ICML2023

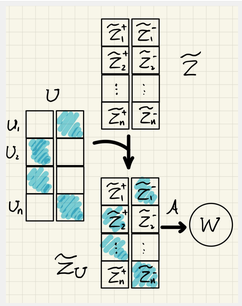
for updated paper, scan

Tighter Information-Theoretic Generalization Bounds from Supersamples

Ziqiao Wang and Yongyi Mao

University of Ottawa

supersample and CMI bound (Steinke-Zakynthinou, 2020)

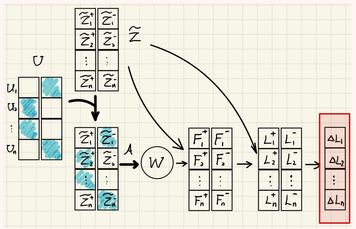


- Sample $2n$ instances ((X, Y) pairs) to fill matrix $\widetilde{Z}$
- Use a random binary mask U_i to select an instance from row i of $\widetilde{Z}$ to include in the training set $\widetilde{Z}_U$: $U_i = 0$ selects the left one.
- Apply learning algorithm $\mathcal{A}$ to $\widetilde{Z}_U$ to obtain model parameter $W = \mathcal{A}(\widetilde{Z}_U)$.
- For bounded loss in $[0, 1]$, the generalization error (expected over W and $\widetilde{Z}_U$) is bounded by

$$|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2I(W;U_i|\widetilde{Z})}$$

This paper: tightest bounds and novel perspectives for this setting

from CMI, f-CMI, e-CMI, to loss-difference CMI



- $F_i^+ := f_W(\widetilde{X}_i^+)$, $F_i^- := f_W(\widetilde{X}_i^-)$, $F_i := (F_i^+, F_i^-)$
  $\Rightarrow$ f-CMI Bound: $|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{I(F_i;U_i|\widetilde{Z})}$ (Harutyunyan et al.,2021)
- $L_i^+ := \ell(W, \widetilde{Z}_i^+)$, $L_i^- := \ell(W, \widetilde{Z}_i^-)$, $L_i := (L_i^+, L_i^-)$
  $\Rightarrow$ e-CMI Bound: $|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{I(L_i;U_i|\widetilde{Z})}$ (Hellström-Durisi,2022)
- This paper: $\Delta L_i := L_i^- - L_i^+$

Theorem A: ld-CMI Bound

$$|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{I(\Delta L_i;U_i|\widetilde{Z})}$$

ld-CMI bound is tighter than e-CMI, f-CMI and CMI bounds

Noting the Markov Chain (conditioned on $\widetilde{Z}$): $U_i - W - F_i - L_i - \Delta L_i$ we see

$$I(W;U_i|\widetilde{Z}) \geq I(F_i;U_i|\widetilde{Z}) \geq I(L_i;U_i|\widetilde{Z}) \geq I(\Delta L_i;U_i|\widetilde{Z})$$

ld-CMI Benefits from reducing loss-pairs to their equivalence classes.

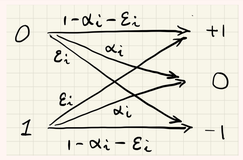
an even tighter bound: loss-difference MI bound

Theorem B: ld-MI bound

$$|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{I(\Delta L_i;U_i)}$$

Since $U_i \perp \widetilde{Z}$, $I(\Delta_i;U) \leq I(\Delta L_i;U) + I(U_i;\widetilde{Z}|\Delta L_i) = I(\Delta L_i;U|\widetilde{Z})$

loss-difference MI bound chracterizes generalization error for interpolation algorithms under 0/1-loss



- $I(\Delta L_i;U_i)$ measures the rate of communication for channel $U_i \to \Delta L_i$.
- Consider 0/1-loss. Then $\Delta L_i \in \{-1, 0, +1\}$.
- α_i: $\mathbb{P}\{\text{trainLoss} = \text{testLoss in row } i \text{ of } \widetilde{Z}\}$
- ϵ_i: $\mathbb{P}\{\text{trainLoss} > \text{testLoss in row } i \text{ of } \widetilde{Z}\}$

- Consider an interpolation algorithm: $\epsilon_i = 0$. $\Rightarrow$ Binary Erasure Channel.

