uOttawa

Tighter Information-Theoretic Generalization Bounds from Supersamples

ICML
International Conference On Machine Learning

Ziqiao Wang[1] Yongyi Mao[1]

[1]University of Ottawa

July 25, 2023

Background & Contributions

- Traditional generalization bounds (e.g., VC-dim, Rademacher complexity ...) are *vacuous* in DL.

Background & Contributions

- Traditional generalization bounds (e.g., VC-dim, Rademacher complexity ...) are *vacuous* in DL.
- Information-theoretic generalization bounds can be non-vacuous since they are both **distribution-dependent** and **algorithm-dependent bounds**.

Background & Contributions

- Traditional generalization bounds (e.g., VC-dim, Rademacher complexity ...) are *vacuous* in DL.
- Information-theoretic generalization bounds can be non-vacuous since they are both **distribution-dependent** and **algorithm-dependent bounds**.
- Our contribution: New **Conditional Mutual Information (CMI)** bounds which are **either theoretically or empirically tighter** than previous CMI bounds for the **same supersample** setting.

Supersample Setting

Let $\widetilde{Z}$ drawn i.i.d. from μ and $U = (U_1, U_2, \ldots, U_n)^T \sim \mathrm{Unif}(\{0,1\}^n)$.

$$\text{Supersample } \widetilde{Z} = \begin{array}{|c|c|} \widetilde{Z}_{1,0} & \widetilde{Z}_{1,1} \\ \widetilde{Z}_{2,0} & \widetilde{Z}_{2,1} \\ \vdots & \vdots \\ \widetilde{Z}_{n,0} & \widetilde{Z}_{n,1} \end{array} \xrightarrow{U} S = \begin{array}{|c|} \widetilde{Z}_{1,U_1} \\ \widetilde{Z}_{2,U_2} \\ \vdots \\ \widetilde{Z}_{n,U_n} \end{array} \xrightarrow{\mathcal{A}} W$$

$$\mathrm{Err} \triangleq \mathbb{E}_{W,S}\left[\mathbb{E}_{Z\sim\mu}[\ell(w,Z)] - \frac{1}{n}\sum_{i=1}^{n}\ell(w,Z_i)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{W,U_i,\widetilde{Z}}\left[(-1)^{U_i}\left(\ell(W,\widetilde{Z}_{i,1}) - \ell(W,\widetilde{Z}_{i,0})\right)\right]$$

Supersample Setting

Let $\widetilde{Z}$ drawn i.i.d. from μ and $U = (U_1, U_2, \ldots, U_n)^T \sim \mathrm{Unif}(\{0,1\}^n)$.

$$\text{Supersample } \widetilde{Z} = \begin{array}{|c|c|} \widetilde{Z}_{1,0} & \widetilde{Z}_{1,1} \\ \widetilde{Z}_{2,0} & \widetilde{Z}_{2,1} \\ \vdots & \vdots \\ \widetilde{Z}_{n,0} & \widetilde{Z}_{n,1} \end{array} \xrightarrow{U} S = \begin{array}{|c|} \widetilde{Z}_{1,U_1} \\ \widetilde{Z}_{2,U_2} \\ \vdots \\ \widetilde{Z}_{n,U_n} \end{array} \xrightarrow{\mathcal{A}} W$$

$$\mathrm{Err} \triangleq \mathbb{E}_{W,S}\left[\mathbb{E}_{Z\sim\mu}[\ell(w,Z)] - \frac{1}{n}\sum_{i=1}^{n}\ell(w,Z_i)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{W,U_i,\widetilde{Z}}\left[(-1)^{U_i}\left(\ell(W,\widetilde{Z}_{i,1}) - \ell(W,\widetilde{Z}_{i,0})\right)\right]$$

Lemma (Steinke and Zakynthinou [2020])

Assume the loss is bounded between $[0,1]$, *we have* $|\mathrm{Err}| \leq \sqrt{\frac{2I(W;U|\widetilde{Z})}{n}}$.

