Exploring Generalization in Machine Learning through Information-Theoretic Lens

IMA Annual Workshop

Ziqiao Wang
Under the Supervision of
Prof. Yongyi Mao

University of Ottawa
School of Electrical Engineering and Computer Science

August 5, 2023

Contents

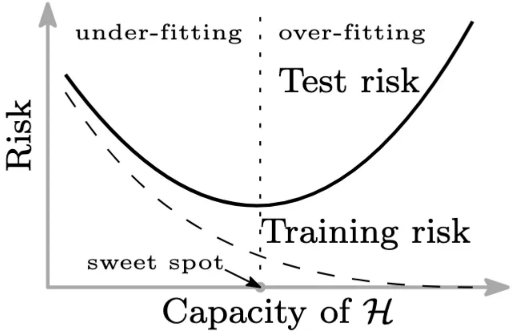
Motivation

Motivation

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

Motivation

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



Classical Viewpoint of Generalization

Motivation

- Generalization measures (e.g., VC-dim and Rademacher complexity) in classical statistical learning theory cannot explain the success of modern deep neural networks [Zhang et al., 2017].
  \# of parameters > # of training data & can even perfectly fit random labels
  $\Longrightarrow$ high capacity
  $\Longrightarrow$ still perform well on unseen data

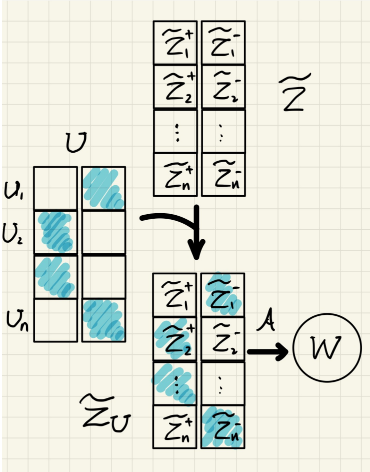
Motivation

- Generalization measures (e.g., VC-dim and Rademacher complexity) in classical statistical learning theory cannot explain the success of modern deep neural networks [Zhang et al., 2017].
  \# of parameters > # of training data & can even perfectly fit random labels
  $\Longrightarrow$ high capacity
  $\Longrightarrow$ still perform well on unseen data
- Algorithm & Distribution-dependent $\Longrightarrow$ non-vacuous generalization bound

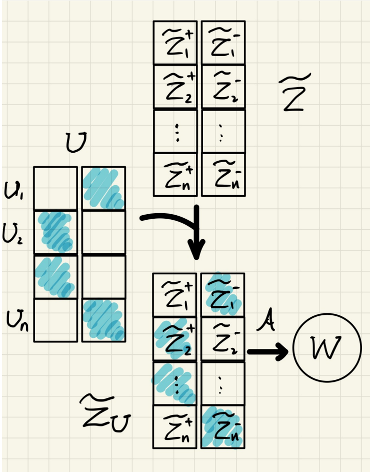
Tighter Information-Theoretic Generalization Bounds from Supersamples (*ICML'23*)

- 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
- Let $U = (U_1, U_2, \ldots, U_n)^T \sim \text{Unif}(\{0,1\}^n)$.
- Learning algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$
- $\text{Err} \triangleq \mathbb{E}_{W,S}\left[\mathbb{E}_{Z\sim\mu}[\ell(w,Z)] - \frac{1}{n}\sum_{i=1}^{n}\ell(w,Z_i)\right]$



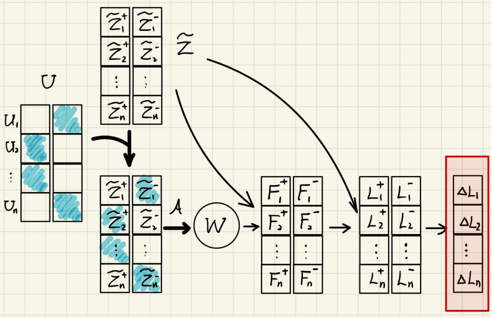
- Let $\widetilde{Z}$ drawn i.i.d. from μ
- Let $U = (U_1, U_2, \ldots, U_n)^T \sim \mathrm{Unif}(\{0,1\}^n)$.
- Learning algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$
- $\mathrm{Err} \triangleq \mathbb{E}_{W,S}\left[\mathbb{E}_{Z\sim\mu}[\ell(w,Z)] - \frac{1}{n}\sum_{i=1}^{n}\ell(w,Z_i)\right]$

