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 12, 2023

Overview

What is Generalization?

- Our ultimate interest is the **testing performance** of the learned model

What is Generalization?

- Our ultimate interest is the **testing performance** of the learned model
- Generalization error = testing error - training error

What is Generalization?

ICML International Conference On Machine Learning

- Our ultimate interest is the **testing performance** of the learned model
- Generalization error = testing error - training error
- Ideally, we wish to have training error ≈ 0 and generalization error ≈ 0

What is Generalization?

- Our ultimate interest is the **testing performance** of the learned model
- Generalization error = testing error - training error
- Ideally, we wish to have training error≈ 0 and generalization error≈ 0
- In practice, we cannot access to the unknown distribution of data

What is Generalization?

- Our ultimate interest is the **testing performance** of the learned model
- Generalization error = testing error - training error
- Ideally, we wish to have training error ≈ 0 and generalization error ≈ 0
- In practice, we cannot access to the unknown distribution of data $\Longrightarrow$ small training loss and small generalization bound/guarantee gives a small testing error.

What is Generalization Bound?

- High-probability generalization bound:

$$P(\mathit{ts\_error} - \mathit{tr\_error} \geq \epsilon) \leq \delta.$$

Or equivalently, w.p. $\geq 1 - \delta$, we have

$$\mathit{ts\_error} - \mathit{tr\_error} \leq \epsilon.$$

Typically,

$$\epsilon \leq \mathcal{O}(\frac{\text{Complexity Measure}}{n}).$$

What is Generalization Bound?

- Rademacher Complexity [Bartlett and Mendelson, 2002]:
  Given a function class $\mathcal{F} = \{f : \mathcal{Z} \to \mathbb{R}\}$ and a sample $S = \{Z_i\}_{i=1}^n$, the empirical Rademacher Complexity is

$$\hat{\mathfrak{R}}_n(\mathcal{F}) \triangleq \mathbb{E}_{\varepsilon_{1:n}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(Z_i)\right],$$

  where $\varepsilon_i \sim \mathrm{Unif}(\{-1,1\})$ is called Rademacher variable.
  $\Longrightarrow$ It measures the ability of functions from $\mathcal{F}$ to fit random noise.

Failure in Modern Deep Learning

- Zhang, Chiyuan, et al. ”Understanding deep learning requires rethinking generalization.” ICLR 2017:
  Deep neural networks (DNN) can perfectly fit random labels

Failure in Modern Deep Learning

- Zhang, Chiyuan, et al. "Understanding deep learning requires rethinking generalization." ICLR 2017:
  Deep neural networks (DNN) can perfectly fit random labels
  $\Longrightarrow$ It implicitly shows the Rademacher complexity of DNN is very large

Failure in Modern Deep Learning

- Zhang, Chiyuan, et al. ”Understanding deep learning requires rethinking generalization.” ICLR 2017:
  Deep neural networks (DNN) can perfectly fit random labels
  $\Longrightarrow$ It implicitly shows the Rademacher complexity of DNN is very large
  $\Longrightarrow$ $ts\_error - tr\_error \leq \mathcal{O}(\frac{\hat{\mathfrak{R}}_n(\mathcal{F})}{n})$ is *vacuous*!

Failure in Modern Deep Learning

- Zhang, Chiyuan, et al. "Understanding deep learning requires rethinking generalization." ICLR 2017:
  Deep neural networks (DNN) can perfectly fit random labels
  $\Longrightarrow$ It implicitly shows the Rademacher complexity of DNN is very large
  $\Longrightarrow$ $ts\_error - tr\_error \leq \mathcal{O}(\frac{\hat{\mathfrak{R}}_n(\mathcal{F})}{n})$ is *vacuous*!
- **We need new generalization bounds in deep learning!**