CMI, f-CMI and e-CMI

- Using the superscripts $+$ and $-$ to replace the 0 and 1: e.g, let $\widetilde{Z}_i = (\widetilde{Z}_i^+, \widetilde{Z}_i^-)$
- $L_i \triangleq (L_i^+, L_i^-) = (\ell(W, \widetilde{Z}_i^+), \ell(W, \widetilde{Z}_i^-))$
- $\Delta L_i = L_i^- - L_i^+$

$$
\begin{array}{|c|c|}\hline \widetilde{Z}_1^+ & \widetilde{Z}_1^- \\ \widetilde{Z}_2^+ & \widetilde{Z}_2^- \\ \vdots & \vdots \\ \widetilde{Z}_n^+ & \widetilde{Z}_n^- \\ \hline \end{array}
\overset{f_W}{\Longrightarrow}
\begin{array}{|c|c|}\hline f_W(\widetilde{X}_1^+) & f_W(\widetilde{X}_1^-) \\ f_W(\widetilde{X}_2^+) & f_W(\widetilde{X}_2^-) \\ \vdots & \vdots \\ f_W(\widetilde{X}_n^+) & f_W(\widetilde{X}_n^-) \\ \hline \end{array}
\overset{\ell}{\Longrightarrow}
\begin{array}{|c|c|}\hline \ell(W, \widetilde{Z}_1^+) & \ell(W, \widetilde{Z}_1^-) \\ \ell(W, \widetilde{Z}_2^+) & \ell(W, \widetilde{Z}_2^-) \\ \vdots & \vdots \\ \ell(W, \widetilde{Z}_n^+) & \ell(W, \widetilde{Z}_n^-) \\ \hline \end{array}
\overset{\Delta}{\Longrightarrow}
\begin{array}{|c|}\hline \Delta L_1 \\ \Delta L_2 \\ \vdots \\ \Delta L_n \\ \hline \end{array}
$$

$$
\underbrace{I(W; U_i|\widetilde{Z})}_{\text{CMI}} \geq \underbrace{I(f_W(\widetilde{Z}_i); U_i|\widetilde{Z})}_{f-\text{CMI [Harutyunyan et al., 2021]}} \geq \underbrace{I(L_i; U_i|\widetilde{Z})}_{\text{e}-\text{CMI [Hellström and Durisi, 2022]}}
$$

CMI, f-CMI and e-CMI

- Using the superscripts $+$ and $-$ to replace the 0 and 1: e.g, let $\widetilde{Z}_i = (\widetilde{Z}_i^+, \widetilde{Z}_i^-)$
- $L_i \triangleq (L_i^+, L_i^-) = (\ell(W, \widetilde{Z}_i^+), \ell(W, \widetilde{Z}_i^-))$
- $\Delta L_i = L_i^- - L_i^+$

$$
\begin{array}{|c|c|}\hline \widetilde{Z}_1^+ & \widetilde{Z}_1^- \\ \widetilde{Z}_2^+ & \widetilde{Z}_2^- \\ \vdots & \vdots \\ \widetilde{Z}_n^+ & \widetilde{Z}_n^- \\ \hline \end{array}
\xRightarrow{f_W}
\begin{array}{|c|c|}\hline f_W(\widetilde{X}_1^+) & f_W(\widetilde{X}_1^-) \\ f_W(\widetilde{X}_2^+) & f_W(\widetilde{X}_2^-) \\ \vdots & \vdots \\ f_W(\widetilde{X}_n^+) & f_W(\widetilde{X}_n^-) \\ \hline \end{array}
\xRightarrow{\ell}
\begin{array}{|c|c|}\hline \ell(W, \widetilde{Z}_1^+) & \ell(W, \widetilde{Z}_1^-) \\ \ell(W, \widetilde{Z}_2^+) & \ell(W, \widetilde{Z}_2^-) \\ \vdots & \vdots \\ \ell(W, \widetilde{Z}_n^+) & \ell(W, \widetilde{Z}_n^-) \\ \hline \end{array}
\xRightarrow{\Delta}
\begin{array}{|c|}\hline \Delta L_1 \\ \Delta L_2 \\ \vdots \\ \Delta L_n \\ \hline \end{array}
$$

$$
\underbrace{I(W; U_i|\widetilde{Z})}_{\text{CMI}} \geq \underbrace{I(f_W(\widetilde{Z}_i); U_i|\widetilde{Z})}_{f-\text{CMI [Harutyunyan et al., 2021]}} \geq \underbrace{I(L_i; U_i|\widetilde{Z})}_{\text{e}-\text{CMI [Hellström and Durisi, 2022]}} \geq \underbrace{I(\Delta L_i; U_i|\widetilde{Z})}_{\text{ld}-\text{CMI (Ours)}}
$$

