Appendices

A Some Useful Lemmas

In this paper, there are some equivalent forms of the generalization error we will study, e.g., Eq. (2) and Eq. (5) in the main text, which are presented in the following lemma.

Lemma A.1. *Let $W_i = \widetilde{W}_{i,U_i}$ and $\overline{W}_i = \widetilde{W}_{i,\overline{U}_i}$. For any learning algorithm $\mathcal{A}$, the following equations hold*

$$\begin{aligned}\mathcal{E}_\mu(\mathcal{A}) =& \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\widetilde{Z}_i^+,\widetilde{W}_i}\left[\ell(\widetilde{W}_i^-,\widetilde{Z}_i^+) - \ell(\widetilde{W}_i^+,\widetilde{Z}_i^+)\right], && (12)\\ =& \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\widetilde{W}_i}\left[\mathbb{E}_{\widehat{Z}_i,U_i|\widetilde{W}_i}\left[(-1)^{U_i}\left(\ell(\widetilde{W}_i^-,\widehat{Z}_i) - \ell(\widetilde{W}_i^+,\widehat{Z}_i)\right)\right]\right], && (13)\\ =& \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\widetilde{Z}_i^+}\left[\mathbb{E}_{W_i,\overline{W}_i,U_i|\widetilde{Z}_i^+}\left[(-1)^{U_i}\left(\ell(\overline{W}_i,\widetilde{Z}_i^+) - \ell(W_i,\widetilde{Z}_i^+)\right)\right]\right]. && (14)\end{aligned}$$

Proof. This lemma is a consequence of Lemma 2.1, with further utilizing some symmetric properties. Recall Eq. (1) in Lemma 2.1,

$$\begin{aligned}\mathcal{E}_\mu(\mathcal{A}) =& \mathbb{E}_{\widetilde{Z}_{[n]}^+,\widetilde{Z}_{[n]}^-}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}_{\widetilde{W}_i^-|\widetilde{Z}_{[n]}^+,\widetilde{Z}_i^-}\ell(\widetilde{W}_i^-,\widetilde{Z}_i^+) - \mathbb{E}_{\widetilde{W}^+|\widetilde{Z}_{[n]}^+}\ell(\widetilde{W}^+,\widetilde{Z}_i^+)\right]\right],\\ =& \mathbb{E}_{\widetilde{Z}_{[n]}^+,\widetilde{W}}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\ell(\widetilde{W}_i^-,\widetilde{Z}_i^+) - \ell(\widetilde{W}^+,\widetilde{Z}_i^+)\right]\right],\\ =& \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\widetilde{Z}_i^+,\widetilde{W}_i}\left[\ell(\widetilde{W}_i^-,\widetilde{Z}_i^+) - \ell(\widetilde{W}^+,\widetilde{Z}_i^+)\right].\end{aligned}$$

Note that Eq. (2) in the main text is from the second equation above, which is used to derive individual IOMI bounds in Section 3.

Similar to the standard setting for CMI bounds, where the role of each $\widetilde{Z}_i^+$ and $\widetilde{Z}_i^-$ can be exchanged, a key observation here is that for each i, $\widetilde{W}_i^+$ and $\widetilde{W}_i^-$ can also be exchanged arbitrarily. That is to say,

$$\mathcal{E}_\mu(\mathcal{A}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\widetilde{Z}_i^-,\widetilde{W}_i}\left[\ell(\widetilde{W}_i^+,\widetilde{Z}_i^-) - \ell(\widetilde{W}_i^-,\widetilde{Z}_i^-)\right] \tag{15}$$

also holds true. Notice that we do not change the definitions of any the random variable, e.g., $\widetilde{W}^+ = \mathcal{A}(\widetilde{Z}_{[n]}^+, R)$ and $\widetilde{W}_i^- = \mathcal{A}(\widetilde{Z}_{[n]\sim i}^+, R)$.

What differs from the standard CMI is that the roles of the whole sequences $\widetilde{Z}_{[n]}^+$ and $\widetilde{Z}_{[n]}^-$ are not exchangeable with each other. Here, when we exchange each $\widetilde{Z}_i^+$ and $\widetilde{Z}_i^-$, we need to keep the other positions in S unchanged.

By introducing $U_i \sim \mathrm{Unif}(\{0,1\})$, we have

$$\begin{aligned}\mathcal{E}_\mu(\mathcal{A}) =& \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\widetilde{Z}_i^+,\widetilde{W}_i}\left[\ell(\widetilde{W}_i^-,\widetilde{Z}_i^+) - \ell(\widetilde{W}_i^+,\widetilde{Z}_i^+)\right],\\ =& \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\widetilde{Z}_i,\widetilde{W}_i,U_i}\left[\ell(\widetilde{W}_{i,\overline{U}_i},\widetilde{Z}_{i,U_i}) - \ell(\widetilde{W}_{i,U_i},\widetilde{Z}_{i,U_i})\right].\end{aligned}$$

To obtain Eq. (13), notice that $\widehat{Z}_i = \widetilde{Z}_{i,U_i}$, we have

$$\begin{aligned}\mathcal{E}_\mu(\mathcal{A}) =& \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\widehat{Z}_i,\widetilde{W}_i,U_i}\left[\ell(\widetilde{W}_{i,\overline{U}_i},\widehat{Z}_i) - \ell(\widetilde{W}_{i,U_i},\widehat{Z}_i)\right]\\ =& \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\widehat{Z}_i,\widetilde{W}_i,U_i}\left[(-1)^{U_i}\left(\ell(\widetilde{W}_i^-,\widehat{Z}_i) - \ell(\widetilde{W}_i^+,\widehat{Z}_i)\right)\right].\end{aligned} \tag{16}$$

This, as we have already seen in Eq. (5) in the main text, is used to derive hypotheses-conditioned CMI bounds in Section 4. It's easy to see that when $U_i = 0$, Eq. (16) becomes Eq. (12), and when $U_i = 1$, we obtain Eq. (15) via Eq. (16).

To obtain Eq. (14), we let $W_i = \widetilde{W}_{i,U_i}$, $\overline{W}_i = \widetilde{W}_{i,\overline{U}_i}$, and fix $\widehat{Z}_i = \widetilde{Z}_i^+$. Similarly,

$$\mathcal{E}_\mu(\mathcal{A}) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\widetilde{Z}_i^+}\left[\mathbb{E}_{W_i,\overline{W}_i,U_i|\widetilde{Z}_i^+}\left[(-1)^{U_i}\left(\ell(\overline{W}_i,\widetilde{Z}_i^+) - \ell(W_i,\widetilde{Z}_i^+)\right)\right]\right].$$

This is used to derive supersample-conditioned CMI bounds in Section 4. It's easy to see that both $U_i = 0$ and $U_i = 1$ will give us Eq. (12). □

Like all the previous information-theoretic bounds, the following lemma is widely used in our paper.

Lemma A.2 (Donsker-Varadhan (DV) variational representation of KL divergence [44, Theorem 3.5])**.** *Let Q, P be probability measures on Θ, for any bounded measurable function $f : \Theta \to \mathbb{R}$, we have* $\mathrm{D_{KL}}(Q||P) = \sup_f \mathbb{E}_{\theta\sim Q}\left[f(\theta)\right] - \ln \mathbb{E}_{\theta\sim P}\left[\exp f(\theta)\right]$.

We also invoke some other lemmas as given below.

Lemma A.3 (Hoeffding's Lemma [26])**.** *Let $X \in [a, b]$ be a bounded random variable with mean μ. Then, for all $t \in \mathbb{R}$, we have $\mathbb{E}\left[e^{tX}\right] \leq e^{t\mu + \frac{t^2(b-a)^2}{8}}$.*

Lemma A.4 (Popoviciu's inequality [45])**.** *Let M and m be upper and lower bounds on the values of any random variable X, then* $\mathrm{Var}(X) \leq \frac{(M-m)^2}{4}$.

The following lemma is from [35, Lemma 2.8], we provide a self-contained proof.

Lemma A.5. *Let $h(x) = \frac{e^x - x - 1}{x^2}$ be the Bernstein function. If a random variable X satisfies $\mathbb{E}[X] = 0$ and $X \leq b$, then $\mathbb{E}\left[e^X\right] \leq e^{h(b)\mathbb{E}[X^2]}$.*

Proof. It's easy to verify that $h(x)$ is an increasing function for $x > 0$. Thus, $h(x) \leq h(b)$ for $x \leq b$. Then,

$$e^x = x + 1 + x^2 h(x) \leq x + 1 + x^2 h(b).$$

For the bounded random variable X with zero mean, we have

$$\mathbb{E}\left[e^X\right] \leq \mathbb{E}[X] + 1 + \mathbb{E}\left[X^2 h(b)\right] \leq e^{h(b)\mathbb{E}[X^2]}.$$

The last inequality is by $e^x \geq x + 1$. This completes the proof. □

B Further Elaborations on SCH Stability

We note that the reason we introduce four types of SCH stability in Definition 2.1 is that solely using β_2 in our bounds might be too loose, as it considers the supremum over all sources of randomness. By incorporating SCH stabilities, we aim to demonstrate that theoretically, we can achieve significantly tighter stability parameters.

The basic set up is as follows. Assume a random sample S gives rise to W. For each $Z_i \in S$, we construct S^i by replacing Z_i with another independently drawn instance; call training result W^i, the neighbor of W.

