Rethinking Information-theoretic Generalization: Loss Entropy Induced PAC Bounds

Yuxin Dong

School of Computer Science and Technology
Xi'an Jiaotong University

April, 2024

Problem Setting

- Dataset: $S = \{Z_i\}_{i=1}^n \in \mathcal{Z}^n$, sampled i.i.d from μ.
  - e.g. regression: $Z_i = (X_i, Y_i)$, $X_i \in \mathbb{R}^m$, $Y_i \in \mathbb{R}$.
- Hypothesis: $W \in \mathcal{W}$.
  - e.g. neural networks: $\mathcal{W} \subset \mathbb{R}^d$, d: number of tunable parameters.
- Loss function: $\ell : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}^+$.
  - e.g. square loss: $\ell(w, z) = (f_w(x) - y)^2$.
- Learning algorithm: $\mathcal{A} : \mathcal{Z}^n \mapsto \mathcal{W}$.
  - e.g. stochastic gradient descent (SGD).

Definition of Generalization Error

- Population risk (test loss) $L(w)$:
  - $L(w) = \mathbb{E}_Z[\ell(w, Z)]$.
- Empirical risk (training loss) $L_S(w)$:
  - $L_S(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i)$.
- Population risk decomposition for $W = \mathcal{A}(S)$:
  - $L(W) = L_S(W) + \underbrace{(L(W) - L_S(W))}_{\Delta(W,S):\ \text{Generalization Error}}.$

Generalization Analysis Techniques

- Uniform convergence: $\sup_{w \in \mathcal{W}}\{L(W) - L_S(W)\}$.
  - Distribution-agnostic: VC-dimension.
  - Distribution-dependent: Rademacher complexity.

- Algorithm-dependent techniques:
  - Algorithm stability [Hardt et al., 2016]: How does the learning algorithm respond to input perturbations?
  - Information theory [Xu and Raginsky, 2017]: How much information is captured by the learning algorithm?

Comparison of Different Techniques

Method	Algorithm Stability	Information Theory
Assumptions	Lipschitz Condition Smoothness (Strong) Convexity	Subgaussian (Bounded) Interpolating Regime
Convergence Rate	Non-convex: $O(\frac{1}{\sqrt{n}})$ Convex: $O(\frac{1}{n})$	General: $O(\frac{1}{\sqrt{n}})$ Interpolating: $O(\frac{1}{n})$
Tractability	Not computable	Computable

1. Introduction
2. **Data-independent Bounds**
3. Data-dependent Bounds
4. Proof Sketch
5. References

Generalization by Compressed Representation

For some fixed $w \in \mathcal{W}$, with probability at least $1 - \delta$ over the draw of S:

- [Shwartz-Ziv et al., 2018]:
  - $\Delta(w, S) \leq \sqrt{\frac{2^{I(X,T)} + \log\left(\frac{1}{\delta}\right)}{2n}}$.
- [Kawaguchi et al., 2023]:
  - $\Delta(w, S) \leq O\left(\sqrt{\frac{I(X;T|Y) + \log\left(\frac{1}{\delta}\right)}{n}}\right)$.

Existing Problems

- $I(X;T)$ can be infinite in some cases.
  - e.g. invertible encoder with continuous input: $f_w^{-1} : \mathcal{T} \mapsto \mathcal{X}$, such that $f_w^{-1}(f_w(X)) = X$.
  - Workaround: Assume discrete inputs; use lossy activations.
- $I(X;T)$ is generally hard to estimate.
  - Both X and T are high-dimensional variables.
  - Workaround: Monte-Carlo sampling-based estimators; the reparameterization trick.

Motivation

Deeper representations are highly compressed.

$$
\begin{tikzcd}
Y \arrow[d] \arrow[rrrr] & & & & \\
X \arrow[r] & T_1 \arrow[r] & \cdots \arrow[r] & T_l \arrow[r] & L
\end{tikzcd}
$$

The loss is basically the last-layer representation.

- $I(X; T_1|Y) \geq \cdots \geq I(X; T_l|Y) \geq I(X; L|Y)$.
- For deterministic networks: $H(L|X, Y) = 0$.
- $H(L|Y) = H(L|Y) - H(L|X, Y) = I(X; L|Y) \leq I(X; T|Y)$.