Lemma 1 (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



- $F_i^+ := f_W(\widetilde{X}_i^+)$, $F_i^- := f_W(\widetilde{X}_i^-)$, $F_i := (F_i^+, F_i^-)$ $\Rightarrow$ **f-CMI Bound**: $|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{I(F_i; U_i|\widetilde{Z})}$ [Harutyunyan et al., 2021]
- $L_i^+ := \ell(W, \widetilde{Z}_i^+)$, $L_i^- := \ell(W, \widetilde{Z}_i^-)$, $L_i := (L_i^+, L_i^-)$ $\Rightarrow$ **e-CMI Bound**: $|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{I(L_i; U_i|\widetilde{Z})}$ [Hellström and Durisi, 2022]
- This paper: $\Delta L_i := L_i^- - L_i^+$ $\Rightarrow$ **ld-CMI**: $I(\Delta L_i; U_i|\widetilde{Z})$

Theorem 1

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

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

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

Generalization Bounds via Loss Difference

Theorem 1

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

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

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

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!

Theorem 2

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

Theorem 2

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

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

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

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

Theorem 3

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

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

Sharpness Based MI Bound

Theorem 3

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

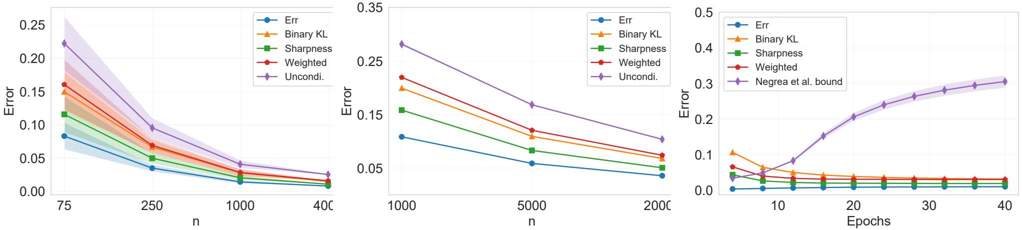
$$F_i(\lambda) \triangleq \mathbb{E}_{W,Z_i}\left[\ell(W,Z_i) - (1+\lambda)\mathbb{E}_{W|Z_i}\ell(W,Z_i)\right]^2.$$

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

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

- $L_n = 0 \to F(\lambda) = 0$, but $L_n = 0 \not\leftarrow F(\lambda) = 0$;
- 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)

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: $C_1F(\lambda)+\sum_{i=1}^{n}\frac{I(L_i^+;U_i)}{C_2 n}$.

On the Generalization of Models Trained with SGD: Information-Theoretic Bounds and Implications

(*ICLR'22*)

Main Contributions

- New information-theoretic upper bounds for the generalization error of machine learning models trained with SGD
- New and simple regularization scheme

Main Result

Theorem 4

The generalization error of SGD is upper bounded by

$$\mathrm{Err} \leq \mathcal{O}\left(\sqrt[3]{\sum_{t=1}^{T} \frac{\mathbb{E}\left[\mathbb{V}_t(W_{t-1})\right]\mathbb{E}\left[\mathrm{Tr}\left(\mathrm{H}_{W_T}(Z)\right)\right]}{n}}\right) \tag{6}$$

- Gradient Dispersion: $\mathbb{V}_t(w) \triangleq \mathbb{E}_S\left[||g(w, B_t) - \mathbb{E}_{W,Z}\left[\nabla_w \ell(W, Z)\right]||_2^2\right]$

Implication: Gaussian Model Perturbation (GMP)

- We hope the empirical risk surface at w^* is flat, or insensitive to a small perturbation of w^*.

$$\min_w L_s(w) + \rho \mathbb{E}_{\Delta \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)} [L_s(w + \Delta) - L_s(w)],$$

  where ρ is a hyper-parameter.

- Replacing the expectation above with its stochastic approximation using k realizations of Δ gives rise to the following optimization problem.

$$\min_w \frac{1}{b} \sum_{z \in B} \left((1 - \rho)\ell(w, z) + \rho \frac{1}{k} \sum_{i=1}^{k} (\ell(w + \delta_i, z)) \right).$$

Implication: GMP on VGG16

Method	SVHN	CIFAR-10	CIFAR-100
ERM	96.86±0.060	93.68±0.193	72.16±0.297
Dropout	97.04±0.049	93.78±0.147	72.28±0.337
L. S.	96.93±0.070	93.71±0.158	72.51±0.179
Flooding	96.85±0.085	93.74±0.145	72.07±0.271
MixUp	96.91±0.057	**94.52±0.112**	73.19±0.254
Adv. Tr.	97.06±0.091	93.51±0.130	70.88±0.145
AMP[1]	**97.27±0.015**	94.35±0.147	74.40±0.168
GMP^3	<u>97.18±0.057</u>	94.33±0.094	<u>74.45±0.256</u>
GMP^{10}	97.09±0.068	<u>94.45±0.158</u>	**75.09±0.285**

Top-1 classification accuracy acc.(%) of VGG16. We run experiments 10 times and report the mean and the standard deviation of the testing accuracy.