In a), γ_1-SCH-A stability measures the difference between the loss of w and the expected loss of its neighbor W^i at a worst z and the worst possible w. While in (b), γ_2-SCH-B stability measures the square of this difference, not in the worst case, but in an average case, where the average is over an independently Z' for the loss evaluation, the training sample, and the algorithm randomness. Since "average is smaller than worst", $\gamma_2 \leq \gamma_1$.

In c), we consider the difference between the loss of W and the loss of its neighbor when evaluated at the worst possible Z_i that when included in S gives rise to W. The expected value of this difference is γ_3-SCH-C stability.

In d), γ_4-SCH-D stability measures the expected squared difference between the loss of W and the loss of its neighbor when evaluated at Z_i (a member of S). For a similar "average smaller worst" reason, one expects that $\gamma_4 \leq \gamma_3$.

We expect that γ_2, γ_3, and γ_4 are all smaller than β_1. This is because in β_1, we consider the worst evaluated instance, whereas in the other cases, we take the expectation over all instances. Additionally, in Theorem 4.1, we expect that $\mathbb{E}_{\widetilde{W}_i} \Delta_1(\widetilde{W}_i)^2 \leq \beta_1^2$, this is because β_1-stability holds for all the possible s and s^i, namely it holds for all the (w, w^i) pair (that shares the same randomness) while in $\mathbb{E}_{\widetilde{W}_i} \Delta_1(\widetilde{W}_i)^2$, we take the expectation of these pairs.

We expect $\gamma_2 \leq \gamma_4$ due to the following reason: first by Jensen's inequality, we have $\mathbb{E}_{S,R,Z'}\left[\ell(W, Z') - \mathbb{E}_{W^i|W}\ell(W^i, Z')\right]^2 \leq \mathbb{E}_{W,W^i,Z'}\left[\ell(W, Z') - \ell(W^i, Z')\right]^2$, then since Z' is an independent of both W and W', Z' can be regarded as a testing point for both W and W', we could expect that the expectation of $\ell(W, Z') - \ell(W^i, Z')$ is small. While in $\mathbb{E}_{S,Z_i',R}\left[\ell(W, Z_i) - \ell(W^i, Z_i)\right]^2$, Z_i is a training point for obtaining W, so $\ell(W, Z_i)$ could be small in general, and Z_i is a testing point for W^i. Therefore, it is reasonable to expect $\mathbb{E}_{W,W^i,Z'}\left[\ell(W, Z') - \ell(W^i, Z')\right]^2 \leq \mathbb{E}_{S,Z_i',R}\left[\ell(W, Z_i) - \ell(W^i, Z_i)\right]^2$, namely $\gamma_2 \leq \gamma_4$.

As a concrete example, let ℓ be zero-one loss and assume $\mathcal{A}$ is an interpolating algorithm and and randomly makes predictions for unseen data. By Jensen's inequality, $\gamma_2^2 \leq \mathbb{E}_{W,W^i,Z'}\left[\ell(W, Z') - \ell(W^i, Z')\right]^2 = \mathbb{E}_{W,Z'}\left[\ell(W, Z')\right] - 2\mathbb{E}_{W,W^i,Z'}\left[(\ell(W, Z')\ell(W^i, Z'))\right] + \mathbb{E}_{W^i,Z'}\left[\ell(W^i, Z')\right]^2$, where we use $\ell^2 = \ell$ for zero-one loss. Since Z' is an unseen data for both W and W^i, we have $\gamma_2^2 \leq \mathbb{E}_{W^i,Z'}\left[\ell(W^i, Z')\right]^2 + \frac{1}{2} - \frac{1}{2} = \mathbb{E}_{W^i,Z'}\left[\ell(W^i, Z')\right]^2$. While in this case $\gamma_4^2 = \mathbb{E}_{W^i,Z_i}\left[\ell(W^i, Z_i)\right]^2$ so $\gamma_2 \leq \gamma_4$.

C Omitted Proofs and Additional Discussions in Section 3

C.1 Proof of Theorem 3.1

Proof. Let $g(\tilde{w}^+, \tilde{z}_i^+) = \mathbb{E}_{\widetilde{W}_i^-|\tilde{w}^+}\left[\ell(\widetilde{W}_i^-, \tilde{z}_i^+)\right] - \ell(\tilde{w}^+, \tilde{z}_i^+)$ be the average loss difference between $\tilde{w}^+$ and its neighboring hypothesis, and let $f = t \cdot g$ for $t > 0$ in Lemma A.2. Let $\widetilde{Z}_i^{+'}$ be an independent copy of $\widetilde{Z}_i^+$, then

$$\mathbb{E}_{\widetilde{W}^+,\widetilde{Z}_i^+}\left[g(\widetilde{W}^+, \widetilde{Z}_i^+)\right] \leq \inf_{t>0} \frac{I(\widetilde{W}^+; \widetilde{Z}_i^+) + \log \mathbb{E}_{\widetilde{W}^+,\widetilde{Z}_i^{+'}}\left[e^{tg(\widetilde{W}^+, \widetilde{Z}_i^{+'})}\right]}{t}. \tag{17}$$

Since $\widetilde{Z}_i^{+'}$ is independent of both $\widetilde{W}_i^-$ and $\widetilde{W}^+$, and $\widetilde{W}_i^-$ and $\widetilde{W}^+$ are identically distributed, we have

$$\mathbb{E}_{\widetilde{W}^+,\widetilde{Z}_i^{+'}}\left[g(\widetilde{W}^+, \widetilde{Z}_i^{+'})\right] = \mathbb{E}_{\widetilde{W}_i^-,\widetilde{Z}_i^{+'}}\left[\ell(\widetilde{W}_i^-, \widetilde{Z}_i^{+'})\right] - \mathbb{E}_{\widetilde{W}^+,\widetilde{Z}_i^{+'}}\left[\ell(\widetilde{W}^+, \widetilde{Z}_i^{+'})\right] = 0.$$

By the definition of γ_1-SCH-A stability,

$$\sup_{\tilde{w}^+, z}\left|\mathbb{E}_{\widetilde{W}_i^-|\tilde{w}^+}\left[\ell(\widetilde{W}_i^-, z)\right] - \ell(\tilde{w}^+, z)\right| \leq \gamma_1,$$

so $g(\widetilde{W}^+, \widetilde{Z}_i^{+'})$ is a zero-mean random variable bounded in $[-\gamma_1, \gamma_1]$. By Lemma A.3, we have

$$\begin{aligned}\mathbb{E}_{\widetilde{W}^+,\widetilde{Z}_i^+}\left[g(\widetilde{W}^+, \widetilde{Z}_i^+)\right] \leq& \inf_{t>0} \frac{I(\widetilde{W}^+; \widetilde{Z}_i^+) + \log \mathbb{E}_{\widetilde{W}^+,\widetilde{Z}_i^{+'}}\left[e^{tg(\widetilde{W}^+,\widetilde{Z}_i^{+'})}\right]}{t}\\ \leq& \inf_{t>0} \frac{I(\widetilde{W}^+; \widetilde{Z}_i^+) + \frac{t^2\gamma_1^2}{2}}{t}\\ =&\sqrt{2\gamma_1^2 I(\widetilde{W}^+; \widetilde{Z}_i^+)},\end{aligned}$$

where the last equality is obtained by optimizing the bound over t, i.e. letting $t = \sqrt{\frac{I(\widetilde{W}^+;\widetilde{Z}_i^+)}{2\gamma_1^2}}$.

Recall Eq. (12) in Lemma A.1 and applying Jensen's inequality to the absolute function, the first bound is then obtained by

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n}\sum_{i=1}^n \left|\mathbb{E}_{\widetilde{W}^+,\widetilde{Z}_i^+}\left[g(\widetilde{W}^+, \widetilde{Z}_i^+)\right]\right| \leq \frac{\gamma_1}{n}\sum_{i=1}^n \sqrt{2I(\widetilde{W}^+; \widetilde{Z}_i^+)},$$

Furthermore, by the chain rule of mutual information,

$$I(\widetilde{W}_i^-; \widetilde{Z}_i^+|\widetilde{W}^+) + I(\widetilde{W}^+; \widetilde{Z}_i^+) = I(\widetilde{W}^+; \widetilde{Z}_i^+|\widetilde{W}_i^-) + I(\widetilde{W}_i^-; \widetilde{Z}_i^+). \tag{18}$$

Notice that $I(\widetilde{W}_i^-; \widetilde{Z}_i^+) = 0$ in the RHS, we have

$$I(\widetilde{W}^+; \widetilde{Z}_i^+) \leq I(\widetilde{W}^+; \widetilde{Z}_i^+|\widetilde{W}_i^-),$$

which will give us the second bound. This concludes the proof. □

Remark C.1 (Comparison with Mutual Information Stability [46, 23])**.** *To compare with the mutual information stability $I\left(\widetilde{W}^+; \widetilde{Z}_i^+|\widetilde{Z}_{[n]\setminus i}^+\right)$, recall Eq.(18): $I(\widetilde{W}^+; \widetilde{Z}_i^+|\widetilde{W}^-) = I(\widetilde{W}_i^-; \widetilde{Z}_i^+|\widetilde{W}^+) + I(\widetilde{W}^+; \widetilde{Z}_i^+)$, and similarly we also have $I(\widetilde{W}^+; \widetilde{Z}_i^+|\widetilde{Z}_{[n]\setminus i}^+) = I(\widetilde{Z}_{[n]\setminus i}^+; \widetilde{Z}_i^+|\widetilde{W}^+) + I(\widetilde{W}^+; \widetilde{Z}_i^+)$. Thus, we only need to compare $I(\widetilde{W}_i^-; \widetilde{Z}_i^+|\widetilde{W}^+)$ and $I(\widetilde{Z}_{[n]\setminus i}^+; \widetilde{Z}_i^+|\widetilde{W}^+)$. Notice that for a deterministic $\mathcal{A}$, we have $I(\widetilde{W}_i^-; \widetilde{Z}_i^+|\widetilde{W}^+) \leq I(\widetilde{Z}_{[n]\setminus i}^+, \widetilde{Z}_i^-; \widetilde{Z}_i^+|\widetilde{W}^+)$. Since $\widetilde{Z}_i^- \perp\!\!\!\perp \left(\widetilde{W}^+, \widetilde{Z}_{[n]}^+\right)$, we further have $I(\widetilde{Z}_{[n]\setminus i}^+, \widetilde{Z}_i^-; \widetilde{Z}_i^+|\widetilde{W}^+) = I(\widetilde{Z}_{[n]\setminus i}^+; \widetilde{Z}_i^+|\widetilde{W}^+)$, which gives us the desired result:*

$$I(\widetilde{W}_i^+; \widetilde{Z}_i^+|\widetilde{W}^-) \leq I\left(\widetilde{W}^+; \widetilde{Z}_i^+|\widetilde{Z}_{[n]\setminus i}^+\right).$$

