How Tight can PAC-Bayes be in the Small Data Regime?

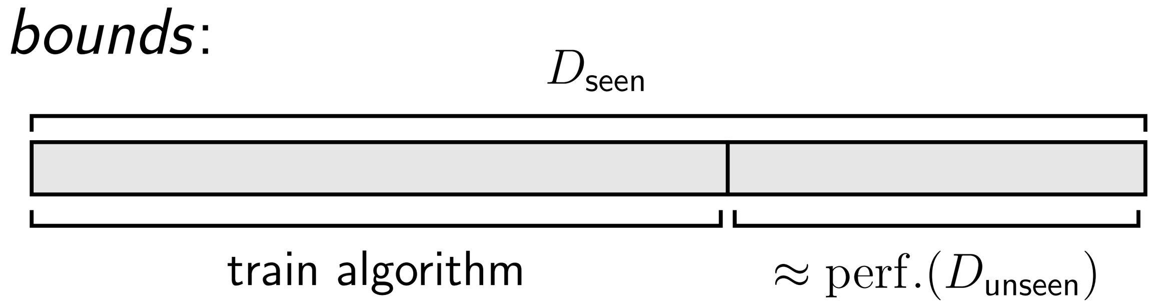
Andrew Y. K. Foong*[1], Wessel P. Bruinsma*[1,2], David R. Burt[1], Richard E. Turner[1]
[1]University of Cambridge [2]Invenia Labs {ykf21, wpb23, drb62, ret26}cam.ac.uk

- Investigate whether **PAC-Bayes** can give tighter bounds than **test set bounds**.
- **Characterise limits** of the well-known generic PAC-Bayes theorem of Germain et al.
- **Meta-learning** experiments on synthetic data to obtain tightest bounds possible.
- PAC-Bayes **tighter than some but not all** test set bounds.

Motivation and Test Set Bounds

Generalisation bounds *guarantee* and potentially *explain* generalisation.

Test set bounds:



Chernoff test set bound:

$$\Pr\left(\mathrm{kl}(R_{S_{\text{test}}}(h), R_D(h)) \leq \tfrac{1}{N_{\text{test}}}\log\tfrac{1}{\delta}\right) \geq 1-\delta$$

where $\mathrm{kl}(q,p) := q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}$.

✗ Can't explain *why* generalisation occurred.

✗ Sacrifices train data: *especially bad in small-data regime*!

Can we get tighter bounds with PAC-Bayes?

Generic PAC-Bayes Theorem

Bounds the gen. risk of *randomised* classifiers, $\overline{R}_D(Q)$.

Generic PAC-Bayes theorem (Germain et. al., 2009).

Choose a convex function $\Delta\colon [0,1]^2 \to \mathbb{R}\cup\{+\infty\}$. With probability $1-\delta$, for all posterior distributions Q,

$$\Delta(\overline{R}_S(Q), \overline{R}_D(Q)) \leq \tfrac{1}{N}\left[\mathrm{KL}(Q\|P) + \log\left(\tfrac{1}{\delta}\mathcal{I}_\Delta(N)\right)\right],$$

where $\mathcal{I}_\Delta(N) := \sup_{r\in[0,1]} \sum_{k=0}^{N} \binom{N}{k} r^k (1-r)^{N-k} e^{N\Delta(k/N,r)}$.

Obtains well-known bounds as special cases:

- PAC-Bayes-kl: $\Delta(q,p) = \mathrm{kl}(q,p) := q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}$.
- Catoni: $\Delta(q,p) = C_\beta(q,p) := -\log(1+p(e^{-\beta}-1)) - \beta q$.

Limits of Generic PAC-Bayes Theorem

- *Generic PAC-Bayes theorem bound:* $\overline{p}_\Delta$.
- *Conjectured PAC-Bayes-kl:* $\underline{p}$, modify PAC-Bayes-kl:

$$\mathrm{kl}(\overline{R}_S(Q), \overline{R}_D(Q)) \leq \tfrac{1}{N}\left[\mathrm{KL}(Q\|P) + \log\tfrac{1}{\delta} + \cancel{\log(2\sqrt{N})}\right].$$

Our contribution:

