Block-Sample MAC-Bayes Generalization Bounds

Matthias Frey, Jingge Zhu, and Michael C. Gastpar

Department of Electrical and Electronic Engineering, The University of Melbourne
Laboratory for Information in Networked Systems, École Polytechnique Fédérale de Lausanne (EPFL)

The Fourteenth International Conference on Learning Representations (ICLR 2026)

Generalization Error

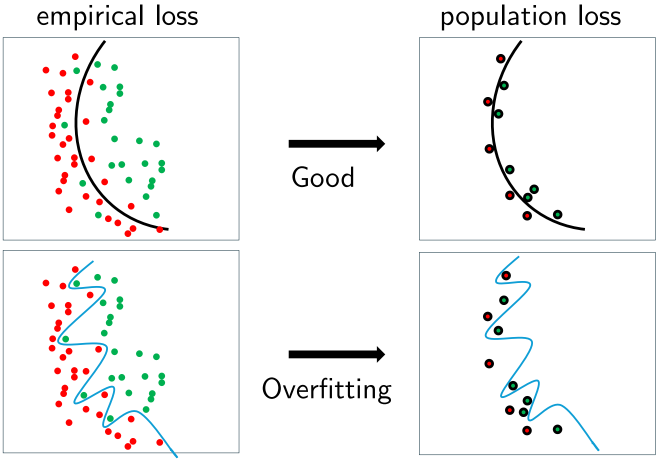


Figure credit: Muhan Guan

Expected generalization error $\mathrm{gen} := \mathbb{E}_{P_{S,W}}\Big(\underbrace{L(W)}_{\text{population loss}} - \underbrace{\hat{L}(W,S)}_{\text{empirical loss}} \Big)$

Classical Bounds

Bounds are typically in terms of a (generalized) comparator function and include an information/divergence term, a moment-generating function term, and the number of training samples.

PAC-Bayes (Probably Approximately Correct)

Low error probability: $P_S\left(d(\mathbb{E}_{P_{W|S}}\hat{L}(W,S), \mathbb{E}_{P_{W|S}}L(W)) \geq \frac{D(P_{W|S}||Q_W)+I(n,d)+\log\frac{1}{\delta}}{n}\right) \leq \delta$

MAC-Bayes (Mean Approximately Correct)

$$\mathbb{E}_{P_S} d(\mathbb{E}_{P_{W|S}}\hat{L}(W,Z), \mathbb{E}_{P_{W|S}}L(W)) \leq \frac{\mathbb{E}_{P_S} D(P_{W|S}||Q_W)+I'(n,d)}{n}$$

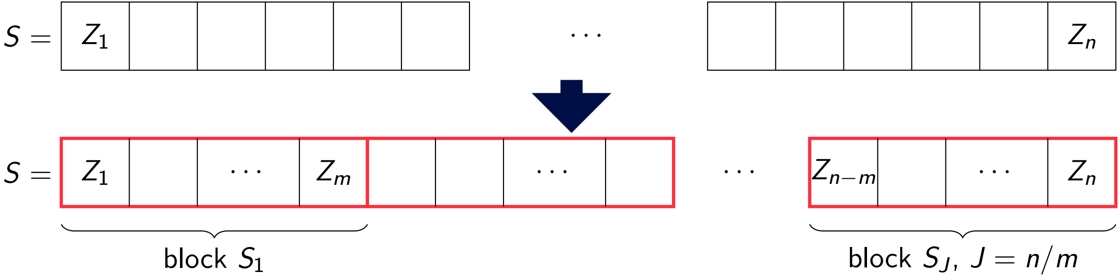
MAC-Bayes: **expectation** only, but can be tighter.