C.2 Proof of Theorem 3.2

Proof. The proof is nearly the same to the proof of Theorem 3.1, except that now the randomness of the algorithm is given for each DV auxiliary function, so the randomness of $\widetilde{W}_i$ is completely controlled by $\widetilde{Z}$.

Let $g(\tilde{w}^+, \tilde{z}_i^+, r) = \mathbb{E}_{\widetilde{W}_i^-|\tilde{w}^+,r}\left[\ell(\widetilde{W}_i^-, \tilde{z}_i^+)\right] - \ell(\tilde{w}^+, \tilde{z}_i^+)$ and let $f = t\cdot g$ for $t > 0$ in Lemma A.2. Let $\widetilde{Z}_i^{+'}$ be an independent copy of $\widetilde{Z}_i^+$, then

$$\mathbb{E}_{\widetilde{W}^+,\widetilde{Z}_i^+|r}\left[g(\widetilde{W}^+, \widetilde{Z}_i^+, r)\right] \leq \inf_{t>0} \frac{I(\widetilde{W}^+; \widetilde{Z}_i^+|R = r) + \log \mathbb{E}_{\widetilde{W}^+,\widetilde{Z}_i^{+'}|r}\left[e^{tg(\widetilde{W}^+,\widetilde{Z}_i^{+'},r)}\right]}{t}.$$

Notice that

$$\mathbb{E}_{\widetilde{W}^+,\widetilde{Z}_i^{+'}|r}\left[g(\widetilde{W}^+, \widetilde{Z}_i^{+'}, r)\right] = \mathbb{E}_{\widetilde{W}_i^-,\widetilde{Z}_i^{+'}|r}\left[\ell(\widetilde{W}_i^-, \widetilde{Z}_i^{+'})\right] - \mathbb{E}_{\widetilde{W}_i^+,\widetilde{Z}_i^{+'}|r}\left[\ell(\widetilde{W}_i^+, \widetilde{Z}_i^{+'})\right] = 0$$

still holds since $\widetilde{Z}_i^+$ and $\widetilde{Z}_i^-$ are i.i.d. drawn.

Thus, $g(\widetilde{W}^+, \widetilde{Z}_i^{+'}, r)$ is a zero-mean random variable bounded in $[-\beta_2, \beta_2]$. By Lemma A.3, the remaining part is routine:

$$\mathbb{E}_{\widetilde{W}^+,\widetilde{Z}_i^+|r}\left[g(\widetilde{W}^+, \widetilde{Z}_i^+, r)\right] \leq \sqrt{2\beta_2^2 I(\widetilde{W}^+; \widetilde{Z}_i^+|R=r)}.$$

Thus,

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n}\sum_{i=1}^n \left|\mathbb{E}_{\widetilde{W}^+,\widetilde{Z}_i^+,R}\left[g(\widetilde{W}^+, \widetilde{Z}_i^+, R)\right]\right| \leq \frac{\beta_2}{n}\sum_{i=1}^n \mathbb{E}_R\sqrt{2I^R(\widetilde{W}^+; \widetilde{Z}_i^+)},$$

This completes the proof. □

C.3 Proof of Theorem 3.3

Proof. Let $h(x) = \frac{e^x - x - 1}{x^2}$ be the Bernstein function. Similar to the proof of Theorem 3.1, we let $g(\tilde{w}^+, \tilde{z}_i^+) = \mathbb{E}_{\widetilde{W}_i^-|\tilde{w}^+}\left[\ell(\widetilde{W}_i^-, \tilde{z}_i^+)\right] - \ell(\tilde{w}^+, \tilde{z}_i^+)$. We have already known that $\mathbb{E}_{\widetilde{W}^+,\widetilde{Z}_i^{+'}}\left[g(\widetilde{W}^+, \widetilde{Z}_i^{+'})\right] = 0$ and $\left|g(\widetilde{W}^+, \widetilde{Z}_i^{+'})\right| \leq \gamma_1$. By Lemma A.5,

$$\begin{aligned}\log \mathbb{E}_{\widetilde{W}^+,\widetilde{Z}_i^{+'}}\left[e^{tg(\widetilde{W}^+,\widetilde{Z}_i^{+'})}\right] \leq & h(\gamma_1 t)t^2 \mathbb{E}_{\widetilde{W}^+,\widetilde{Z}_i^{+'}}\left[\left(\mathbb{E}_{\widetilde{W}_i^-|\widetilde{W}^+}\left[\ell(\widetilde{W}_i^-, \widetilde{Z}_i^{+'})\right] - \ell(\widetilde{W}^+, \widetilde{Z}_i^{+'})\right)^2\right] \\ \leq & h(\gamma_1 t)t^2\gamma_2^2,\end{aligned}$$

where the second inequality is by the definition of γ_2-SCH-B stability.

Plugging the above into Eq. (17),

$$\begin{aligned}\mathbb{E}_{\widetilde{W}^+,\widetilde{Z}_i^+}\left[g(\widetilde{W}^+, \widetilde{Z}_i^+)\right] \leq & \inf_{t>0}\frac{I(\widetilde{W}^+; \widetilde{Z}_i^+) + \log \mathbb{E}_{\widetilde{W}^+,\widetilde{Z}_i^{+'}}\left[e^{tg(\widetilde{W}^+,\widetilde{Z}_i^{+'})}\right]}{t} \\ \leq & \inf_{t>0}\frac{I(\widetilde{W}^+; \widetilde{Z}_i^+)}{t} + h(\gamma_1 t)t\gamma_2^2.\end{aligned}$$

Usually we have $\gamma_2^2 \leq \gamma_1^2 \leq \gamma_1$, we let $t = 1/\gamma_1$, then

$$h(\gamma_1 t)t\gamma_2^2 = \frac{h(1)\gamma_2^2}{\gamma_1} \approx 0.72\frac{\gamma_2^2}{\gamma_1}.$$

Thus,

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{\gamma_1}{n}\sum_{i=1}^n I(\widetilde{W}^+; \widetilde{Z}_i^+) + \frac{0.72\gamma_2^2}{\gamma_1}.$$

This concludes the proof. □

D Omitted Proofs in Section 4

D.1 Proof of Theorem 4.1

Proof. We now prove the first bound. Let $g(\tilde{w}_i, \hat{z}_i, u_i) = (-1)^{u_i}\left(\ell(\tilde{w}_i^-, \hat{z}_i) - \ell(\tilde{w}_i^+, \hat{z}_i)\right)$. By Lemma A.2, we have

$$\mathbb{E}_{\widehat{Z}_i,U_i|\tilde{w}_i}\left[g(\tilde{w}_i, \widehat{Z}_i, U_i)\right] \leq \inf_{t>0}\frac{I(\widehat{Z}_i; U_i|\widetilde{W}_i = \tilde{w}_i) + \log \mathbb{E}_{\widehat{Z}_i,U_i'|\tilde{w}_i}\left[e^{tg(\tilde{w}_i,\widehat{Z}_i,U_i')}\right]}{t}. \quad (19)$$

Since $U_i' \perp\!\!\!\perp \widehat{Z}_i$, we have $\mathbb{E}_{U_i'}\left[g(\tilde{w}_i, \hat{z}_i, U_i')\right] = \mathbb{E}_{U_i'}\left[(-1)^{U_i'}\left(\ell(\tilde{w}_i^-, \hat{z}_i) - \ell(\tilde{w}_i^+, \hat{z}_i)\right)\right] = 0$ for any $\tilde{w}_i$ and $\hat{z}_i$. Ergo,

$$\mathbb{E}_{\widehat{Z}_i|\tilde{w}_i}\left[\mathbb{E}_{U_i'}\left[g(\tilde{w}_i, \widehat{Z}_i, U_i')\right]\right] = 0.$$

By the definition of $\Delta_1(\tilde{w}_i)$,

$$\left|g(\tilde{w}_i, \widehat{Z}_i, U_i')\right| = \left|\ell(\tilde{w}_i^-, \widehat{Z}_i) - \ell(\tilde{w}_i^+, \widehat{Z}_i)\right| \leq \sup_{z_i \in \mathcal{Z}_{\tilde{w}_i}} \left|\ell(\tilde{w}_i^-, z_i) - \ell(\tilde{w}_i^+, z_i)\right| \leq \Delta_1(\tilde{w}_i).$$