We see $I(\Delta L_i;U_i) = (1-\alpha_i)\log 2\text{nats}$ (= channel capacity), and

Theorem C: MI Characterization of Err for Interpolation Algorithm

$$|\mathrm{Err}| = \sum_{i=1}^{n}\frac{I(\Delta L_i;U_i)}{n\log 2} = 1 - \textstyle\sum_{i=1}^{n}\alpha_i/n$$

weighted generalization error from correlating with shifted Radermacher sequence

- R: population risk expected over W.
- $\widehat{R}_n$: empirical risk expected over W and training set of size n
- For any $\rho > 0$, define ρ-weighted generalization error (Catoni, 2007) as

$$\mathrm{Err}_\rho := R - (1+\rho)\widehat{R}_n$$

and define ρ-shifted Rademacher random variable

$$\tilde{\epsilon}_i := (-1)^{1-U_i} - \frac{\rho}{\rho+2}$$

Symmetry Lemma

$$\mathrm{Err}_\rho = \tfrac{\rho+2}{n}\sum_{i=1}^{n}\mathbb{E}_{L_i^+,\tilde{\epsilon}_i}\tilde{\epsilon}_i L_i^+$$

Such a perspective allows the development of fast-rate bounds.

fast-rate MI bounds

Theorem D: Fast-Rate MI Bounds

1. There exist $\rho, \delta > 0$ such that $\mathrm{Err}_\rho \leq \sum_{i=1}^{n}\frac{I(L_i^+;U_i)}{\delta n}$
2. $R \leq \widehat{R}_n + \sum_{i=1}^{n}\frac{4I(L_i^+;U_i)}{n} + 4\sqrt{\sum_{i=1}^{n}\frac{\widehat{R}_n I(L_i^+;U_i)}{n}}$.
3. For interpolation algorithms, $\mathrm{Err} \leq \sum_{i=1}^{n}\frac{2I(L_i^+;U_i)}{n\ln 2}$.

variance-based fast-rate MI bound

- $L_S(W)$: emprical risk of W on training sample S
- For $\gamma \in (0, 1)$, γ-variance is defined as

$$V(\gamma) := \mathbb{E}_{W,S}\frac{1}{n}\sum_{i=1}^{n}\left(\ell(W, Z_i) - (1+\gamma)L_S(W)\right)^2.$$

Theorem E: Variance-Based MI Bound for 0/1-Loss

There exist $\rho, \delta > 0$ such that $\mathrm{Err} \leq \rho V(\gamma) + \sum_{i=1}^{n}\frac{I(L_i^+;U_i)}{n\delta}$.

sharpness-based fast-rate MI bounds

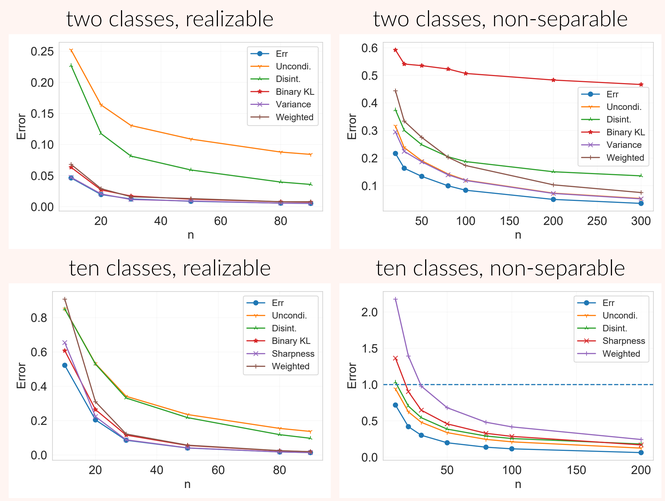
For any $\lambda \in (0, 1)$,

- $F_i(\lambda) \triangleq \mathbb{E}_{W,Z_i}\ell(W, Z_i) - (1+\lambda)\mathbb{E}_{W|Z_i}\ell(W, Z_i)^2$.
- λ-sharpness is defined as $F(\lambda) = \frac{1}{n}\sum_{i=1}^{n}F_i(\lambda)$

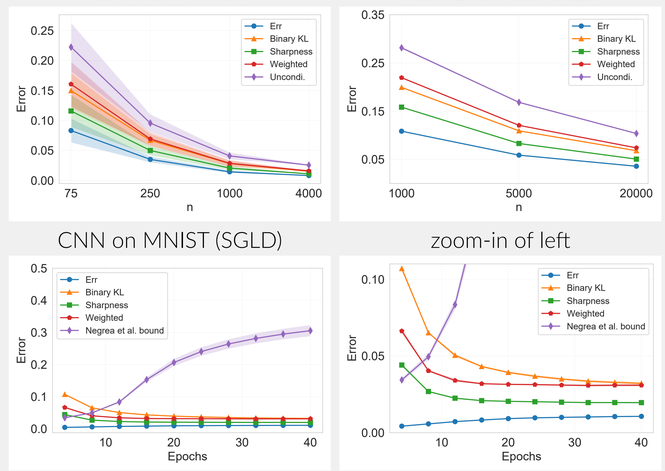
Theorem F: Sharpness-Based MI Bound for 0/1-Loss

There exist $\rho, \delta > 0$ such that $\mathrm{Err} \leq \rho F(\lambda) + \sum_{i=1}^{n}\frac{I(L_i^+;U_i)}{\delta n}$

experimental comparison: linear classifier on synthetic data



experimental comparison: neural networks on real data



zwang286@uottawa.ca, ymao@uottawa.ca