Sample-Conditioned Hypothesis Stability Sharpens Information-Theoretic Generalization Bounds

Ziqiao Wang [1] Yongyi Mao [1]

[1]University of Ottawa

Background

- Learning algorithm $\mathcal{A}: S \to W$ i.e. mapping a training sample to a hypothesis.
- Gen. Err. $= \mathbb{E}$ [Test Err. $-$ Train Err.] $\leq$ Gen. Bound.
- Information-theoretic (IT) bounds belong to the class of Gen. Bound.

Limitations of IT Generalization bounds

- Original input-output mutual information (IOMI) (e.g., $I(W;S)$ in [5]) based bound can $\to \infty$.
  $\Longrightarrow$ solved by conditional mutual information (CMI) $I(W;U|\widetilde{Z})$ in [4].
- Slow convergence rate, e.g., $\mathcal{O}(1/\sqrt{n})$ $\Longrightarrow$ mitigated by [3, 6] and so on.
- Non-vanishing in Stochastic Convex Optimization (SCO) problems [2]!

Contributions

Our contribution: Incorporating stability-based analysis into IT framework which improves both stability-based bounds and IT bounds.

Key Observation from Algorithmic Stability

- Given $S = \{Z_i\}_{i=1}^n$ and Z_i':

$$Z_1, \ldots, \boxed{Z_i}, \ldots, Z_n \xrightarrow{\mathcal{A}} \boxed{W} \Rightarrow \text{Loss of } (W, Z)$$

$$Z_1, \ldots, \boxed{Z_i'}, \ldots, Z_n \xrightarrow{\mathcal{A}} \boxed{W^{-i}} \Rightarrow \text{Loss of } (W^{-i}, Z)$$

- $\mathcal{A}$ is Stable $\iff$ Loss of (W^{-i}, Z) is close to Loss of (W, Z).
  - Uniform Stability [1]:
    $\sup_{W,W^{-i},Z} |\text{Loss of } (W,Z) - \text{Loss of } (W^{-i},Z)| \leq$ Unif. Stability Param.
  - Sample-Conditioned Hypothesis (SCH) Stability in this paper
    $\mathbb{E}_{W,W^{-i}}\left[\sup_Z |\text{Loss of } (W,Z) - \text{Loss of } (W^{-i},Z)|\right] \leq$ SCH Stability Param.,
    where Z can be either Z_i or Z_i'.

$\Downarrow$

Some terminologies

- Evaluated Data $Z \in (Z_i, Z_i')$;
- (Neighboring) Hypothesis pair: (W, W^{-i})
- Membership: e.g. $\mathbb{1}\{\text{Evaluated Data} = Z_i\}$

$\Downarrow$

Main Theorem (informal.)

If $\mathcal{A}$ is stable, then

$$\text{Gen. Err.} \precsim \text{Stability Param} \times \underbrace{\sqrt{I\left(\text{Evaluated Data; Membership}|\text{Hypothesis Pair}\right)}}_{\text{New CMI}}$$

$\Downarrow$

Generalization, in this context, pertains to the ability to infer, given (W, W^{-i}) and Evaluated Data, whether the Evaluated Data corresponds to Z_i or Z_i'.

Novel Construction: "Neighboring-Hypothesis" Matrix

- Original "supersample" matrix construction in CMI [4]:

$$\text{Supersample}=\begin{bmatrix} \text{Data}_{1,0} & \text{Data}_{1,1} \\ \text{Data}_{2,0} & \text{Data}_{2,1} \\ \vdots & \vdots \\ \text{Data}_{n,0} & \text{Data}_{n,1} \end{bmatrix} \xRightarrow{\text{Mask}(U)} \begin{bmatrix} \text{Data}_{1,0} & \blacksquare \\ \blacksquare & \text{Data}_{2,1} \\ \vdots & \vdots \\ \text{Data}_{n,0} & \blacksquare \end{bmatrix} \xRightarrow{\text{Train. Set}} \begin{bmatrix} \text{Data}_{1,0} \\ \text{Data}_{2,1} \\ \vdots \\ \text{Data}_{n,0} \end{bmatrix} \xrightarrow{\mathcal{A}} W$$

$\xRightarrow{\text{CMI}}$ $I(W;U|\text{Supersample})$: membership inference of train. set.