Thus, $g(\tilde{w}_i, \widehat{Z}_i, U_i')$ is a zero-mean random variable bounded in $[-\Delta_1(\tilde{w}_i), \Delta_1(\tilde{w}_i)]$ for a fixed $\tilde{w}_i$. By Lemma A.3, we have

$$\mathbb{E}_{\widehat{Z}_i, U_i'|\tilde{w}_i}\left[e^{tg(\tilde{w}, \widehat{Z}_i, U_i')}\right] \leq e^{\frac{t^2\Delta_1(\tilde{w}_i)^2}{2}}.$$

Plugging the above into Eq. (19),

$$\begin{aligned}\mathbb{E}_{\widehat{Z}_i, U_i|\tilde{w}_i}\left[g(\tilde{w}_i, \widehat{Z}_i, U_i)\right] \leq & \inf_{t>0} \frac{I(\widehat{Z}_i; U_i|\widetilde{W}_i = \tilde{w}_i) + \frac{t^2\Delta_1(\tilde{w}_i)^2}{2}}{t} \\ =& \Delta_1(\tilde{w}_i)\sqrt{2I(\widehat{Z}_i; U_i|\widetilde{W}_i = \tilde{w}_i)}.\end{aligned} \tag{20}$$

Recall Eq. (14) in Lemma A.1 and by Jensen's inequality for the absolute function, the first bound is obtained:

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\widetilde{W}_i}\left[\Delta_1(\widetilde{W}_i)\sqrt{2I^{\widetilde{W}_i}(\widehat{Z}_i; U_i)}\right]. \tag{21}$$

To prove the second bound, we return to Eq. (20), and take expectation over $\widetilde{W}_i$ first. By Jensen's inequality,

$$\begin{aligned}\mathbb{E}_{\widehat{Z}_i, U_i, \widetilde{W}_i}\left[g(\tilde{W}_i, \widehat{Z}_i, U_i)\right] \leq & \inf_{t>0} \frac{I(\widehat{Z}_i; U_i|\widetilde{W}_i) + \frac{t^2\mathbb{E}_{\widetilde{W}_i}[\Delta(\widetilde{W}_i)^2]}{2}}{t} \\ =& \sqrt{2\mathbb{E}_{\widetilde{W}_i}\left[\Delta(\widetilde{W}_i)^2\right]I(\widehat{Z}_i; U_i|\widetilde{W}_i)}.\end{aligned} \tag{22}$$

Therefore, we have the second bound as below

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n}\sum_{i=1}^n \sqrt{2\mathbb{E}_{\widetilde{W}_i}\left[\Delta(\widetilde{W}_i)^2\right]I(\widehat{Z}_i; U_i|\widetilde{W}_i)}. \tag{23}$$

For the second part of Theorem 4.1, notice that it's valid to let $\gamma_3 = \mathbb{E}_{\widetilde{W}_i}\left[\Delta(\widetilde{W}_i)\right]$, then recall Eq. (21),

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\widetilde{W}_i}\left[\Delta(\widetilde{W}_i)\sqrt{2I^{\widetilde{W}_i}(\widehat{Z}_i; U_i)}\right] \leq \frac{\sqrt{2}\gamma_3}{n}\sum_{i=1}^n \sqrt{\sup_{\tilde{w}_i \in (\mathcal{W}_s)^2_{s \in \mathcal{Z}^n}} I^{\tilde{w}_i}(\widehat{Z}_i; U_i)}.$$

This completes the proof. □

D.2 Proof of Theorem 4.2

Proof. The proof is similar to [18, Theorem 2.1]. By the chain rule,

$$I(\widehat{Z}_i; U_i, \widetilde{W}_i) = I(\widehat{Z}_i; U_i|\widetilde{W}_i) + I(\widehat{Z}_i; \widetilde{W}_i). \tag{24}$$

Since $H(\widehat{Z}_i|U_i, \widetilde{W}_i) = H(\widehat{Z}_i|W_i, U_i, \widetilde{W}_i) = H(\widehat{Z}_i|W_i)$, we have $I(\widehat{Z}_i; U_i, \widetilde{W}_i) = H(\widehat{Z}_i) - H(\widehat{Z}_i|U_i, \widetilde{W}_i) = H(\widehat{Z}_i) - H(\widehat{Z}_i|W_i) = I(\widehat{Z}_i; W_i)$. Thus, $I(\widehat{Z}_i; U_i, \widetilde{W}_i) = I(\widehat{Z}_i; W_i)$. Recall Eq. (24) and by the non-negativity of mutual information, we have $I(\widehat{Z}_i; U_i|\widetilde{W}_i) \leq I(W_i; \widehat{Z}_i)$. Note that $I(W_i; \widehat{Z}_i) = I(\widetilde{W}_i^+; \widetilde{Z}_i^+) = I(W; Z_i)$. This completes the proof. □

D.3 Proof of Theorem 4.3

Proof. We first return to Eq. (19) in the previous proof, and we have already known that $g(\tilde{w}_i, \widehat{Z}_i, U_i')$ is a zero-mean random variable bounded in $[-\Delta_1(\tilde{w}_i), \Delta_1(\tilde{w}_i)]$ for a fixed $\tilde{w}_i$.

By Lemma A.5, we have

$$\begin{aligned}\log \mathbb{E}_{\widehat{Z}_i,U_i'|\tilde{w}_i}\left[e^{tg(\tilde{w}_i,\widehat{Z}_i,U_i')}\right] \leq & h\left(\Delta_1(\tilde{w}_i)t\right) t^2 \mathbb{E}_{\widehat{Z}_i,U_i'|\tilde{w}_i}\left[g(\tilde{w}_i,\widehat{Z}_i,U_i')^2\right] \\ = & h\left(\Delta_1(\tilde{w}_i)t\right) t^2 \mathbb{E}_{\widehat{Z}_i|\tilde{w}_i}\left[\left(\ell(\tilde{w}_i^-,\widehat{Z}_i)-\ell(\tilde{w}_i^+,\widehat{Z}_i)\right)^2\right].\end{aligned}$$

Plugging the above into Eq. (19),

$$\mathbb{E}_{\widehat{Z}_i,U_i|\tilde{w}_i}\left[g(\tilde{w}_i,\widehat{Z}_i,U_i)\right] \leq \inf_{t>0} \frac{I(\widehat{Z}_i;U_i|\widetilde{W}_i=\tilde{w}_i)}{t} + h\left(\Delta_1(\tilde{w}_i)t\right) t \mathbb{E}_{\widehat{Z}_i|\tilde{w}_i}\left[\left(\ell(\tilde{w}_i^-,\widehat{Z}_i)-\ell(\tilde{w}_i^+,\widehat{Z}_i)\right)^2\right]. \tag{25}$$

Let $t=\frac{1}{\Delta_1(\tilde{w}_i)}$, we have

$$\mathbb{E}_{\widehat{Z}_i,U_i|\tilde{w}_i}\left[g(\tilde{w}_i,\widehat{Z}_i,U_i)\right] \leq \Delta_1(\tilde{w}_i) I(\widehat{Z}_i;U_i|\widetilde{W}_i=\tilde{w}_i) + 0.72\frac{\mathbb{E}_{\widehat{Z}_i|\tilde{w}_i}\left[\left(\ell(\tilde{w}_i^-,\widehat{Z}_i)-\ell(\tilde{w}_i^+,\widehat{Z}_i)\right)^2\right]}{\Delta_1(\tilde{w}_i)}.$$

Let $\Lambda(\tilde{w}_i)=\mathbb{E}_{\widehat{Z}_i|\tilde{w}_i}\left[\left(\ell(\tilde{w}_i^-,\widehat{Z}_i)-\ell(\tilde{w}_i^+,\widehat{Z}_i)\right)^2\right]\Big/\Delta_1(\tilde{w}_i)^2$, then

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\widetilde{W}_i}\left[\Delta_1(\widetilde{W}_i)\left(I^{\widetilde{W}_i}(\widehat{Z}_i;U_i)+0.72\Lambda(\widetilde{W}_i)\right)\right].$$

For the second part, if $\mathcal{A}$ is further β_2-uniform stable, recall Eq. (25) and by the non-decreasing property of h, we have

$$\mathbb{E}_{\widehat{Z}_i,U_i|\tilde{w}_i}\left[g(\tilde{w}_i,\widehat{Z}_i,U_i)\right] \leq \inf_{t>0} \frac{I(\widehat{Z}_i;U_i|\widetilde{W}_i=\tilde{w}_i)}{t} + h\left(\beta_2 t\right) t \mathbb{E}_{\widehat{Z}_i|\tilde{w}_i}\left[\left(\ell(\tilde{w}_i^-,\widehat{Z}_i)-\ell(\tilde{w}_i^+,\widehat{Z}_i)\right)^2\right].$$

Let $t=\frac{1}{\beta_2}$ and taking expectation over $\widetilde{W}_i$, we have

$$\begin{aligned}\mathbb{E}_{\widehat{Z}_i,U_i,\widetilde{W}_i}\left[g(\widetilde{W}_i,\widehat{Z}_i,U_i)\right] \leq & \beta_2 I(\widehat{Z}_i;U_i|\widetilde{W}_i) + 0.72\frac{\mathbb{E}_{\widehat{Z}_i,\widetilde{W}_i}\left[\left(\ell(\widetilde{W}_i^-,\widehat{Z}_i)-\ell(\widetilde{W}_i^+,\widehat{Z}_i)\right)^2\right]}{\beta_2} \\ = & \beta_2 I(\widehat{Z}_i;U_i|\widetilde{W}_i) + 0.72\frac{\gamma_4^2}{\beta_2},\end{aligned}$$

where the equality is by the definition of γ_4-SCH-D stability.

