delta$, for any $m > 0$, Q being $\texttt{QSB}(\ell, C)$:

$$\mathsf{R}_{\mathcal{D}}(\mathrm{P}_{-\gamma \hat{\mathsf{R}}_{\mathcal{S}}}) \leq 2\hat{\mathsf{R}}_{\mathcal{S}}(\mathrm{P}_{-\gamma \hat{\mathsf{R}}_{\mathcal{S}}}) + \mathcal{O}\left(C \frac{\gamma^2 \, \mathbb{E}_{h \sim \mathrm{P}_{-\gamma \hat{\mathsf{R}}_{\mathcal{S}}}} \left[\|\nabla_h \hat{\mathsf{R}}_{\mathcal{S}}(h)\|^2 \right] + \log(1/\delta)}{m} + \frac{1}{C} \mathop{\mathbb{E}}_{\mathbf{z} \sim \mathcal{D}} \left[\mathop{\mathbb{E}}_{h \sim \mathrm{P}_{-\gamma \hat{\mathsf{R}}_{\mathcal{S}}}} \left(\|\nabla_h \ell(h, \mathbf{z})\|^2 \right) \right] \right).$$

KL small if a flat minima on $\hat{\mathsf{R}}_{\mathcal{S}}$ is reached

TAKE-HOME MESSAGES

1. Gibbs posterior generalises well if they reach a flat minima on both $\hat{R}_{\mathcal{S}}$ and $R_{\mathcal{D}}$.
2. Flatness of the minimum on $\hat{R}_{\mathcal{S}}$ controls the expansion of KL.

RESULT FOR DETERMINISTIC PREDICTORS

Drawback: results hold for probabilistic predictors

RESULT FOR DETERMINISTIC PREDICTORS

Drawback: results hold for probabilistic predictors

Answer: Exploit the 2-Wasserstein distance to obtain guarantees valid for deterministic predictors (Diracs)

Key tool: a novel change of measure inequality

For any f gradient lipschitz, any P, Q:

$$\mathbb{E}_{h\sim \mathrm{Q}}[f(h)] \leq \frac{G}{2} W_2^2(Q,P) + \mathbb{E}_{h\sim \mathrm{P}}[f(h)] + D\mathbb{E}_{h\sim \mathrm{Q}}[||\nabla f(h)||].$$

NB: a variant of this formula with a KL is attainable if $\mathrm{Q} << \mathrm{P}$ and P is `L-Sob` !

Assumption

- A relaxation of gradient-lipschitz loss.
- $\mathrm{P} \propto exp(-V(h))dh$

Theorem

Let $\delta \in (0,1)$ and $P \in \mathcal{M}(\mathcal{H})$ a data-free prior. Assume $\mathcal{H}$ has a finite diameter $D > 0$, $\ell \geq 0$ and that for any m, the generalisation gap $\Delta_{\mathcal{S}_m}$ is G gradient-Lipschitz. Assume that $\mathbb{E}_{h\sim \mathrm{P}}\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[\ell(h,z)^2] \leq \sigma^2$, then the following holds with probability at least $1-\delta$, for any $m > 0$ and any Q:

$$\mathsf{R}_D(\mathrm{Q}) \leq \hat{\mathsf{R}}_{\mathcal{S}_m}(\mathrm{Q}) + \frac{G}{2}W_2^2(\mathrm{Q},\mathrm{P}) + \sqrt{\frac{2\sigma^2 \log\left(\frac{1}{\delta}\right)}{m}} + D\mathbb{E}_{h\sim\mathrm{Q}}\left(\left\|\nabla_h \mathsf{R}_{\mathcal{D}}(h) - \nabla_h \hat{\mathsf{R}}_{\mathcal{S}_m}(h)\right\|\right)$$

CONCLUSION

- We mathematically quantify the impact of flat minima in generalisation!
- The QSB condition is verified on basic neural nets (classification) with constant C sharper than 1!
- A crucial future lead: understanding why optimisation procedures on deep nets lead to flat minima: **here, we are only able to explain why flat minima generalise well, not how we reach them.**

Full paper available at `https://arxiv.org/abs/2402.08508`

Thank you for your attention!

REFERENCES I

- Andreas Maurer. A note on the PAC-Bayesian theorem. *arXiv*. cs/0411099. (2004).

- Yevgeny Seldin, François Laviolette, Nicolò Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. PAC-Bayesian Inequalities for Martingales. *IEEE Transactions on Information Theory*. (2012).