- Our "neighboring-hypothesis" matrix construction:

$$\begin{bmatrix} \text{Data}_{1,0} & \blacksquare \\ \text{Data}_{2,0} & \blacksquare \\ \vdots & \vdots \\ \text{Data}_{n,0} & \blacksquare \end{bmatrix} \xrightarrow{\mathcal{A}} W, \quad \begin{bmatrix} \blacksquare & \text{Data}_{1,1} \\ \text{Data}_{2,0} & \blacksquare \\ \vdots & \vdots \\ \text{Data}_{n,0} & \blacksquare \end{bmatrix} \xrightarrow{\mathcal{A}} \underbrace{W^{-1}, \cdots, W^{-n}}_{n \text{ new neighboring hypotheses}} \Longrightarrow \begin{bmatrix} W & W^{-1} \\ W & W^{-2} \\ \vdots & \vdots \\ W & W^{-n} \end{bmatrix}$$

- The generalization game:

$$\begin{bmatrix} \text{Data}_{1,0} & \text{Data}_{1,1} \\ \text{Data}_{2,0} & \text{Data}_{2,1} \\ \vdots & \vdots \\ \text{Data}_{n,0} & \text{Data}_{n,1} \end{bmatrix} + \begin{bmatrix} \square & \blacksquare \\ \blacksquare & \square \\ \vdots & \vdots \\ \square & \blacksquare \end{bmatrix} \Longrightarrow \begin{bmatrix} \text{Data}_{1,0} & \blacksquare \\ \blacksquare & \text{Data}_{2,1} \\ \vdots & \vdots \\ \text{Data}_{n,0} & \blacksquare \end{bmatrix} + \begin{bmatrix} W & W^{-1} \\ W & W^{-2} \\ \vdots & \vdots \\ W & W^{-n} \end{bmatrix} \Longrightarrow \begin{bmatrix} \text{Tr. Err}_1 & \text{Ts. Err}_1 \\ \text{Ts. Err}_2 & \text{Tr. Err}_2 \\ \vdots & \vdots \\ \text{Ts. Err}_n & \text{Tr. Err}_n \end{bmatrix}$$

$\Longrightarrow$ Gen. Err.$=\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\text{Ts. Err}_i - \text{Tr. Err}_i\right]$ ($\leq$ Stability Param.)

$\Downarrow$

- Our main theorem is asking:

$$\text{Given} \begin{bmatrix} W & W^{-1} \\ W & W^{-2} \\ \vdots & \vdots \\ W & W^{-n} \end{bmatrix} \text{and} \begin{bmatrix} \text{Data}_{1,0} \\ \text{Data}_{2,1} \\ \vdots \\ \text{Data}_{n,0} \end{bmatrix}, \text{can you infer} \begin{bmatrix} \square & \blacksquare \\ \blacksquare & \square \\ \vdots & \vdots \\ \square & \blacksquare \end{bmatrix}?$$

More Technical? All about Bounding CGF

Recall Donsker-Varadhan (DV) lemma:

$$\text{Gen. Err.} \leq \inf_{t>0} \frac{\text{IOMI or CMI} + \text{CGF}}{t}.$$

Let f_{DV} be so-called DV auxiliary function, then

$$\text{CGF} = \log \mathbb{E}\left[\exp\left(t \cdot f_{\text{DV}}\right)\right] \leq \text{Some Concentration Bound}.$$

- Typical choices of f_{DV} in previous works:

$$f_{\text{DV}} = \begin{cases} \text{Single Loss of } (w, z') \\ \text{Loss of } (w, z') - \text{Expected Loss of } (w, Z') \\ \text{Loss of Data 1} - \text{Loss of Data 2 for the same } w \end{cases},$$

where in the last choose, Data 1 is chosen uniformly from a data pair, e.g., (Z^0, Z^1), decided by a $U \sim \text{Bern}(1/2) \Longrightarrow$ Data 1 $= Z^U$, Data 2$=Z^{1-U}$.