Thus,

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{1}{n}\sum_{i=1}^{n}\beta_2 I(\widehat{Z}_i;U_i|\widetilde{W}_i) + 0.72\frac{\gamma_4^2}{\beta_2}.$$

This concludes the proof. □

D.4 Proof of Theorem 4.4

We present a stronger version of Theorem 4.4.

Theorem D.1. *Under the same conditions in Theorem 4.1, and we further assume that $\mathcal{A}$ is γ_2-SCH-B stable and symmetric with respect to S, i.e. it does not depend on the order of the elements in the training sample. Let $\bar{\Delta}_1(\widetilde{W})=\frac{1}{n}\sum_{i=1}^{n}\Delta_1(\widetilde{W}_i)^2$, we have*

$$\mathbb{E}_{W,S}\left[\left(L_S(W)-L_\mu(W)\right)^2\right] \leq \frac{6}{n}\mathbb{E}_{\widetilde{W}}\left[\bar{\Delta}_1(\widetilde{W})\left(I^{\widetilde{W}}(E;U)+\frac{\log 3}{2}\right)\right]+\frac{1}{n}+4\gamma_2^2.$$

Then Theorem 4.4 is a corollary of Theorem D.1.

Proof of Theorem 4.4. For β_2-uniform stable algorithm, by $\bar{\Delta}_1(\widetilde{W}) \leq \beta_2^2$ and $\gamma_2^2 \leq \beta_2^2$, we have

$$\begin{aligned}\mathbb{E}_{W,S}\left[\left(L_S(W) - L_\mu(W)\right)^2\right] \leq & \frac{6\beta_2^2}{n}\left(I(E;U|\widetilde{W}) + \frac{\log 3}{2}\right) + \frac{1}{n} + 4\beta_2^2 \\ = & 4\beta_2^2\left(\frac{1.5I(E;U|\widetilde{W}) + 0.82}{n} + 1\right) + \frac{1}{n}.\end{aligned}$$

This completes the proof. □

Before we prove Theorem D.1, we need to first obtain the following lemma.

Lemma D.1. *Under the same conditions in Theorem 4.1, let* $\bar{\Delta}_1(\widetilde{W}) = \frac{1}{n}\sum_{i=1}^n \Delta_1(\widetilde{W}_i)^2$, *we have*

$$\mathbb{E}_{W,S}\left[\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{W^i|W}\left[\ell(W^i, Z_i)\right] - L_S(W)\right)^2\right] \leq \frac{3}{n}\mathbb{E}_{\widetilde{W}}\left[\bar{\Delta}_1(\widetilde{W})\left(I^{\widetilde{W}}(E;U) + \frac{\log 3}{2}\right)\right].$$

Proof of Lemma D.1. Here we borrow some proof techniques used in [58, Thm. 2].

Let $g(\tilde{w}, \hat{z}_i, u_i) = (-1)^{u_i}\left(\ell(\tilde{w}_i^-, \hat{z}_i) - \ell(\tilde{w}_i^+, \hat{z}_i)\right)$ and let $G \sim \mathcal{N}(0,1)$ be an independent standard Gaussian random variable. Let $f = t \cdot (\frac{1}{n}\sum_{i=1}^n g)^2$ in Lemma A.2, then

$$\begin{aligned}\mathbb{E}_{E,U|\tilde{w}}\left[\left(\frac{1}{n}\sum_{i=1}^n g(\tilde{w}, \widehat{Z}_i, U_i)\right)^2\right] \leq & \inf_{t>0} \frac{I(E;U|\widetilde{W} = \tilde{w}) + \log \mathbb{E}_{E,U'|\tilde{w}}\left[e^{t\left(\frac{1}{n}\sum_{i=1}^n g(\tilde{w}, \widehat{Z}_i, U_i')\right)^2}\right]}{t} \\ = & \inf_{t>0} \frac{I(E;U|\widetilde{W} = \tilde{w}) + \log \mathbb{E}_{E,U'|\tilde{w}}\left[\mathbb{E}_G\left[e^{\frac{G\sqrt{2t}}{n}\sum_{i=1}^n g(\tilde{w}, \widehat{Z}_i, U_i')}\right]\right]}{t} \qquad (26) \\ = & \inf_{t>0} \frac{I(E;U|\widetilde{W} = \tilde{w}) + \log \mathbb{E}_{G,E|\tilde{w}}\left[\prod_{i=1}^n \mathbb{E}_{U_i'}\left[e^{\frac{G\sqrt{2t}}{n} g(\tilde{w}, \widehat{Z}_i, U_i')}\right]\right]}{t} \\ \leq & \inf_{t>0} \frac{I(E;U|\widetilde{W} = \tilde{w}) + \log \mathbb{E}_G\left[e^{\frac{G^2 t\sum_{i=1}^n \Delta_1(\tilde{w}_i)^2}{n^2}}\right]}{t} \qquad (27) \\ \leq & \inf_{t\in\left(0, \frac{n^2}{2\sum_{i=1}^n \Delta_1(\tilde{w}_i)^2}\right)} \frac{I(E;U|\widetilde{W} = \tilde{w}) + \log\left(1\Big/\sqrt{1 - \frac{2t\sum_{i=1}^n \Delta_1(\tilde{w}_i)^2}{n^2}}\right)}{t} \qquad (28) \\ = & \inf_{t\in\left(0, \frac{n^2}{2\sum_{i=1}^n \Delta_1(\tilde{w}_i)^2}\right)} \frac{I(E;U|\widetilde{W} = \tilde{w}) - \frac{1}{2}\log\left(1 - \frac{2t\sum_{i=1}^n \Delta_1(\tilde{w}_i)^2}{n^2}\right)}{t},\end{aligned}$$

where Eq. (26) is by the moment generating function of Gaussian distribution: $\mathbb{E}_G\left[e^{\lambda G}\right] = e^{\frac{\lambda^2}{2}}$ for all $\lambda \in \mathbb{R}$, Eq. (27) is by Lemma A.3 and Eq. (28) is by the moment generating function of chi-squared distribution: $\mathbb{E}_G\left[e^{\lambda G^2}\right] \leq \frac{1}{\sqrt{1-2\lambda}}$ for $\lambda < \frac{1}{2}$.

Let $t = \frac{n^2}{3\sum_{i=1}^n \Delta_1(\tilde{w}_i)^2}$ be substituted to the last equation above, we have

$$\mathbb{E}_{E,U|\tilde{w}}\left[\left(\frac{1}{n}\sum_{i=1}^n g(\tilde{w}, \widehat{Z}_i, U_i)\right)^2\right] \leq \frac{3}{n^2}\sum_{i=1}^n \Delta_1(\tilde{w}_i)^2\left(I(E;U|\widetilde{W} = \tilde{w}) + \frac{\log 3}{2}\right). \qquad (29)$$

Let $\bar{\Delta}_1(\tilde{w}) = \frac{1}{n}\sum_{i=1}^n \Delta_1(\tilde{w}_i)^2$, and taking expectation over $\widetilde{W}$ for both sides,

$$\mathbb{E}_{E,U,\widetilde{W}}\left[\left(\frac{1}{n}\sum_{i=1}^n g(\widetilde{W},\widehat{Z}_i,U_i)\right)^2\right] \leq \frac{3}{n}\mathbb{E}_{\widetilde{W}}\left[\bar{\Delta}_1(\widetilde{W})\left(I^{\widetilde{W}}(E;U)+\frac{\log 3}{2}\right)\right]. \tag{30}$$

Applying Jensen's inequality to the square function, we have

$$\mathbb{E}_{W,S}\left[\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{W^i|W}\left[\ell(W^i,Z_i)\right]-L_S(W)\right)^2\right] \leq \mathbb{E}_{E,U,\widetilde{W}}\left[\left(\frac{1}{n}\sum_{i=1}^n g(\widetilde{W},\widehat{Z}_i,U_i)\right)^2\right].$$

Combining Eq. (30) with the inequality above will concludes the proof. □

We are now in a position to prove Theorem D.1.

Proof of Theorem D.1.