- Ilya O. Tolstikhin and Yevgeny Seldin. PAC-Bayes-Empirical-Bernstein Inequality. *Advances in Neural Information Processing Systems (NeurIPS)*. (2013).

- Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*. (2016).

- Gintare Karolina Dziugaite and Daniel Roy. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. *Conference on Uncertainty in Artificial Intelligence (UAI)*. (2017).

REFERENCES II

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring Generalization in Deep Learning. *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. (2017). url: https://proceedings.neurips.cc/paper/2017/hash/10ce03a1ed01077e3e289f3e53c72813-Abstract.html.

Ilja Kuzborskij and Csaba Szepesvári. Efron-Stein PAC-Bayesian Inequalities. *arXiv*. abs/1909.01931. (2019).

Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M. Roy. In search of robust measures of generalization. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. (2020). url: https://proceedings.neurips.cc/paper/2020/hash/86d7c8a08b4aaa1bc7c599473f5dddda-Abstract.html.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic Generalization Measures and Where to Find Them. *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. (2020). url: https://openreview.net/forum?id=SJgIPJBFvH.

REFERENCES III

David Stutz, Matthias Hein, and Bernt Schiele. Relating Adversarially Robust Generalization to Flat Minima. *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. *IEEE*. (2021). doi: 10.1109/ICCV48922.2021.00771. url: https://doi.org/10.1109/ICCV48922.2021.00771.

Ben Chugg, Hongjian Wang, and Aaditya Ramdas. A Unified Recipe for Deriving (Time-Uniform) PAC-Bayes Bounds. *Journal of Machine Learning Research*. (2023). url: http://jmlr.org/papers/v24/23-0401.html.

Michael Gastpar, Ido Nachum, Jonathan Shafer, and Thomas Weinberger. Fantastic Generalization Measures are Nowhere to be Found. (2023).

Kaiyue Wen, Zhiyuan Li, and Tengyu Ma. Sharpness Minimization Algorithms Do Not Only Minimize Sharpness To Achieve Better Generalization. *Thirty-seventh Conference on Neural Information Processing Systems*. (2023). url: https://openreview.net/forum?id=Dkmpa6wCIx.

REFERENCES I

Andreas Maurer. A note on the PAC-Bayesian theorem. *arXiv*. cs/0411099. (2004).

Yevgeny Seldin, François Laviolette, Nicolò Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. PAC-Bayesian Inequalities for Martingales. *IEEE Transactions on Information Theory*. (2012).

Ilya O. Tolstikhin and Yevgeny Seldin. PAC-Bayes-Empirical-Bernstein Inequality. *Advances in Neural Information Processing Systems (NeurIPS)*. (2013).

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*. (2016).

Gintare Karolina Dziugaite and Daniel Roy. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. *Conference on Uncertainty in Artificial Intelligence (UAI)*. (2017).

REFERENCES II

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring Generalization in Deep Learning. *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. (2017). url: https://proceedings.neurips.cc/paper/2017/hash/10ce03a1ed01077e3e289f3e53c72813-Abstract.html.

Ilja Kuzborskij and Csaba Szepesvári. Efron-Stein PAC-Bayesian Inequalities. *arXiv*. abs/1909.01931. (2019).

Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M. Roy. In search of robust measures of generalization. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. (2020). url: https://proceedings.neurips.cc/paper/2020/hash/86d7c8a08b4aaa1bc7c599473f5dddda-Abstract.html.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic Generalization Measures and Where to Find Them. *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. (2020). url: https://openreview.net/forum?id=SJgIPJBFvH.

REFERENCES III

David Stutz, Matthias Hein, and Bernt Schiele. Relating Adversarially Robust Generalization to Flat Minima. *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. *IEEE*. (2021). doi: 10.1109/ICCV48922.2021.00771. url: https://doi.org/10.1109/ICCV48922.2021.00771.

Ben Chugg, Hongjian Wang, and Aaditya Ramdas. A Unified Recipe for Deriving (Time-Uniform) PAC-Bayes Bounds. *Journal of Machine Learning Research*. (2023). url: http://jmlr.org/papers/v24/23-0401.html.

Michael Gastpar, Ido Nachum, Jonathan Shafer, and Thomas Weinberger. Fantastic Generalization Measures are Nowhere to be Found. (2023).

Kaiyue Wen, Zhiyuan Li, and Tengyu Ma. Sharpness Minimization Algorithms Do Not Only Minimize Sharpness To Achieve Better Generalization. *Thirty-seventh Conference on Neural Information Processing Systems*. (2023). url: https://openreview.net/forum?id=Dkmpa6wCIx.

Thank you for your attention!