- In this paper:

$$f_{\text{DV}} = \begin{cases} \text{Loss of } (w, z') - \text{Conditional Expected Loss of } (W^{-i}, z') \\ \text{Loss of Hypothesis 1} - \text{Loss of Hypothesis 2 for the same } z \end{cases},$$

where in the last choose, Hypothesis 1 is chosen uniformly from a neighboring hypothesis pair, e.g., (W^0, W^1), decided by a $U \sim \text{Bern}(1/2)$ $\Longrightarrow$ Hypothesis 1 $= W^U$, Hypothesis 2$=W^{1-U}$.

Application: Stochastic Covex Optimization Problems

SCO setting: Hypothesis set is convex; Objective function is convex.

- In convex-Lipschitz-bounded (CLB) counterexamples (which is a subset of SCO problems) given by [2]:

$$\text{Gen. Err.} \leq \mathcal{O}(1/\sqrt{n}).$$

- Previous IOMI or CMI bound: $\mathcal{O}\left(\alpha\sqrt{\frac{\text{IOMI or CMI}}{n}}\right)$, where α usually satisfies that CGF $\leq \frac{t^2\alpha^2}{2}$.
  e.g., α can be a SubGaussian variance proxy or

$$\alpha = \sup_{\text{Hypothesis, Data Pair}} |\text{Loss of Data 1} - \text{Loss of Data 2}|.$$

- [2] shows that
  $\alpha = \mathcal{O}(1)$ (=Lip. Param.$\times$Diam. of Hypothesis Domain)
  and Previous IOMI$\geq$ Previous CMI$= \mathcal{O}(n)$.
  $\Longrightarrow \mathcal{O}\left(\alpha\sqrt{\frac{\text{IOMI or CMI}}{n}}\right) \in \mathcal{O}(1) \Longrightarrow$ Fail to explain the learnability.
- Our new CMI bound:
  Stability Param. $= \mathcal{O}(1/\sqrt{n})$
  and New CMI$= \mathcal{O}(1)$.
  $\Longrightarrow$ New CMI Bound $\in \mathcal{O}(1/\sqrt{n})$ $\Longrightarrow$Can explain the learnability!
- Wait, Stability Param. itself can serve as a generalization bound, why do we need IOMI or CMI?
  There is another CLB example in our paper where Stability Param. $= \mathcal{O}(1/\sqrt{n})$ but Gen. Err. $\leq$ New CMI Bound $= \mathcal{O}(1/n)$ Check it!

Concluding Remarks

1. Take Home Messages: Selecting the Suitable DV Auxiliary Function for Varied Problem Contexts.
2. There are additional choices for SCH stability, allowing us to establish connections with the Bernstein condition or achieve faster-rate bounds in certain cases.
3. Our new CMI maintains the same expressiveness as the original CMI and preserves its boundedness property. The comparison between the new CMI and the original CMI in a broader context remains an open question.

References

[1] Olivier Bousquet and André Elisseeff. Stability and generalization. The Journal of Machine Learning Research, 2: 499–526, 2002.

[2] Mahdi Haghifam, Borja Rodríguez-Gálvez, Ragnar Thobaben, Mikael Skoglund, Daniel M Roy, and Gintare Karolina Dziugaite. Limitations of information-theoretic generalization bounds for gradient descent methods in stochastic convex optimization. In International Conference on Algorithmic Learning Theory, pages 663–706. PMLR, 2023.

[3] Fredrik Hellström and Giuseppe Durisi. Fast-rate loss bounds via conditional information measures with applications to neural networks. In 2021 IEEE International Symposium on Information Theory (ISIT), 2021.

[4] Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In Conference on Learning Theory. PMLR, 2020.

[5] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. Advances in Neural Information Processing Systems, 2017.

[6] Ruida Zhou, Chao Tian, and Tie Liu. Exactly tight information-theoretic generalization error bound for the quadratic gaussian problem. 2023 IEEE International Symposium on Information Theory (ISIT), 2023.