$$\begin{aligned}
&\mathbb{E}_{W,S}\left[\left(L_S(W)-L_\mu(W)\right)^2\right]\\
=&\mathbb{E}_{W,S}\left[\left(\frac{1}{n}\sum_{i=1}^n \ell(W,Z_i)-\mathbb{E}_{Z'}\left[\ell(W,Z')\right]\right)^2\right]\\
=&\mathbb{E}_{W,S}\left[\left(\frac{1}{n}\sum_{i=1}^n \ell(W,Z_i)-\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{W^i|W}\left[\ell(W^i,Z_i)\right]+\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{W^i|W}\left[\ell(W^i,Z_i)\right]-\mathbb{E}_{Z'}\left[\ell(W,Z')\right]\right)^2\right]\\
\leq& \underbrace{2\,\mathbb{E}_{W,S}\left[\left(\frac{1}{n}\sum_{i=1}^n \ell(W,Z_i)-\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{W^i|W}\left[\ell(W^i,Z_i)\right]\right)^2\right]}_{B_1}\\
&\quad+\underbrace{2\,\mathbb{E}_{W,S}\left[\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{W^i|W}\left[\ell(W^i,Z_i)\right]-\mathbb{E}_{Z'}\left[\ell(W,Z')\right]\right)^2\right]}_{B_2},
\end{aligned}$$

where the last inequality is by $(x+y)^2 \leq 2x^2+2y^2$. Notice that B_1 can be bounded by using Lemma D.1. We now focus on B_2. Since $\mathbb{E}_{Z'}\left[\ell(W,Z')\right]=\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{Z'}\left[\ell(W,Z')\right]$, we have

$$\begin{aligned}
B_2 =&\mathbb{E}_{W,S}\left[\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{W^i|W}\left[\ell(W^i,Z_i)\right]-\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{Z'}\left[\ell(W,Z')-\mathbb{E}_{W^i|W}\left[\ell(W^i,Z')\right]+\mathbb{E}_{W^i|W}\left[\ell(W^i,Z')\right]\right]\right)^2\right]\\
=&\mathbb{E}_{W,S}\left[\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{W^i|W}\left[\ell(W^i,Z_i)-\mathbb{E}_{Z'}\left[\ell(W^i,Z')\right]\right]-\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{Z'}\left[\ell(W,Z')-\mathbb{E}_{W^i|W}\left[\ell(W^i,Z')\right]\right]\right)^2\right]\\
\leq&\underbrace{2\,\mathbb{E}_{W,S}\left[\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{W^i|W}\left[\ell(W^i,Z_i)-\mathbb{E}_{Z'}\left[\ell(W^i,Z')\right]\right]\right)^2\right]}_{B_3}\\
&\quad+\underbrace{2\,\mathbb{E}_{W}\left[\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{Z'}\left[\ell(W,Z')-\mathbb{E}_{W^i|W}\left[\ell(W^i,Z')\right]\right]\right)^2\right]}_{B_4}.
\end{aligned}$$

For B_3, we apply Jensen's inequality to move the expectation over W^i outside of the square function,

$$\begin{aligned}B_3 \leq & \mathbb{E}_{W^1,W^2,\ldots,W^n,S}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\ell(W^i,Z_i)-\mathbb{E}_{Z'}\left[\ell(W^i,Z')\right]\right)^2\right]\\ = & \mathbb{E}_{S',S,R}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\ell(\mathcal{A}(S^i,R),Z_i)-\mathbb{E}_{Z'}\left[\ell(\mathcal{A}(S^i,R),Z')\right]\right)^2\right].\end{aligned}$$

Notice that S', S and R are all independent with each other (so W^i, Z_i and Z'_i are also independent with each other). If we further let $\mathcal{A}$ be symmetric, namely W^i does not dependent on i, then the inequality above is equivalent to

$$\begin{aligned}B_3 \leq & \mathbb{E}_{W,S'}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\ell(W,Z'_i)-\mathbb{E}_{Z'}\left[\ell(W,Z')\right]\right)^2\right]\\ = & \mathbb{E}_{W}\left[\mathbb{E}_{S'}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\ell(W,Z'_i)-\mathbb{E}_{Z'}\left[\ell(W,Z')\right]\right)^2\right]\right].\end{aligned}$$

Hence, the inner expectation in the RHS above is just the variance of the sample mean of n i.i.d bounded random variables. Recall that $\ell(\cdot,\cdot)\in[0,1]$, thereby

$$B_3 \leq \mathbb{E}_W\left[\frac{\mathrm{Var}_{Z'}(\ell(W,Z'))}{n}\right]\leq\frac{1}{4n},$$

where the second inequality is by Lemma A.4.

Then, for B_4, we also apply Jensen's inequality to the square function, and by the definition of γ_2-SCH-B stability, we have

$$\begin{aligned}B_4 \leq & \mathbb{E}_{W,Z'}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\ell(W,Z')-\mathbb{E}_{W^i|W}\left[\ell(W^i,Z')\right]\right)^2\right]\\ \leq & \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{W,Z'}\left[\left(\ell(W,Z')-\mathbb{E}_{W^i|W}\left[\ell(W^i,Z')\right]\right)^2\right]\leq\gamma_2^2.\end{aligned}$$

Putting everthing together, we have

$$\begin{aligned}\mathbb{E}_{W,S}\left[\left(L_S(W)-L_\mu(W)\right)^2\right] \leq & 2B_1+2B_2\\ \leq & 2B_1+4B_3+4B_4\\ \leq & 2B_1+\frac{1}{n}+4\gamma_2\\ \leq & \frac{6}{n}\mathbb{E}_{\widetilde{W}}\left[\bar{\Delta}_1(\widetilde{W})\left(I^{\widetilde{W}}(E;U)+\frac{\log 3}{2}\right)\right]+\frac{1}{n}+4\gamma_2^2,\end{aligned}$$

where the last inequality is by Lemma D.1. This completes the proof. □

D.5 Proof of Theorem 4.5

Proof. Let $g(\tilde{z}_i^+,w_i,\bar{w}_i,u_i)=(-1)^{u_i}\left(\ell(\bar{w}_i,\tilde{z}_i^+)-\ell(w_i,\tilde{z}_i^+)\right)$. Again, by Lemma A.2, we have

$$\mathbb{E}_{W_i,\overline{W}_i,U_i|\tilde{z}_i^+}\left[g(\tilde{z}_i^+,W_i,\overline{W}_i,U_i)\right]\leq\inf_{t>0}\frac{I(W_i,\overline{W}_i;U_i|\widetilde{Z}_i^+=\tilde{z}_i^+)+\log\mathbb{E}_{W_i,\overline{W}_i,U'_i|\tilde{z}_i^+}\left[e^{tg(\tilde{z}_i^+,W_i,\overline{W}_i,U'_i)}\right]}{t}. \quad (31)$$

Similar to the previous proofs, it's easy to see that $g(\tilde{z}_i^+, W_i, \overline{W}_i, U_i')$ is a zero-mean random variable bounded in $[-\Delta_2(\tilde{z}_i^+), \Delta_2(\tilde{z}_i^+)]$. Thus,

$$\mathbb{E}_{W_i,\overline{W}_i,U_i|\tilde{z}_i^+}\left[g(\tilde{z}_i^+, W_i, \overline{W}_i, U_i)\right] \le \inf_{t>0} \frac{I(W_i,\overline{W}_i;U_i|\widetilde{Z}_i^+=\tilde{z}_i^+)+\frac{t^2\Delta_2(\tilde{z}_i^+)^2}{2}}{t}. \tag{32}$$

To prove the first bound, we let $t=\sqrt{\frac{I(W_i,\overline{W}_i;U_i|\widetilde{Z}_i^+=\tilde{z}_i^+)}{2\Delta_2(\tilde{z}_i^+)^2}}$, then

$$\mathbb{E}_{W_i,\overline{W}_i,U_i|\tilde{z}_i^+}\left[g(\tilde{z}_i^+, W_i, \overline{W}_i, U_i)\right] \le \Delta_2(\tilde{z}_i^+)\sqrt{2I(W_i,\overline{W}_i;U_i|\widetilde{Z}_i^+=\tilde{z}_i^+)}.$$

Recall Eq. (14) in Lemma A.1, hence,

$$|\mathcal{E}_\mu(\mathcal{A})| \le \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\widetilde{Z}_i^+}\left[\Delta_2(\widetilde{Z}_i^+)\sqrt{2I^{\widetilde{Z}_i^+}(W_i,\overline{W}_i;U_i)}\right]. \tag{33}$$

To prove the second bound, we take expectation over $\widetilde{Z}_i^+$ for Eq. (32),

$$\mathbb{E}_{W_i,\overline{W}_i,U_i,\widetilde{Z}_i^+}\left[g(\widetilde{Z}_i^+, W_i, \overline{W}_i, U_i)\right] \le \inf_{t>0} \frac{I(W_i,\overline{W}_i;U_i|\widetilde{Z}_i^+)+\frac{t^2\mathbb{E}_{\widetilde{Z}_i^+}\left[\Delta_2(\widetilde{Z}_i^+)^2\right]}{2}}{t}.$$

Let $t=\sqrt{\frac{I(W_i,\overline{W}_i;U_i|\widetilde{Z}_i^+)}{2\mathbb{E}_{\widetilde{Z}_i^+}\left[\Delta_2(\widetilde{Z}_i^+)^2\right]}}$, then

$$\mathbb{E}_{W_i,\overline{W}_i,U_i,\widetilde{Z}_i^+}\left[g(\widetilde{Z}_i^+, W_i, \overline{W}_i, U_i)\right] \le \sqrt{2\mathbb{E}_{\widetilde{Z}_i^+}\left[\Delta_2(\widetilde{Z}_i^+)^2\right]I(W_i,\overline{W}_i;U_i|\widetilde{Z}_i^+)}.$$

Ergo,

$$|\mathcal{E}_\mu(\mathcal{A})| \le \frac{1}{n}\sum_{i=1}^n \sqrt{2\mathbb{E}_{\widetilde{Z}_i^+}\left[\Delta_2(\widetilde{Z}_i^+)^2\right]I(W_i,\overline{W}_i;U_i|\widetilde{Z}_i^+)}. \tag{34}$$

This completes the proof. □

E Omitted Proof in Section 6

E.1 Proof of Proposition 1

Proof. By Jensen's inequality and triangle inequality, for any $i \in [n]$, we have

$$\begin{aligned}
&\mathbb{E}_{S,R,Z'}\left[\left(\ell(W,Z')-\mathbb{E}_{W^i|W}\left[\ell(W^i,Z')\right]\right)^2\right]\\
\le&\mathbb{E}_{W,W^i,Z'}\left[\left(\ell(W,Z')-\ell(W^i,Z')\right)^2\right]\\
=&\mathbb{E}_{W,W^i,Z'}\left[\left(\ell(W,Z')-\ell(w^*,Z')+\ell(w^*,Z')-\ell(W^i,Z')\right)^2\right]\\
\le&2\mathbb{E}_{W,Z'}\left[\left(\ell(W,Z')-\ell(w^*,Z')\right)^2\right]+2\mathbb{E}_{W^i,Z'}\left[\left(\ell(W^i,Z')-\ell(w^*,Z')\right)^2\right]\\
\le&4B\mathbb{E}_W\left[(L_\mu(W)-L_\mu(w^*))^{\frac{1}{\kappa}}\right],
\end{aligned}$$

where the last inequality is by the definition of the Bernstein condition. This completes the proof. □

