PAC-Bayes, MAC-Bayes, and Conditional Mutual Information:

Fast rate bounds that handle general VC classes

Peter Grünwald, Thomas Steinke, Lydia Zakynthinou

CWI & Leiden University, Google Research (Brain Team) , Northeastern University

Generalization bounds

- Sample i.i.d dataset Z of size n from unknown distribution $\mathcal{D}$ over $\mathcal{Z}$.
- Loss function $\ell : \Delta(\mathcal{F}) \times \mathcal{Z} \to [0,1]$ indicates quality of (randomized) $f \in \mathcal{F}$.

True: $L(A|Z;\mathcal{D}) = \mathbb{E}_{f\sim A|Z,\ Z'\sim\mathcal{D}}[\ell(f;Z')]$

Empirical: $L(A|Z;Z) = \mathbb{E}_{f\sim A|Z}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(f;Z_i)\right]$

Standard PAC-Bayes/MI bounds

[McAllester 1998, 2003], [Audibert 2004], [Catoni 2007]

$$L(A|Z;\mathcal{D}) - L(A|Z;Z) \trianglelefteq_{\text{whp \& } \mathbb{E}} \sqrt{L(A|Z;Z) \cdot \frac{\mathrm{KL}(A|Z \parallel \pi)}{n}}$$

independent of Z

[Russo and Zhou 2016], [Xu and Raginsky 2017]

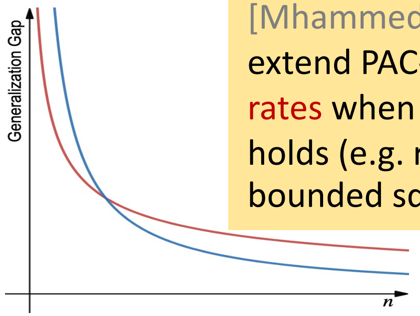
$$\mathbb{E}_Z[L(A|Z;\mathcal{D}) - L(A|Z;Z)] \leq \sqrt{\frac{2 \cdot I(A|Z\,;Z)}{n}}$$

Directions of improvement over standard

1. Do not capture fast rates: $\sim \sqrt{\frac{\text{COMPLEXITY}}{n}}$

   Rewriting PAC-Bayes excess risk bounds:

   $R(A|Z;Z) + \left(\frac{\mathrm{KL}(A|Z \parallel \pi)}{n}\right)^{\gamma}$, where $\gamma \in \left[\frac{1}{2}, 1\right]$.



[Mhammedi, Grünwald, Guedj 2019] extend PAC-Bayes to capture fast rates when a **Bernstein condition** holds (e.g. random label noise, bounded squared error loss).

2. Do not handle general VC classes: bound can be infinite for cases where Uniform Convergence implies generalization [Bassily, Moran, Nachum, Shafer, Yehudayoff 2018], [Livni and Moran 2020]

Conditional Mutual Information:
[Steinke and Zakynthinou 2020] extend MI to handle general VC classes, proposing $CMI_{\mathcal{D}}(A)$. Subsequently [Hellström and Durisi 2020] extend to PAC-Bayes.

Conditional, faster rate PAC-Bayes/MI bound

Theorem. If a γ-Bernstein condition holds, for arbitrary *almost exchangeable data-dependent priors* $\pi|\langle Z_0, Z_1\rangle$

$$L(A(Z_0);\mathcal{D}) - L(A(Z_0);Z_0) \trianglelefteq \left(2-\frac{1}{\gamma}\right)\cdot R(A(Z_0);Z_0) + \left(\frac{\mathbb{E}_{Z_1}[\mathrm{KL}(A(Z_0) \parallel \pi|\langle Z_0, Z_1\rangle)]}{n}\right)^{\gamma}$$

Real dataset $Z_0 \sim \mathcal{D}^n$

Ghost dataset $Z_1 \sim \mathcal{D}^n$

$\langle Z_0, Z_1\rangle = \{Z_{1,0}, Z_{1,1}\}, \{Z_{2,0}, Z_{2,1}\}, \ldots, \{Z_{n,0}, Z_{n,1}\}$

Claim (VC+New bound). For any class $\mathcal{F}$ with VCdim=d, $\exists A$ (ERM with a consistency property) and prior π such that for any $\mathcal{D}$, $\mathrm{KL}(A|Z_0 \parallel \pi|\langle Z_0, Z_1\rangle) \leq d \log 2n$

Main Technical Lemma. Let $S \sim \mathrm{Ber}(1/2)$, $\bar{S} = 1 - S$, and $|r_0|, |r_1| \leq 1$. Then for all $\eta < 1/4$,

$$r_{\bar{S}} - r_S \trianglelefteq C \cdot \eta \cdot r_{\bar{S}}^2$$

Future directions

- Extend to unbounded (e.g. subgaussian) losses.
- Extend to *observable* bound (now might need to know $\gamma, f^*, \mathcal{D}$)

LZ was supported by a Facebook Fellowship, PG by the Dutch Research Council (NWO) via research programme 617.001.651. This work began when TS and LZ were at IBM Research in Almaden.