Notations

- Training dataset: $S = \{Z_i\}_{i=1}^n \in \mathcal{Z}$, drawn i.i.d. from μ
- Hypothesis space: $\mathcal{W} \subseteq \mathbb{R}^d$; Predictor space: $\mathcal{F} = \{f_w : \mathcal{X} \to \mathcal{Y} | w \in \mathcal{W}\}$
- Learning algorithm: $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$ by $P_{W|S}$
- Loss: $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$

Notations

- Training dataset: $S = \{Z_i\}_{i=1}^n \in \mathcal{Z}$, drawn i.i.d. from μ
- Hypothesis space: $\mathcal{W} \subseteq \mathbb{R}^d$; Predictor space: $\mathcal{F} = \{f_w : \mathcal{X} \to \mathcal{Y} | w \in \mathcal{W}\}$
- Learning algorithm: $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$ by $P_{W|S}$
- Loss: $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$
- We're interested in
  - Population risk: $L_\mu(w) \triangleq \mathbb{E}_{Z\sim\mu}[\ell(w, Z)]$; Expected population risk: $L_\mu = \mathbb{E}_W [L_\mu(W)]$
  - Empirical risk: $L_S(w) \triangleq \frac{1}{n} \sum_{i=1}^n \ell(w, Z_i)$; Expected empirical risk: $L_n = \mathbb{E}_{W,S} [L_S(W)]$
  - Expected generalization error: $\mathrm{Err} \triangleq L_\mu - L_n = \mathbb{E}_{W,S}[L_\mu(W) - L_S(W)]$

First MI Bound

Lemma (Xu and Raginsky [2017])

Assume the loss $\ell(w, Z)$ is R-subgaussian[1] for any $w \in \mathcal{W}$. The generalization error of $\mathcal{A}$ is bounded by

$$|\mathrm{Err}| \leq \sqrt{\frac{2R^2}{n} I(W; S)}.$$

[1] A random variable X is R-subgaussian if for any ρ, $\log \mathbb{E} \exp\left(\rho\left(X - \mathbb{E}X\right)\right) \leq \rho^2 R^2/2$.

First MI Bound

Lemma (Xu and Raginsky [2017])

Assume the loss $\ell(w, Z)$ is R-subgaussian[1] for any $w \in \mathcal{W}$. The generalization error of $\mathcal{A}$ is bounded by

$$|\mathrm{Err}| \leq \sqrt{\frac{2R^2}{n} I(W;S)}.$$

Mutual information $I(W;S) \triangleq \mathrm{D_{KL}}(P_{W,S}||P_W \otimes P_S)$.

$\Longrightarrow$ Distribution-dependent and Algorithm-dependent

[1] A random variable X is R-subgaussian if for any ρ, $\log \mathbb{E} \exp\left(\rho\left(X - \mathbb{E}X\right)\right) \leq \rho^2 R^2/2$.

First MI Bound

Lemma (Xu and Raginsky [2017])

Assume the loss $\ell(w, Z)$ is R-subgaussian[1] for any $w \in \mathcal{W}$. The generalization error of $\mathcal{A}$ is bounded by

$$|\mathrm{Err}| \leq \sqrt{\frac{2R^2}{n} I(W; S)}.$$

Mutual information $I(W; S) \triangleq \mathrm{D_{KL}}(P_{W,S} || P_W \otimes P_S)$.

$\Longrightarrow$ Distribution-dependent and Algorithm-dependent

Problem: $I(W; S) = H(W) - H(W|S) \to \infty$ in some cases

[1] A random variable X is R-subgaussian if for any ρ, $\log \mathbb{E} \exp\left(\rho\left(X - \mathbb{E}X\right)\right) \leq \rho^2 R^2/2$.