E.2 Proof of Theorem 6.1

Proof. Let $g(\Delta\ell_i, u_i) = (-1)^{u_i}\Delta\ell_i$. Notice that $|g(\Delta L_i, U_i')| \le \beta_2$ and $g(\Delta L_i, U_i')$ is agian zero-mean, then

$$\begin{aligned}
\mathbb{E}_{\Delta L_i,U_i}\left[g(\Delta L_i,U_i)\right]\le&\frac{I(\Delta L_i;U_i)+\log\mathbb{E}_{\Delta L_i,U_i'}\left[e^{tg(\Delta L_i,U_i')}\right]}{t}\\
\le&\beta_2\sqrt{2I(\Delta L_i;U_i)}.
\end{aligned}$$

Thus,

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{\beta_2}{n}\sum_{i=1}^{n}\sqrt{2I(\Delta L_i;U_i)}.$$

To prove the disintegrated CMI bound, we let g be defined in the same way, and the remaining development is the same with the proof in Theorem 4.5.

For the second inequality, notice that $I(\Delta L_i;U_i) \leq I(\Delta L_i;U_i|\widetilde{Z}_i^+)$ by using the chain rule of mutual information and the independence between $\widetilde{Z}_i^+$ and U_i. In addition, moving the expectation over $\widetilde{Z}_i^+$ inside the square-root function by Jensen's inequality, we have $\mathbb{E}_{\widetilde{Z}_i^+}\sqrt{I^{\widetilde{Z}_i^+}(\Delta L_i;U_i)} \leq \sqrt{I(\Delta L_i;U_i|\widetilde{Z}_i^+)}$. □

E.3 Proof of Theorem 6.2

Proof. Before we prove Theorem 6.2, we first show the following lemma.

Lemma E.1. *For any $i \in [n]$, we have $\sum_{i=1}^{n} I(F_i,\bar{F}_i;U_i|\widetilde{Z}_i^+) \leq I(F_{[n]},\bar{F}_{[n]};U|\widetilde{Z}_{[n]}^+)$.*

Proof of Lemma E.1. First, by $I(F_i,\bar{F}_i;U_i|\widetilde{Z}_i^+) = H(U_i) - H(U_i|F_i,\bar{F}_i,\widetilde{Z}_i^+)$ and $I(F_i,\bar{F}_i;U_i|\widetilde{Z}_{[n]}^+) = H(U_i) - H(U_i|F_i,\bar{F}_i,\widetilde{Z}_{[n]}^+)$, and notice that $H(U_i|F_i,\bar{F}_i,\widetilde{Z}_{[n]}^+) \leq H(U_i|F_i,\bar{F}_i,\widetilde{Z}_i^+)$, we have

$$I(F_i,\bar{F}_i;U_i|\widetilde{Z}_i^+) \leq I(F_i,\bar{F}_i;U_i|\widetilde{Z}_{[n]}^+). \tag{35}$$

Then, using the chain rule,

$$I(F_i,\bar{F}_i;U_i|\widetilde{Z}_{[n]}^+) + I(F_{[n]\setminus i},\bar{F}_{[n]\setminus i};U_i|\widetilde{Z}_{[n]}^+,F_i,\bar{F}_i) = I(F_{[n]},\bar{F}_{[n]};U_i|\widetilde{Z}_{[n]}^+).$$

By the non-negativity of mutual information, we have

$$I(F_i,\bar{F}_i;U_i|\widetilde{Z}_{[n]}^+) \leq I(F_{[n]},\bar{F}_{[n]};U_i|\widetilde{Z}_{[n]}^+). \tag{36}$$

Furthermore, by the independence of each U_i (i.e. $I(U_i;U_{[n]\setminus i}|\widetilde{Z}_{[n]}^+) = 0$), we have

$$\sum_{i=1}^{n} I(F_{[n]},\bar{F}_{[n]};U_i|\widetilde{Z}_{[n]}^+) \leq I(F_{[n]},\bar{F}_{[n]};U|\widetilde{Z}_{[n]}^+). \tag{37}$$

Combining Eq. (35-37) will conclude the proof. □

We now prove Theorem 6.2.

For a given $\widetilde{Z}_{[n]}$, the number of distinct values of their predictions, denoted by k, are upper bounded by the growth function of $\mathcal{F}$ evaluated at n,

$$k \leq \sum_{i=1}^{d}\binom{n}{i} \leq (\frac{en}{d})^d,$$

where the second inequality is by Sauer-Shelah lemma [53, 57] for $n > d+1$.

Thus,

$$I(F_{[n]},\bar{F}_{[n]};U\big|\widetilde{Z}_{[n]}^+) \leq H(F_{[n]},\bar{F}_{[n]}\big|\widetilde{Z}_{[n]}^+) \leq H(F_{[n]}\big|\widetilde{Z}_{[n]}^+) + H(\bar{F}_{[n]}\big|\widetilde{Z}_{[n]}^+) \leq 2d\log\left(\frac{en}{d}\right). \tag{38}$$

By Jensen's inequality and Lemma E.1, we have

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{I(F_i,\bar{F}_i;U_i|\widetilde{Z}_i^+)} \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}I(F_i,\bar{F}_i;U_i|\widetilde{Z}_i^+)} \leq \sqrt{\frac{I(F_{[n]},\bar{F}_{[n]};U|\widetilde{Z}_{[n]}^+)}{n}}.$$

Plugging Eq. (38) into the inequality above,

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{I(F_i,\bar{F}_i;U_i|\widetilde{Z}_i^+)} \leq \mathcal{O}\left(\sqrt{\frac{d}{n}\log\left(\frac{n}{d}\right)}\right),$$

which completes the proof. □

F CLB Exampls

In Example 1, [21, Thm. 17] demonstrates the non-vanishing behavior of individual IOMI and e-CMI. This is primarily attributed to the dimension-dependent nature of IOMI and CMI. Specifically, there are certain dimensional settings where IOMI can grow faster than $\mathcal{O}(n)$, as shown in [21, Thm.4], and CMI approaches a certain fraction of its upper bound, as illustrated in Example 1, resulting in non-vanishing behavior. Specifically, in Example 1, [21] employs the birthday paradox [37, Sec. 5.1] problem to demonstrate that for a large value of d, the probability that no pair of instances in $\widetilde{Z}$ sharing the same non-zero coordinate (referred to as event E_0) is smaller than a constant probability (that could be independent of n). Particularly, it is shown that if $d \geq \frac{2n-1}{1-c^{1/(2n-1)}}$, then $P(E_0) \geq c \geq \left(1-\frac{2n-1}{d}\right)^{2n-1}$. As an example, [21] chooses $d = 2n^2$, resulting in $c \geq 0.1$.

Failure of $I(W;Z_i)$ Consider the case where $d = 2n^2$. For the individual CMI [49, 70], $I(W;U_i|\widetilde{Z}_i)$, we have the following:

$$I(W;U_i|\widetilde{Z}_i) = \log 2 - H(U_i|W,\widetilde{Z}_i) \geq 0.1\cdot\log 2.$$

This inequality holds because when event E_0 does not occur, one can determine the value of U_i completely, as the returned hypothesis is a weighted sum of the sample. In other words, examining the non-zero coordinates of W is sufficient to determine U_i. For an in-depth derivation of this inequality, readers are referred to the updated version of [21], where their corrected proof involves Fano's inequality. Furthermore, using the relation $I(W;Z_i) \geq I(W;U_i|\widetilde{Z}_i)$ [70], we conclude that $I(W;Z_i) \in \Omega(1)$.

Failure of $I(\widehat{Z}_i;U_i|\widetilde{W}_i)$ Notably, our hypotheses-conditioned CMI also does not vanish for the same reason. More precisely, when $\widetilde{W}_i$ and $\widehat{Z}_i$ are given, there exists a constant probability (independent of n) that allows us to fully determine the returned hypothesis based on $\widehat{Z}_i$, thereby determining the value of U_i.

Failure of $I(\Delta L_i;U_i)$ Furthermore, even the loss difference based CMI (e.g., as shown in Theorem 6.1), which provides the tightest CMI measure, still does not vanish. This is attributed to the fact that if the hypothesis W is independent of certain Z, there exists a constant probability where the loss becomes zero (recall that the loss is the negative inner product of W and Z). Consequently, one can determine the value of U_i by observing the sign of the random variable ΔL_i. This also indicates the limitations of e-CMI and f-CMI in capturing the generalization behavior for Example 1.