Generalization Bounds via Loss Difference

Theorem

Assume the loss is bounded between $[0,1]$*, we have*

$$|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\widetilde{Z}} \sqrt{2 I^{\widetilde{Z}}(\Delta L_i; U_i)} \leq \frac{1}{n}\sum_{i=1}^{n} \sqrt{2 I(\Delta L_i; U_i | \widetilde{Z})}, \tag{1}$$

$$|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^{n} \sqrt{2 I(\Delta L_i; U_i)}. \tag{2}$$

Generalization Bounds via Loss Difference

Theorem

Assume the loss is bounded between $[0,1]$, *we have*

$$|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\widetilde{Z}}\sqrt{2I^{\widetilde{Z}}(\Delta L_i; U_i)} \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2I(\Delta L_i; U_i|\widetilde{Z})}, \tag{1}$$

$$|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2I(\Delta L_i; U_i)}. \tag{2}$$

Estimating $I(W; U_i|\widetilde{Z}_i)$ vs $I(\Delta L_i; U_i)$:

- W is a high-dimensional R.V.
- ΔL_i is an one-dimensional R.V. $\Longrightarrow$ Easy-to-Compute!

A Communication View of Generalization

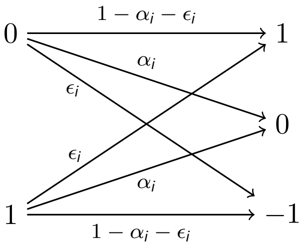


Figure: Channel from U_i to ΔL_i. Zero-one loss assumed.

Theorem

Under zero-one loss and for any interpolating algorithm $\mathcal{A}$, $I(\Delta L_i; U_i) = (1-\alpha_i)\ln 2$ nats for each i, and $|\mathrm{Err}| = L_\mu = \sum_{i=1}^{n} \frac{I(\Delta L_i; U_i)}{n \ln 2}$.

$\Longrightarrow$ Generalization error is exactly determined by the communication rate over the channel in the figure averaged over all such channels.

Fast-Rate MI Bound

Key observation:

$\mathrm{Err} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{W,U_i,\widetilde{Z}}\left[(-1)^{U_i}\left(\ell(W,\widetilde{Z}_i^{+}) - \ell(W,\widetilde{Z}_i^{-})\right)\right] = \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}_{L_i^{+},\varepsilon_i}\left[\varepsilon_i L_i^{+}\right]$, where $\varepsilon_i = (-1)^{\overline{U}_i}$ is the Rademacher variable.

Fast-Rate MI Bound

Key observation:

$\mathrm{Err} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{W,U_i,\widetilde{Z}}\left[(-1)^{U_i}\left(\ell(W,\widetilde{Z}_i^{+}) - \ell(W,\widetilde{Z}_i^{-})\right)\right] = \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}_{L_i^{+},\varepsilon_i}\left[\varepsilon_i L_i^{+}\right]$, where $\varepsilon_i = (-1)^{\overline{U}_i}$ is the Rademacher variable.

Lemma

Consider the weighted generalization error, $\mathrm{Err}_{C_1} \triangleq L_\mu - (1+C_1)L_n$. We have

$$\mathrm{Err}_{C_1} = \frac{2+C_1}{n}\sum_{i=1}^{n}\mathbb{E}_{L_i^{+},\tilde{\varepsilon}_i}\left[\tilde{\varepsilon}_i L_i^{+}\right],$$

where $\tilde{\varepsilon}_i = (-1)^{\overline{U}_i} - \frac{C_1}{C_1+2}$ is a shifted Rademacher variable with mean $-\frac{C_1}{C_1+2}$.

Fast-Rate MI Bound

Theorem

Let $\ell(\cdot,\cdot) \in [0,1]$. *There exist* $C_1, C_2 > 0$ *such that*

$$L_\mu \leq (1+C_1)L_n + \sum_{i=1}^{n} \frac{I(L_i^+;U_i)}{C_2 n}, \tag{3}$$

$$L_\mu \leq L_n + \sum_{i=1}^{n} \frac{4I(L_i^+;U_i)}{n} + 4\sqrt{\sum_{i=1}^{n} \frac{L_n I(L_i^+;U_i)}{n}}. \tag{4}$$

Fast-Rate MI Bound

ICML
International Conference
On Machine Learning

Theorem

Let $\ell(\cdot,\cdot) \in [0,1]$. *There exist* $C_1, C_2 > 0$ *such that*

$$L_\mu \leq (1 + C_1) L_n + \sum_{i=1}^{n} \frac{I(L_i^+; U_i)}{C_2 n}, \tag{3}$$

$$L_\mu \leq L_n + \sum_{i=1}^{n} \frac{4 I(L_i^+; U_i)}{n} + 4\sqrt{\sum_{i=1}^{n} \frac{L_n I(L_i^+; U_i)}{n}}. \tag{4}$$

Faster Rate than Square-Root based Bound

If $L_n \to 0$, then (3)(4) vanish with a faster rate.