Supersample Setting

Supersample $\widetilde{Z} \overset{U}{\Longrightarrow} S = \widetilde{Z}_U = \{\widetilde{Z}_{i,U_i}\}_{i=1}^n$:

$$
\begin{array}{|c|c|}
\hline
\widetilde{Z}_{1,0} & \widetilde{Z}_{1,1} \\
\widetilde{Z}_{2,0} & \widetilde{Z}_{2,1} \\
\vdots & \vdots \\
\widetilde{Z}_{n,0} & \widetilde{Z}_{n,1} \\
\hline
\end{array}
\overset{U}{\Longrightarrow}
\begin{array}{|c|}
\hline
\widetilde{Z}_{1,U_1} \\
\widetilde{Z}_{2,U_2} \\
\vdots \\
\widetilde{Z}_{n,U_n} \\
\hline
\end{array}
$$

where $U = (U_1, U_2, \ldots, U_n)^T \sim \mathrm{Unif}(\{0,1\}^n)$.

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

CMI Bounds

Lemma (Steinke and Zakynthinou [2020])

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

$$|\mathrm{Err}| \leq \sqrt{\frac{2I(W;U|\widetilde{Z})}{n}}.$$

Nice property: $I(W;U|\widetilde{Z}) \leq H(U) = n\ln 2 \implies$ **bounded upper bound.**

CMI, f-CMI and e-CMI

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

CMI, f-CMI and e-CMI

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

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

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

CMI, f-CMI and e-CMI

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

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

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

Generalization Bounds via Loss Difference

ICML International Conference On Machine Learning

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*

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

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

Estimate $I(W; Z_i)$ vs $I(\Delta L_i; U_i)$:

- W and Z_i are high-dimensional R.V.'s
- ΔL_i is an one-dimensional R.V. and U_i is a binary R.V. $\Longrightarrow$ Easy-to-Compute!

A Communication View of Generalization

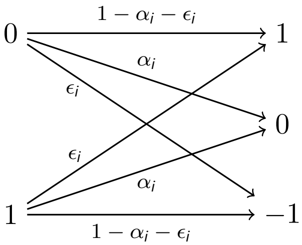


Figure: Channel from U_i to ΔL_i. Zero-one loss assumed.

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.

Generalization Bounds via Single Loss

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 $\varepsilon_i = (-1)^{\overline{U}_i}$.

Generalization Bounds via Single Loss

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 $\varepsilon_i = (-1)^{\overline{U}_i}$.

Recall that $\mathfrak{R}_n(\mathcal{W}) \triangleq \mathbb{E}_S\mathbb{E}_{\varepsilon_{1:n}}\left[\sup_{w\in\mathcal{W}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\ell(w,Z_i)\right] \implies \mathrm{Err} \leq 2\mathfrak{R}_n(\mathcal{W})$.

Generalization Bounds via Single Loss

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 $\varepsilon_i = (-1)^{\overline{U}_i}$.
Recall that $\mathfrak{R}_n(\mathcal{W}) \triangleq \mathbb{E}_S\mathbb{E}_{\varepsilon_{1:n}}\left[\sup_{w\in\mathcal{W}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\ell(w,Z_i)\right] \implies \mathrm{Err} \leq 2\mathfrak{R}_n(\mathcal{W})$.

Theorem

Assume $\ell(\cdot,\cdot)\in[0,1]$, *we have*

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

Bounds only depend on a single column of $\widetilde{Z}$; Still easy-to-compute.

Fast-Rate MI Bound

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

Fast-Rate MI Bound

Consider the weighted generalization error, $\mathrm{Err}_{C_1} \triangleq L_\mu - (1+C_1)L_n$.
$\Longrightarrow$ widely used in the PAC-Bayes literature.

Fast-Rate MI Bound

Consider the weighted generalization error, $\mathrm{Err}_{C_1} \triangleq L_\mu - (1 + C_1)L_n$.
$\Longrightarrow$ widely used in the PAC-Bayes literature.

