

Background

- Learning algorithm $\mathcal{A} : S \rightarrow W$ i.e. mapping a training sample to a hypothesis.
- Gen. Err. = $\mathbb{E} [\text{Test Err.} - \text{Train Err.}] \leq \text{Gen. Bound.}$
- 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 $\rightarrow \infty$ 😞.
 \Rightarrow solved by conditional mutual information (CMI) $I(W; U | \tilde{Z})$ in [4] 😊.
- Slow convergence rate, e.g., $\mathcal{O}(1/\sqrt{n})$ 😞 \Rightarrow 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 :
 $Z_1, \dots, Z_i, \dots, Z_n \xrightarrow{\mathcal{A}} W \Rightarrow \text{Loss of } (W, Z)$
 $Z_1, \dots, Z'_i, \dots, Z_n \xrightarrow{\mathcal{A}} W^{-i} \Rightarrow \text{Loss of } (W^{-i}, Z)$
- \mathcal{A} is Stable \iff Loss of (W^{-i}, Z) is close to Loss of (W, Z) .
 - Uniform Stability [1]:
 $\sup_{W, W^{-i}, Z} |\text{Loss of } (W, Z) - \text{Loss of } (W^{-i}, Z)| \leq \text{Unif. Stability Param.}$
 - Sample-Conditioned Hypothesis (SCH) Stability in this paper
 $\mathbb{E}_{W, W^{-i}} [\sup_Z |\text{Loss of } (W, Z) - \text{Loss of } (W^{-i}, Z)|] \leq \text{SCH Stability Param.,}$
 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

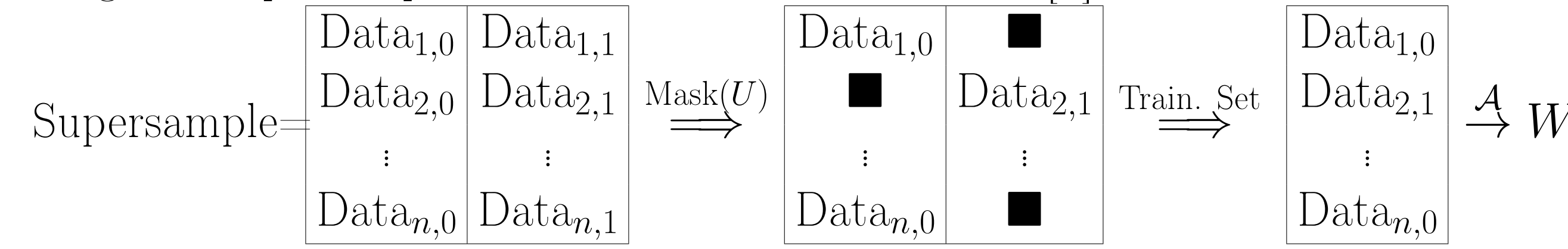
$$\text{Gen. Err.} \lesssim \text{Stability Param.} \times \underbrace{\sqrt{I(\text{Evaluated Data; Membership} | \text{Hypothesis Pair})}}_{\text{New CMI}}$$

\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 .

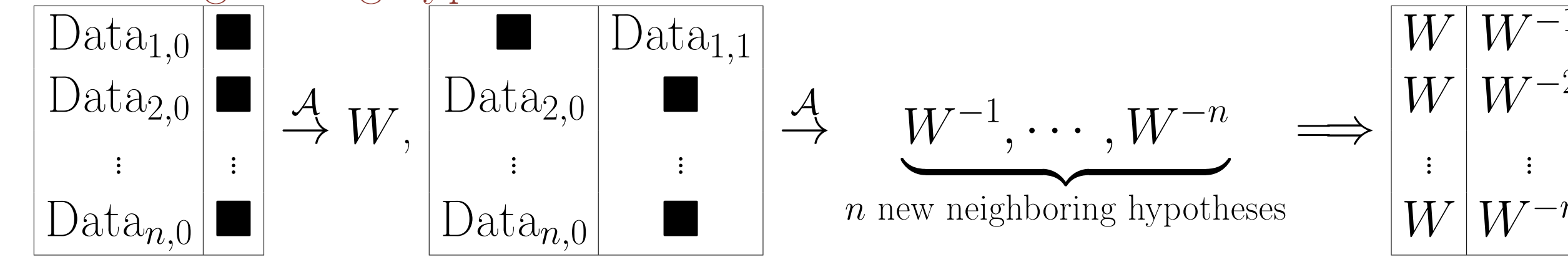
Novel Construction: “Neighboring-Hypothesis” Matrix

- Original “supersample” matrix construction in CMI [4]:

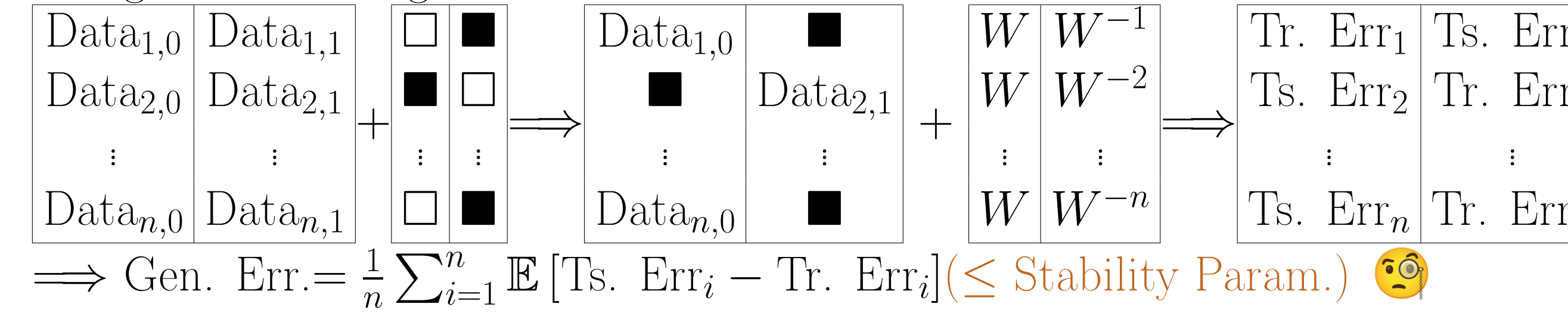


$\xRightarrow{\text{CMI}} I(W; U | \text{Supersample})$: membership inference of train. set.

- Our “neighboring-hypothesis” matrix construction:



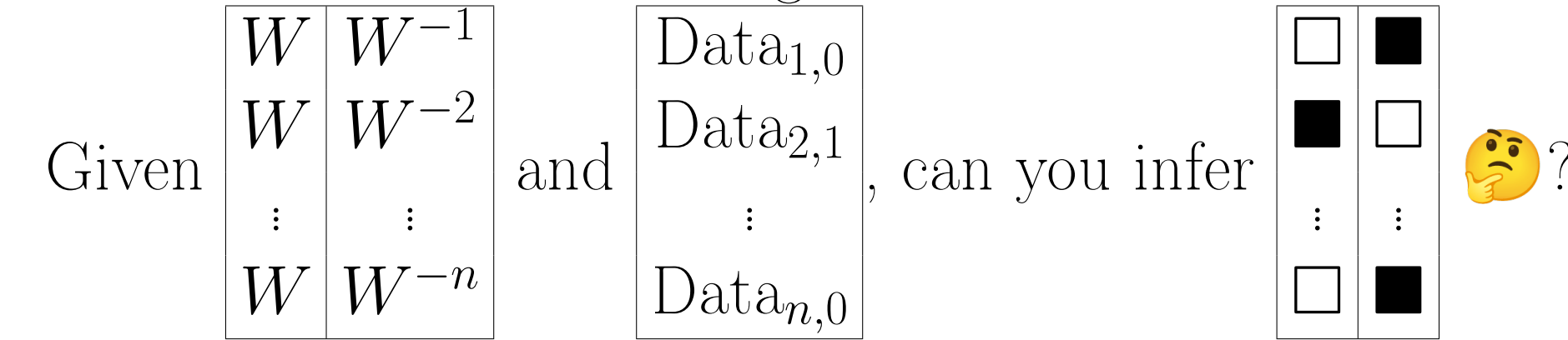
- The generalization game:



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

\Downarrow

- Our main theorem is asking:



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

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

- 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 $U \sim \text{Bern}(1/2) \Rightarrow \text{Data 1} = Z^U, \text{Data 2} = Z^{1-U}$.

- 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 $U \sim \text{Bern}(1/2) \Rightarrow \text{Hypothesis 1} = W^U, \text{Hypothesis 2} = W^{1-U}$.

Application: Stochastic Convex 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

$$\text{that CGF} \leq \frac{t^2 \alpha^2}{2}.$$

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

$$\alpha = \mathcal{O}(1) (= \text{Lip. Param.} \times \text{Diam. of Hypothesis Domain})$$

and Previous IOMI \geq Previous CMI = $\mathcal{O}(n)$.

$$\Rightarrow \mathcal{O}\left(\alpha \sqrt{\frac{\text{IOMI or CMI}}{n}}\right) \in \mathcal{O}(1) \Rightarrow \text{Fail to explain the learnability} \text{ 🤖}.$$

- Our new CMI bound:

$$\text{Stability Param.} = \mathcal{O}(1/\sqrt{n})$$

and New CMI = $\mathcal{O}(1)$.

$$\Rightarrow \text{New CMI Bound} \in \mathcal{O}(1/\sqrt{n}) \Rightarrow \text{Can explain the learnability} \text{ 😊!}$$

- 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

- Take Home Messages: Selecting the Suitable DV Auxiliary Function for Varied Problem Contexts.
- There are additional choices for SCH stability, allowing us to establish connections with the Bernstein condition or achieve faster-rate bounds in certain cases.
- 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

- Olivier Bousquet and André Elisseeff. Stability and generalization. The Journal of Machine Learning Research, 2: 499–526, 2002.
- 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.
- 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.
- Thomas Steinke and Lydia Zakyntinou. Reasoning about generalization via conditional mutual information. In Conference on Learning Theory. PMLR, 2020.
- Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. Advances in Neural Information Processing Systems, 2017.
- 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.