Sharpness Based MI Bound

Theorem

For any $\lambda \in (0,1)$, the "λ-sharpness" at position i of the training set is defined as

$$F_i(\lambda) \triangleq \mathbb{E}_{W,Z_i}\left[\ell(W,Z_i) - (1+\lambda)\mathbb{E}_{W|Z_i}\ell(W,Z_i)\right]^2.$$

Let $F(\lambda) = \frac{1}{n}\sum_{i=1}^{n} F_i(\lambda)$. Assume $\ell(\cdot,\cdot) \in \{0,1\}$, $\lambda \in (0,1)$. Then, there exist $C_1, C_2 > 0$ such that

$$\mathrm{Err} \leq C_1 F(\lambda) + \sum_{i=1}^{n} \frac{I(L_i^+; U_i)}{C_2 n}. \tag{5}$$

Sharpness Based MI Bound

Theorem

For any $\lambda \in (0,1)$, the "λ-sharpness" at position i of the training set is defined as

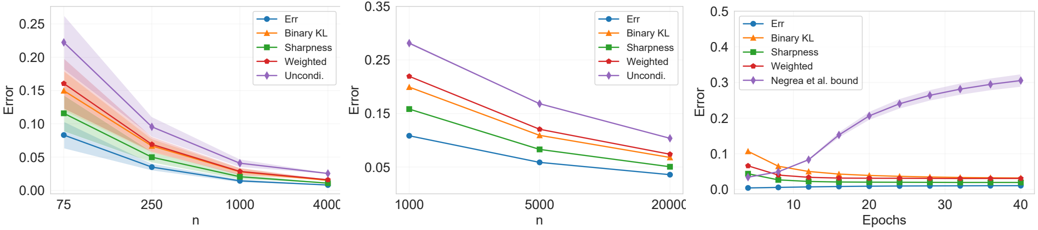
$$F_i(\lambda) \triangleq \mathbb{E}_{W,Z_i}\left[\ell(W,Z_i) - (1+\lambda)\mathbb{E}_{W|Z_i}\ell(W,Z_i)\right]^2.$$

Let $F(\lambda) = \frac{1}{n}\sum_{i=1}^{n} F_i(\lambda)$. Assume $\ell(\cdot,\cdot) \in \{0,1\}$, $\lambda \in (0,1)$. Then, there exist $C_1, C_2 > 0$ such that

$$\mathrm{Err} \leq C_1 F(\lambda) + \sum_{i=1}^{n} \frac{I(L_i^+; U_i)}{C_2 n}. \tag{5}$$

- $L_n = 0 \rightarrow F(\lambda) = 0$, but $L_n = 0 \nleftarrow F(\lambda) = 0$;
- For any fixed C_1 and C_2, Eq. (5) is tighter than Eq. (3).

Experiments



(a) CNN on MNIST (b) ResNet on CIFAR10 (c) SGLD (MNIST)

Figure: Uncondi.: $\frac{1}{n}\sum_{i=1}^{n}\sqrt{2I(\Delta L_i; U_i)}$; Binary KL: Hellström and Durisi [2022]; Weighted: $\sum_{i=1}^{n}\frac{4I(L_i^+;U_i)}{n} + 4\sqrt{\sum_{i=1}^{n}\frac{L_n I(L_i^+;U_i)}{n}}$; Sharpness: $C_1 F(\lambda) + \sum_{i=1}^{n}\frac{I(L_i^+;U_i)}{C_2 n}$.

References I

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

Hrayr Harutyunyan, Maxim Raginsky, Greg Ver Steeg, and Aram Galstyan. Information-theoretic generalization bounds for black-box learning algorithms. In *Advances in Neural Information Processing Systems*, 2021.

Fredrik Hellström and Giuseppe Durisi. A new family of generalization bounds using samplewise evaluated CMI. In *Advances in Neural Information Processing Systems*, 2022.

Thank You!