Lemma

The weighted generalization error can be rewritten as

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

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

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

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

Faster Rate than Square-Root based Bound

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

Variance Based MI Bound

Inspired by [Seldin et al., 2012, Tolstikhin and Seldin, 2013],

Definition (γ-Variance)

For any $\gamma \in (0,1)$, γ-variance for a learning algorithm is defined as

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

Variance Based MI Bound

Inspired by [Seldin et al., 2012, Tolstikhin and Seldin, 2013],

Definition (γ-Variance)

For any $\gamma \in (0,1)$, γ-variance for a learning algorithm is defined as

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

Lemma

Under the <u>zero-one</u> loss assumption, we have $V(\gamma) = L_n - (1-\gamma^2)\mathbb{E}_{W,S}\left[L_S^2(W)\right]$.

Variance Based MI Bound

Lemma

For any $C_1 > 0$, *we have* $\mathrm{Err} - C_1 V(\gamma) \leq \frac{2+C_1\gamma^2}{n}\sum_{i=1}^n \mathbb{E}_{L_i^+,\tilde{\varepsilon}_i}\left[\tilde{\varepsilon}_i L_i^+\right]$, *where* $\tilde{\varepsilon}_i = \varepsilon_i - \frac{C_1\gamma^2}{C_1\gamma^2+2}$ *is the shifted Rademacher variable with mean* $-\frac{C_1\gamma^2}{C_1\gamma^2+2}$.

Variance Based MI Bound

ICML International Conference On Machine Learning

Lemma

For any $C_1 > 0$*, we have* $\mathrm{Err} - C_1 V(\gamma) \le \frac{2+C_1\gamma^2}{n}\sum_{i=1}^{n}\mathbb{E}_{L_i^+,\tilde{\varepsilon}_i}\left[\tilde{\varepsilon}_i L_i^+\right]$*, where* $\tilde{\varepsilon}_i = \varepsilon_i - \frac{C_1\gamma^2}{C_1\gamma^2+2}$ *is the shifted Rademacher variable with mean* $-\frac{C_1\gamma^2}{C_1\gamma^2+2}$.

Theorem

Assume $\ell(\cdot,\cdot) \in \{0,1\}$, $\gamma \in (0,1)$*. Then, there exist* $C_1, C_2 > 0$ *such that*

$$\mathrm{Err} \le C_1 V(\gamma) + \sum_{i=1}^{n}\frac{I(L_i^+;U_i)}{nC_2}. \tag{5}$$

Variance Based MI Bound

Compared with previous fast-rate bound:

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

$$\mathrm{Err} \leq C_1 V(\gamma) + \sum_{i=1}^{n} \frac{I(L_i^+; U_i)}{n C_2}$$

$$\Longrightarrow L_\mu \leq (1 + C_1) L_n - C_1 (1 - \gamma^2) \mathbb{E}_{W,S}\left[L_S^2(W)\right] + \sum_{i=1}^{n} \frac{I(L_i^+; U_i)}{C_2 n}.$$

Variance Based MI Bound

Compared with previous fast-rate bound:

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

$$\mathrm{Err} \leq C_1 V(\gamma) + \sum_{i=1}^{n} \frac{I(L_i^+;U_i)}{nC_2}$$

$$\implies L_\mu \leq (1+C_1)L_n - C_1(1-\gamma^2)\mathbb{E}_{W,S}\left[L_S^2(W)\right] + \sum_{i=1}^{n} \frac{I(L_i^+;U_i)}{C_2 n}.$$

- $L_n = 0 \rightarrow V(\gamma) = 0$, but $L_n = 0 \not\leftarrow V(\gamma) = 0$;
- For the fixed C_1 and C_2, variance-based bound is tighter than the previous bound with the gap being at least $C_1(1-\gamma^2)\mathbb{E}_{W,S}\left[L_S^2(W)\right]$.