Idea: Dividing the Training Sample into Blocks

Learning algorithm: $P_{W|S}$, 'Prior': $Q_W \Rightarrow$ Divergence term: $D\left(P_{W|S} || Q_W\right)$

Idea: Dividing the Training Sample into Blocks

Learning algorithm: $P_{W|S}$, 'Prior': Q_W $\Rightarrow$ Divergence term: $D\left(P_{W|S}||Q_W\right)$



$P_{W|S_j} := \mathbb{E}_{P_{S_1,\ldots,S_{j-1},S_{j+1},\ldots,S_J}} P_{W|S} \Rightarrow$ Divergence terms: $D\left(P_{W|S_j}||Q_W\right)$

Results

Theorem

Generalization error bound in terms of a (generalized) comparator function in terms of loss moment-generating function term, the number of training samples, and the block size:

$$\mathbb{E}_{P_S}\, d\left(\mathbb{E}_{P_{W|S}}\hat{L}(W,S), \mathbb{E}_{P_{W|S}}L(W)\right) \leq \frac{\frac{n}{m}\log\Phi_m\left(\frac{\lambda m}{n}\right)+\sum_{j=1}^{n/m}\mathbb{E}_{P_{S_j}} D\left(P_{W|S_j}||Q_W\right)}{\lambda}$$

Corollary

Loss bounded in $[0,1]$, $C_\beta(r,s) := -\log\left(1-(1-\exp(-\beta))s\right) - \beta r$ (Catoni function). Then

$$\mathbb{E}_{P_S} C_\beta\left(\mathbb{E}_{P_{W|S}}\hat{L}(W,S), \mathbb{E}_{P_{W|S}}L(W)\right) \leq \frac{1}{n}\sum_{j=1}^{n/m}\mathbb{E}_{P_{S_j}} D\left(P_{W|S_j}||Q_W\right).$$

Consequence for generalization error:

$$\mathbb{E}_{P_{S,W}}\left(L(W)-\hat{L}(W,Z)\right) \leq \sqrt{\frac{1}{4n}\sum_{j=1}^{n/m}\mathbb{E}_{P_{S_j}} D\left(P_{W|S_j}||Q_W\right)}$$

Results

Theorem

Generalization error bound in terms of a (generalized) comparator function in terms of loss moment-generating function term, the number of training samples, and the block size:

$$\mathbb{E}_{P_S} d \left(\mathbb{E}_{P_{W|S}} \hat{L}(W,S), \mathbb{E}_{P_{W|S}} L(W) \right) \leq \frac{\frac{n}{m} \log \Phi_m\left(\frac{\lambda m}{n}\right) + \sum_{j=1}^{n/m} \mathbb{E}_{P_{S_j}} D\left(P_{W|S_j} || Q_W\right)}{\lambda}$$

Impossibility Result

Let m be such that $m/n \to 0$ as $n \to \infty$. Then there exist a loss function ℓ with values bounded in $[0,1]$, a sample distribution P_Z, and a learning algorithm $P_{W|S}$ with:

1. Theorem yields vanishing right hand side
2. If $f : [0,\infty) \to \mathbb{R}$ nondecreasing, A_n vanishing as $n \to \infty$, B_n in $o(1/f(n \log n))$, and n large enough, there exists $\delta \in [0,1]$ such that
$$P_S \left(\mathbb{E}_{P_{W|S}} \left(L(W) - \hat{L}(W,S) \right) > A_n + B_n \cdot f\left(\frac{1}{\delta}\right) \right) > \delta.$$

Example

Gaussian mean estimation with truncated loss: $Z_i \sim \mathcal{N}(\mu, 1)$ for some (unknown) $\mu \in (0,1)$, $W := \frac{1}{n}\sum_{i=1}^n Z_i$, and

$$\ell(w,z) := K\left((w-z)^2\right), \quad K(x) := \begin{cases} x, & x \in [0,1) \\ 1, & x \in [1,\infty). \end{cases}$$

- $D\left(P_{W|S} || Q_W\right) = \infty$, so original PAC-Bayes bound vacuous
- Corollary yields $\mathbb{E}_{P_{S,W}}\left(L(W) - \hat{L}(W,Z)\right) \leq \frac{1}{2}\sqrt{\frac{1}{2(n-m)}} = \mathcal{O}(1/\sqrt{n})$ for $m = 1$.

Conclusion and Future Directions

- Block-sample bounds (in general) tighter than standard PAC-Bayes bounds, but hold in expectation only
- Divergence term in bound depends on data distribution (in general unknown; shared property with other related bounds)
  $\Rightarrow$ upper bounds based on general knowledge is necessary for applications