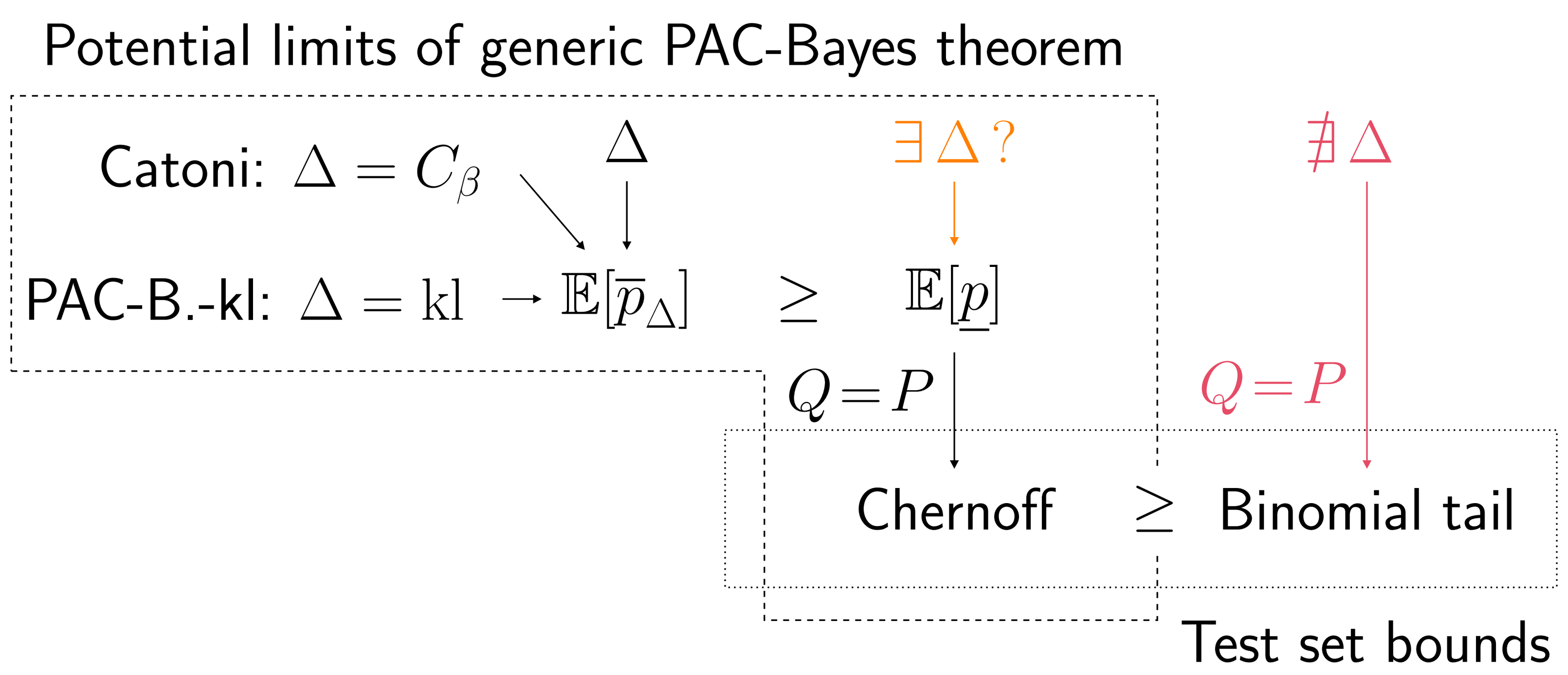
Limits of generic PAC-Bayes theorem, simplified

For any distribution over datasets, prior, and learning alg.,

$$\inf_\Delta \overline{p}_\Delta = \underline{p} \text{ a.s.} \implies \inf_\Delta \mathbb{E}[\overline{p}_\Delta] \geq \mathbb{E}[\underline{p}].$$

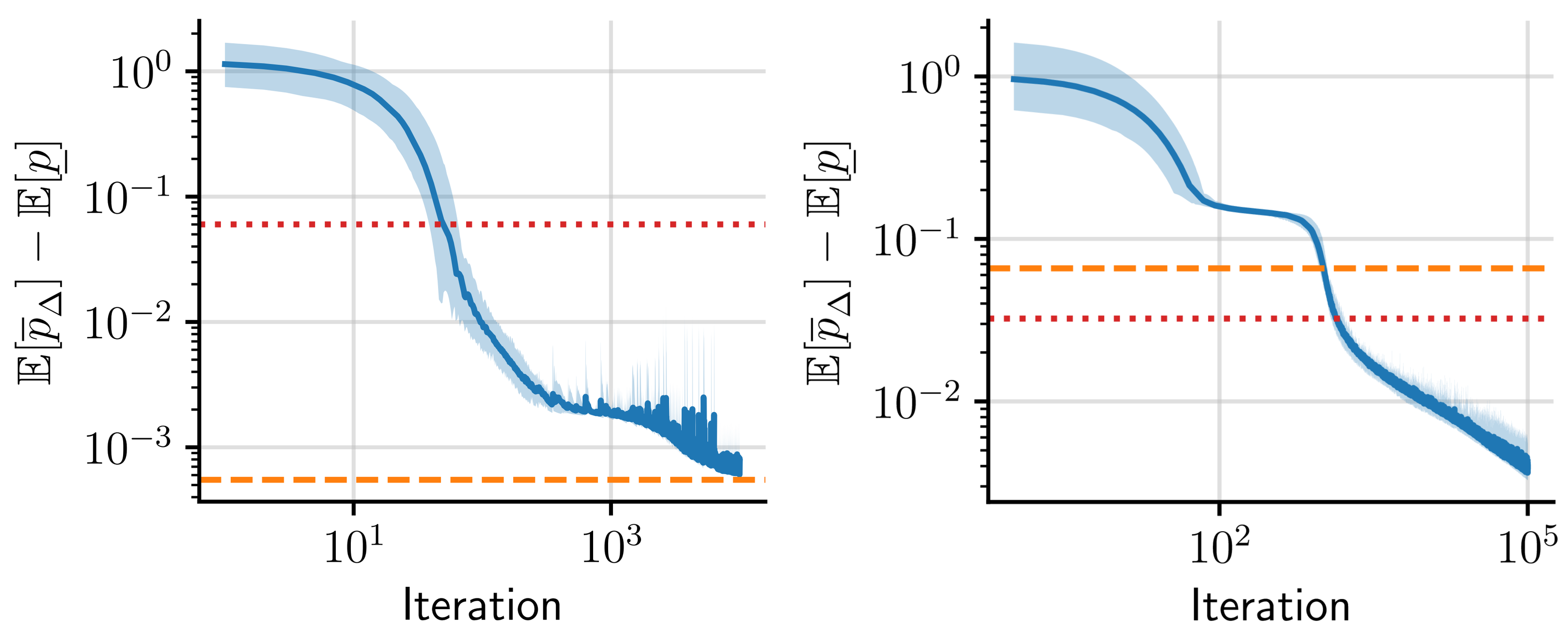
Generic PAC-Bayes bound can never be tighter than conjectured PAC-Bayes-kl.