Sharpness Based MI Bound

Inspired by Yang et al. [2019],

Definition (λ-Sharpness)

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

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

Sharpness Based MI Bound

Inspired by Yang et al. [2019],

Definition (λ-Sharpness)

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

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

Lemma

Assume $\ell(\cdot,\cdot) \in \{0,1\}$, we have $F_i(\lambda) = \mathbb{E}_{W,Z_i}\left[\ell(W,Z_i)\right] - (1-\lambda^2)\mathbb{E}_{Z_i}\left[\mathbb{E}^2_{W|Z_i}\ell(W,Z_i)\right]$.

Sharpness Based MI Bound

Lemma

Let $F(\lambda) = \frac{1}{n}\sum_{i=1}^{n} F_i(\lambda)$. *For any* $C_1 > 0$, *we have*

$$\mathrm{Err} - C_1 F(\lambda) = \frac{C_1+2}{n}\sum_{i=1}^{n}\mathbb{E}_{L_i^+,U_i}\left[\tilde{\varepsilon}_i L_i^+ - \frac{C_1(1-\lambda^2)}{C_1+2}\hat{\varepsilon}_i h(U_i)\right],$$

where $\tilde{\varepsilon}_i = \varepsilon_i - \frac{C_1}{C_1+2}$ *and* $\hat{\varepsilon}_i = \varepsilon_i - 1$ *are the shifted Rademacher variables, and* $h(U_i) = \mathbb{E}_{\widetilde{Z}_i^+|U_i}\left[\mathbb{E}^2_{L_i^+|\widetilde{Z}_i^+,U_i} L_i^+\right]$.

Sharpness Based MI Bound

Theorem

Assume $\ell(\cdot,\cdot) \in \{0,1\}$, $\lambda \in (0,1)$. *Then, there exist* $C_1, C_2 > 0$ *such that*

$$\mathrm{Err} \leq C_1 F(\lambda) + \sum_{i=1}^{n} \frac{I(L_i^+; U_i)}{C_2 n}. \tag{6}$$

Sharpness Based MI Bound

Theorem

Assume $\ell(\cdot,\cdot) \in \{0,1\}$, $\lambda \in (0,1)$. *Then, there exist* $C_1, C_2 > 0$ *such that*

$$\mathrm{Err} \leq C_1 F(\lambda) + \sum_{i=1}^{n} \frac{I(L_i^+; U_i)}{C_2 n}. \tag{6}$$

- $L_n = 0 \to F(\lambda) = 0$, but $L_n = 0 \not\leftarrow F(\lambda) = 0$;
- Sharpness bound can be further bounded:
  $L_\mu \leq (1 + C_1)L_n - C_1(1 - \lambda^2)L_n^2 + \sum_{i=1}^{n} \frac{I(L_i^+;U_i)}{C_2 n}$.
  For any fixed C_1 and C_2, sharpness based bound is tighter than the previous fast-rate bound.

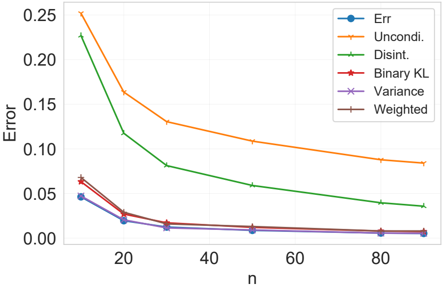
Experiments

We will compare

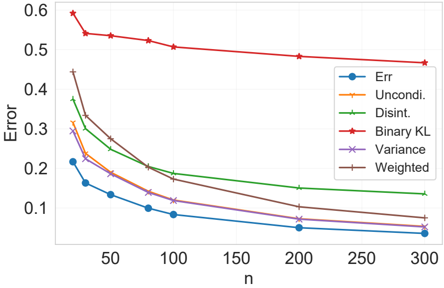
- Uncondi.: $\frac{1}{n}\sum_{i=1}^{n}\sqrt{2I(\Delta L_i; U_i)}$
- Disint.: $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\widetilde{Z}}\sqrt{2I^{\widetilde{Z}}(\Delta L_i; U_i)}$
- Binary KL: Hellström and Durisi [2022]
- Weighted: $\sum_{i=1}^{n}\frac{4I(L_i^+;U_i)}{n} + 4\sqrt{\sum_{i=1}^{n}\frac{L_n I(L_i^+;U_i)}{n}}$
- Variance: $C_1 V(\gamma) + \sum_{i=1}^{n}\frac{I(L_i^+;U_i)}{nC_2}$
- Sharpness: $C_1 F(\lambda) + \sum_{i=1}^{n}\frac{I(L_i^+;U_i)}{C_2 n}$