In Example 2, following the approach in [21], the training sample $S = \{Z_i\}_{i=1}^n \sim \mu^n$ can be represented as $S = \frac{z_0}{R_0}(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n)$, where $\{\varepsilon_i\}_{i=1}^n$ is a sequence of independent Rademacher random variables, i.e., $\varepsilon_i \sim \mathrm{Unif}(\{-1,1\})$. The empirical risk is given by $L_S(W) = -\frac{-L}{nR_0}\langle W,\sum_{i=1}^n\varepsilon_i\rangle$. In this case, the ERM solution is $W_{ERM} = z_0$ if $\mathrm{sign}(\sum_{i=1}^n\varepsilon_i) = 1$, and $W_{ERM} = -z_0$ if $\mathrm{sign}(\sum_{i=1}^n\varepsilon_i) = -1$. It is clear that

$$\sup_{w,w^i,z}\left|\ell(w,z)-\ell(w^i,z)\right| \leq \sup_{w,w^i,z} L||w-w^i|| \leq 2LR_0.$$

Hence, we observe that β_2 is now a constant, whereas IOMI has an upper bound: $\sum_{i=1}^n I(W;Z_i) \leq I(W;S) \leq I(W;\mathrm{sign}(\sum_{i=1}^n \varepsilon_i)) \leq H(\mathrm{sign}(\sum_{i=1}^n \varepsilon_i)) \leq 1$, where the second inequality follows from the Markov chain $S - \mathrm{sign}(\sum_{i=1}^n \varepsilon_i) - W$. This provides us with a generalization bound of $\frac{2LR_0}{\sqrt{n}}$. Meanwhile, the actual generalization error satisfies $\mathcal{E}_\mu(\mathcal{A}) \geq \frac{LR_0}{\sqrt{2n}}$ (see [21, AppendixB] for a derivation). Thus, the IOMI bound is tight up to a constant, and the stability bound β_2 itself is vacuous. It is worth noting that $I(\widehat{Z}_i;U_i|\widetilde{W}_i) \leq I(W;Z_i) \leq 1$ by Theorem4.2, indicating that the CMI bound is also tight.

We would like to note that the failures of chained mutual information bounds [2] are not demonstrated in the counterexamples presented in [21]. Notably, when the hypothesis is quantized, it becomes more challenging to guess U_i or Z_i. Therefore, exploring the potential of chained information-theoretic bounds, which do not necessarily rely on stability notions, could be another avenue to explain the generalization behavior observed in these counterexamples.

G Additional Applications

G.1 Compression Schemes

We now consider the algorithm that has a compression scheme [30]. Formally, a sample compression scheme of size $k \in \mathbb{N}$ is a pair of maps $(\mathcal{A}_1, \mathcal{A}_2)$. Specifically, for all samples s with $n > k$, $\mathcal{A}_1 : \mathcal{Z}^n \to \mathcal{Z}^k$ compresses the sample into a length-k subsequence $\mathcal{A}_1(s) \subseteq s$. Then $\mathcal{A}_2 : \mathcal{Z}^k \to \mathcal{W}$ could be some arbitrary mapping. Hence, $\mathcal{A}(\cdot) = \mathcal{A}_2(\mathcal{A}_1(\cdot))$. Let K be the index set for S selected by $\mathcal{A}_1$, and let $\overline{K}$ be the selected index set for S^i. In this case, our supersample-conditioned CMI has an upper bound: $I(W_i, \overline{W}_i; U_i|\widetilde{Z}_i^+) \leq I(K, \overline{K}; U_i|\widetilde{Z}_i^+) \leq H(K, \overline{K}|\widetilde{Z}_i^+) \leq 2\log\binom{n}{k} \leq 2k\log n$. Then, if $\mathcal{A}$ is further β_2-uniform stable, then we have the generalization bound $\mathcal{E}_\mu(\mathcal{A}) \leq \mathcal{O}(\beta_2\sqrt{k\log n})$. If $\beta_2 < \mathcal{O}(1/\sqrt{n})$, this bound improves the bound in [58]. It is unclear if we can obtain any improved bound for *stable* compression schemes [7], in which case [19] provides an optimal bound that removing the $\log n$ factor for the realizable setting. A main difficulty is that an interpolating algorithm is usually unstable due to the fitting-stability tradeoff [54, Sec. 13.4].

G.2 Distillation Algorithm

The high-probability generalization property of distillation algorithm is studied in [16]. In the first training stage of distillation, we obtain a w_s^* from a highly complex hypothesis space $\mathcal{W}_1$ based on a training sample s. Same to [16], we assume that the first learning stage is α-sensitive, namely $||w_s^* - w_{s^i}^*|| \leq \alpha = \mathcal{O}(1/n)$. In the second stage, the algorithm $\mathcal{A}$ will select a hypothesis that is λ-close to w_s^* from a less complex hypothesis space $\mathcal{W}_2 = \{w \in \mathcal{W} : ||w - w_s^*||_\infty \leq \lambda\}$. Let the loss function ℓ be L-Lipschitz with respect to the first argument. Consequently, $\gamma_3 \leq L||w_s^* - w_{s^i}^*|| \leq L\alpha$. Then, by Theorem 4.1, we have $\mathcal{E}_\mu(\mathcal{A}) \leq L\alpha\frac{1}{n}\sum_{i=1}^n \sqrt{2H(U_i)} = \sqrt{2\log 2}L\alpha$. Notice that the loss here may not necessarily be bounded or sub-Gaussian, rendering previous bounds inapplicable.

G.3 Regularized Empirical Risk Minimization

Regularized Empirical Risk Minimization (ERM) learning rules involve minimizing the empirical risk and a regularization function jointly: $\arg\min_{w\in\mathcal{W}} L_S(w) + f_{\text{reg}}(w)$, where $f_{\text{reg}} : \mathcal{W} \to \mathbb{R}$. Here we specifically consider Tikhonov regularization [54], namely $f_{\text{reg}}(w) = \lambda||w||^2$, where $\lambda > 0$ is a tradeoff coefficient. The regularized ERM algorithm $\mathcal{A}$ aims to find

$$w = \arg\min_{w\in\mathcal{W}} L_S(w) + \lambda||w||^2.$$

This regularization term ensures strong convexity of the training objective. Based on Theorem 4.3, we can derive the following results.

Corollary G.1. *Assume that the loss function ℓ is convex and L-Lipschitz. Then, for the regularized ERM algorithm with Tikhonov regularization, we have*

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{2L^2}{\lambda n}\left(\frac{1}{n}\sum_{i=1}^n I(\widehat{Z};U_i|\widetilde{W}_i) + 0.72\right).$$

Proof of Corollary G.1. By invoking [54, Corollary 13.6], we know that $\gamma_4 \leq \beta_2 = \frac{2L^2}{\lambda n}$. Plugging the value of β_2 will give us the desired result. □

Corollary G.2. *Assume that the loss function is ρ-smooth and nonnegative. Let $\lambda \geq \frac{2\rho}{n}$. Then, for the regularized ERM algorithm with Tikhonov regularization, we have*

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \frac{48\rho\hat{L}_n}{\lambda n}\left(\frac{1}{n}\sum_{i=1}^{n} I(\widehat{Z}; U_i|\widetilde{W}_i) + 0.72\right).$$

Proof of Corollary G.2. By invoking [54, Corollary 13.7], we know that $\gamma_4 \leq \beta_2 = \frac{48\rho\hat{L}_n}{\lambda n}$. Plugging the value of β_2 will give us the desired result. □

Although these bounds do not enhance the convergence rate of $\mathcal{O}(1/n)$ in these settings, they consistently offer tighter results compared to uniform stability-based bounds if $\frac{1}{n}\sum_{i=1}^{n} I(\widehat{Z}; U_i|\widetilde{W}_i) \leq 0.28$. In addition, the expected empirical risk $\hat{L}_n$ appears in the bound of Corollary G.2. While λ has a lower bound, $\hat{L}_n$ could not be arbitrarily small for the regularized ERM.

Notice that previous information-theoretic bounds could not obtain the convergence rate of $\mathcal{O}(1/n)$ as in our results unless ICMI or CMI itself decays with $\mathcal{O}(1/n)$.

H Additional Discussions and Open Problems

Stochastic Gradient Descent (SGD) Since the influential work of [22], stability approaches have been widely employed to provide generalization guarantees for (sub)gradient-based optimization algorithms, such as SGD, under certain conditions like the convex-smooth-Lipschitz loss. More recently, [5] extended the results of [22] to the non-smooth loss function in the SCO setting.

In contrast, information-theoretic (weight/hypothesis-based) bounds are typically used to analyze the noisy version of SGD, known as SGLD [43, 40, 18, 49, 60]. Directly analyzing SGD poses challenges because the returned hypothesis W contains a significant amount of information about S or Z_i, resulting in potentially large (even infinite) mutual information. The prevalent approach to applying information-theoretic bounds to SGD is by introducing noise [41, 61], but this has been shown to yield non-vanishing bounds in [21, Thm. 4].

The combination of information-theoretic bounds with stability for analyzing the generalization of SGD presents a promising future direction. However, some potential difficulties may arise. For instance, if we continue to use the Gaussian noise perturbation method for the weight-based information-theoretic bounds, we would need to characterize the stability property for the perturbed SGD, which might require techniques employed in [38]. Additionally, when combining stability notions with loss difference based CMI (or e-CMI/f-CMI) bounds, as they cannot be unrolled using the chain rule and data processing inequality as in the case of weight-based IOMI/CMI bounds, it may not be possible to bound such CMI terms using trajectory-based quantities. This raises doubts about the potential for obtaining more informative generalization bounds compared to the stability-based bounds themselves.

Generalization Bounds beyond Mutual Information In the information-theoretic literature, it is common to replace mutual information with alternative distributional measures, such as Wasserstein distance based bounds and total variation based bounds [50]. A promising future direction is to incorporate the stability property of algorithms into these bounds, as demonstrated in this work. It is worth noting that obtaining KL divergence-based bounds should be straightforward since they rely on the same foundational Lemma A.2 as discussed in this paper.