[1] $\min_w L_s(w) + \rho \max_\delta L_s(w+\delta) - L_s(w)$

Information-Theoretic Analysis of Unsupervised Domain Adaptation (*ICLR'23*)

Main Contributions

- Novel upper bounds for generalization error of UDA.
- Simple regularization technique for improving generalization of UDA

Additional Notations

- Source data $Z = (X, Y) \sim \mu$ and target data $Z' = (X', Y') \sim \mu'$
- Labeled source sample: $S = \{Z_i\}_{i=1}^{n} \overset{\text{i.i.d}}{\sim} \mu^{\otimes n}$; Unlabelled target sample $S'_{X'} = \{X'_j\}_{j=1}^{m} \overset{\text{i.i.d}}{\sim} P_{X'}^{\otimes m}$
- *Generalization error = testing error of target domain - training error of source domain:*

$$\mathrm{Err} = \mathbb{E}_{W,S,S'_{X'}} \left[R_{\mu'}(W) - R_S(W) \right]$$

MI Bound for EP

Theorem 5

Assume $\ell(f_w(X'), Y')$ is R-subgaussian. Then

$$|\mathrm{Err}| \leq \frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{X'_j}\sqrt{2R^2 I^{X'_j}(W;Z_i)} + \sqrt{2R^2 \mathrm{D}_{\mathrm{KL}}(\mu||\mu')}.$$

Gradient Penalty as an Universal Regularizer

Consider SGLD. At each time step t,

- labelled source mini-batch: Z_{B_t}; unlabelled target mini-batch: X'_{B_t}
- gradient: $G_t = g(W_{t-1}, Z_{B_t}, X'_{B_t})$
- updating rule: $W_t = W_{t-1} - \eta_t G_t + N_t$ where $N_t \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$.

Theorem 6

Under the assumption of Theorem 5. Let the total iteration number be T, then

$$|\mathrm{Err}| \leq \sqrt{\frac{R^2}{n} \sum_{t=1}^{T} \frac{\eta_t^2}{\sigma_t^2} \mathbb{E}_{S'_{X'}, W_{t-1}, S} \left[\left|\left| G_t - \mathbb{E}_{Z_{B_t}} [G_t] \right|\right|^2 \right]} + \sqrt{2R^2 \mathrm{D}_{\mathrm{KL}}(\mu || \mu')}.$$

restrict the gradient norm $\Longrightarrow$ reduce $|\mathrm{Err}|$.

Experimental Results: RotatedMNIST

RotatedMNIST is built based on the MNIST dataset and consists of six domains, which are rotated MNIST images with rotation angle $0°, 15°, 30°, 45°, 60°$ and $75°$.

RotatedMNIST.

	RotatedMNIST (**0°** as source domain)					
Method	**15°**	**30°**	**45°**	**60°**	**75°**	**Ave**
ERM	97.5±0.2	84.1±0.8	53.9±0.7	34.2±0.4	22.3±0.5	58.4
DANN	97.3±0.4	90.6±1.1	68.7±4.2	30.8±0.6	19.0±0.6	61.3
MMD	97.5±0.1	95.3±0.4	73.6±2.1	44.2±1.8	32.1±2.1	68.6
CORAL	97.1±0.3	82.3±0.3	56.0±2.4	30.8±0.2	27.1±1.7	58.7
WD	96.7±0.3	93.1±1.2	64.1±3.3	41.4±7.6	27.6±2.0	64.6
KL	97.8±0.1	97.1±0.2	93.4±0.8	75.5±2.4	68.1±1.8	86.4
ERM-GP	97.5±0.1	86.2±0.5	62.0±1.9	34.8±2.1	26.1±1.2	61.2
KL-GP	98.2±0.2	96.9±0.1	95.0±0.6	**88.0±8.1**	**78.1±2.5**	**91.2**

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 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.

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!