Relationship Between Bounds



Numerical Verification

- Parameterise convex function Δ with a neural network.
- Use ADAM to optimise weights.
- Synthetic distribution over KL and empirical risk term.



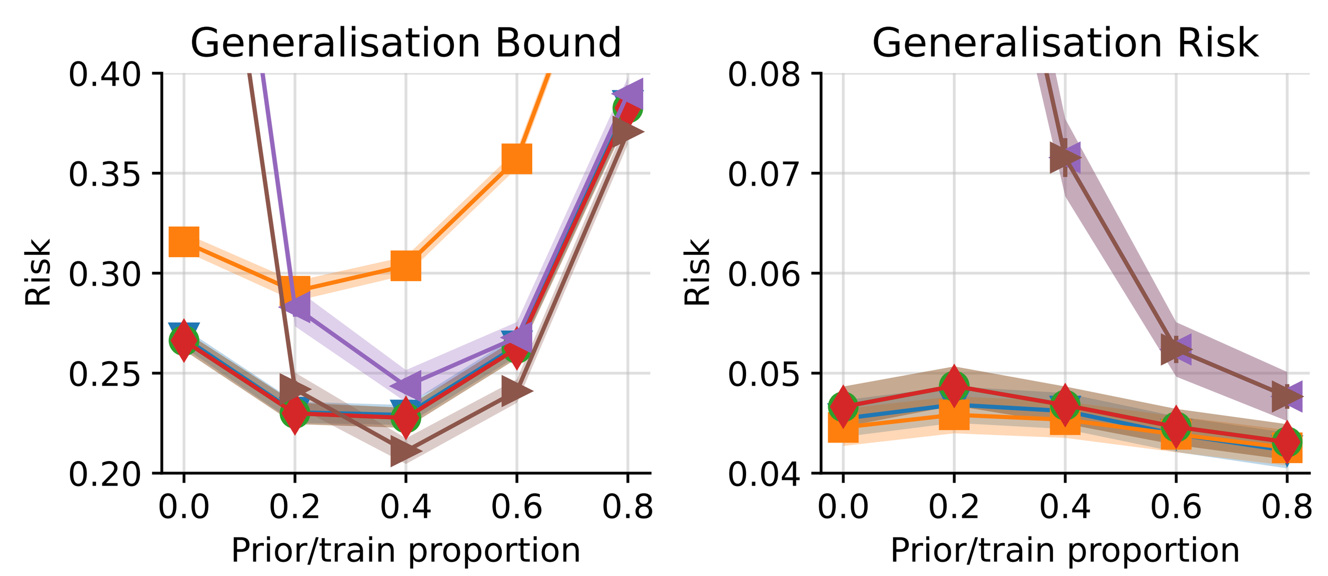
Dotted red: PAC-Bayes-kl bound.

Dashed orange: Catoni bound with optimal β.

Blue: optimised neural network Δ bound.

Empirical Comparison of Tightness

- Compare PAC-B. & test set bounds in 1D classification.
- Train neural processes to meta-learn algorithms that are *adapted to minimise each bound*.



Catoni (▼), MLS (■), conjectured PAC-B.-kl (●), learned convex (♦), Chernoff test set (◄), and binomial tail test set (►)

- PAC-Bayes is competitive with Chernoff test set bound, but looser than binomial tail test set bound.
- PAC-Bayes leads to much lower actual risk.

Check out the paper at https://arxiv.org/abs/2106.03542