Experiments on Synthetic Gaussian Dataset

ICML International Conference On Machine Learning



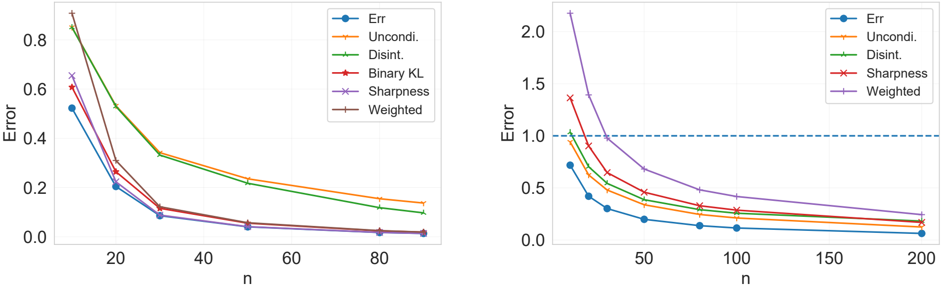
(a) $|\mathcal{Y}| = 2$ (Realizable)



(b) $|\mathcal{Y}| = 2$ (Non-Separable)

Figure: Comparison of bounds on the binary classification task with linear classifier. (a) Binary classification with a separable μ. (b) Binary classification with a non-separable μ.

Experiments on Synthetic Gaussian Dataset



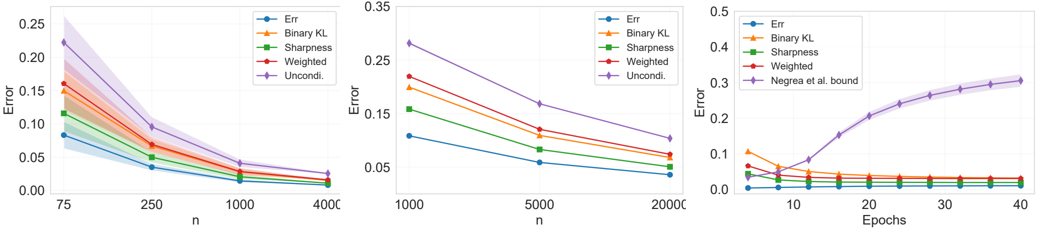
(a) $|\mathcal{Y}| = 10$ (Realizable)

(b) $|\mathcal{Y}| = 10$ (Non-Separable)

Figure: Comparison of bounds on the ten-class classification task with linear classifier. (a) Ten-class classification with a separable μ. (b) Ten-class classification with a non-separable μ.

Experiments on Real datasets

ICML
International Conference
On Machine Learning



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

Figure: Comparison of bounds on two real datasets, MNIST (“4 vs 9”) and CIFAR10.

References I

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov): 463–482, 2002.

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

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.

References II

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

Yevgeny Seldin, François Laviolette, Nicolo Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. Pac-bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012.

Ilya O Tolstikhin and Yevgeny Seldin. Pac-bayes-empirical-bernstein inequality. *Advances in Neural Information Processing Systems*, 26, 2013.

Jun Yang, Shengyang Sun, and Daniel M Roy. Fast-rate pac-bayes generalization bounds via shifted rademacher processes. *Advances in Neural Information Processing Systems*, 32, 2019.

The End