Our Results - Loss Entropy

Theorem 1

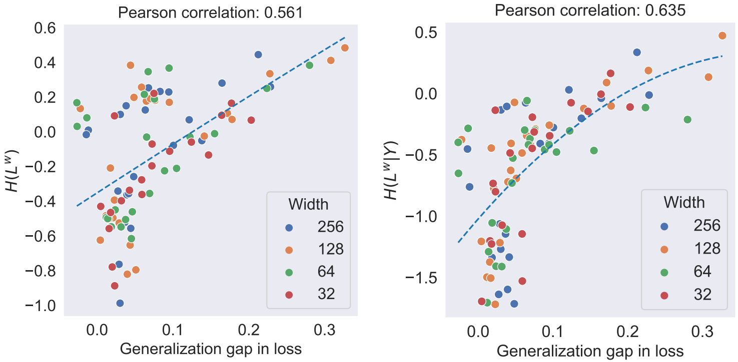
For some fixed $w \in \mathcal{W}$, with probability at least $1-\delta$ over the draw of S:

$$\Delta(w, S) \leq O\left(\sqrt{\frac{H(L|Y) + \log\left(\frac{1}{\delta}\right)}{n}}\right).$$

- $H(L|Y)$ / $H(L)$ is computationally tractable:
    - L is 1-dimensional.
    - Y is discrete or low-dimensional.

Experimental Results

Correlation between $H(L)$ / $H(L|Y)$ and the generalization gap.



Experimental Results

Pearson correlation analysis between the generalization error and different information metrics.

Metric	Correlation
Num. params.	-0.0294
$\|W\|_F$	-0.0871
$I(X; T^w)$	0.3712
$I(X; T^w\|Y)$	0.3842
$I(S; W)$	0.0211
$I(S; W) + I(X; T^w)$	0.3928
$I(S; W) + I(X; T^w\|Y)$	0.4130
$H(L^w)$	$\underline{0.5611}$
$H(L^w\|Y)$	**0.6350**

Connecting Loss and Error

Definition of error (regression / binary classification):

$$E = Y - f_w(X).$$

- One-step loss functions: $L = \phi(E)$
  - e.g. square loss $L = E^2$, absolute loss $L = |E|$.
  - Markov chain: $(X, Y) - E - L$.
  - Data processing inequality: $H(L) \leq H(E)$.
- Loss functions rely on Y: $L = \phi(Yf_w(X))$
  - e.g. cross-entropy, margin-based loss.
  - Markov chain (conditioned on Y): $X - f_w(X) - E - L$.
  - Conditional data processing inequality: $H(L|Y) \leq H(E|Y)$.

Connecting Loss and Error

Corollary 2

For some fixed $w \in \mathcal{W}$, with probability at least $1-\delta$ over the draw of S:

$$\Delta(w, S) \leq O\left(\sqrt{\frac{H(E)+\log\left(\frac{1}{\delta}\right)}{n}}\right).$$

Minimum Error Entropy (MEE) enhances generalization!

Data-independent to Data-dependent

Acquire $W = \mathcal{A}(S)$, with probability at least $1 - \delta$ over the draw of S and W:

$$\sqrt{\frac{H(L) + \log\left(\frac{1}{\delta}\right)}{n}} \xrightarrow[\text{union bound}]{w \in \mathcal{W}} \underbrace{\sqrt{\frac{H(L) + \log(|\mathcal{W}|) + \log\left(\frac{1}{\delta}\right)}{n}}}_{\textbf{Vacuous!}}.$$

Idea: Use the complexity of losses instead of the **hypothesis**.

Problem Setting

The supersample setting [Steinke and Zakynthinou, 2020]:

- Dataset: $\widetilde{S} = \{\widetilde{Z}_i\}_{i=1}^n \in \mathcal{Z}^{n\times 2}$, $\widetilde{Z}_i = \{\widetilde{Z}_i^0, \widetilde{Z}_i^1\}$.
- Dataset separation: $U = \{U_i\}_{i=1}^n \sim \mathrm{Unif}(\{0,1\}^n)$.
    - Training set: $\widetilde{S}_U = \{\widetilde{Z}_i^{U_i}\}_{i=1}^n$, test set: $\widetilde{S}_{\overline{U}} = \{\widetilde{Z}_i^{\overline{U}_i}\}_{i=1}^n$.
- Hypothesis: $W = \mathcal{A}(\widetilde{S}_U)$.
- Loss evaluation: $L_i^0 = \ell(W, \widetilde{Z}_i^0)$, $L_i^1 = \ell(W, \widetilde{Z}_i^1)$.
- Validation error: $\Delta(W, \widetilde{S}) = L_{\widetilde{S}_{\overline{U}}}(W) - L_{\widetilde{S}_U}(W)$.

Types of Generalization Bounds

- Average-case bounds ✗
  - $\mathbb{E}_{W,\widetilde{S},U}\left[\Delta(W,\widetilde{S})\right] \leq \cdots$.
  - Characterize the expected generalization error.
  - Insufficient to analyze single training processes.

- High-probability bounds ✓
  - With high probability, $\Delta(W,\widetilde{S}) \leq \cdots$.
  - Characterize the distribution of generalization error.
  - Provide guarantees for single training processes.

High-probability Generalization Bounds

Acquire $W = \mathcal{A}(S)$, with probability at least $1-\delta$ over the draw of S, U and W:

- Functional CMI [Harutyunyan et al., 2021]:
  - $\Delta(W, \widetilde{S}) \leq \sqrt{\frac{8I(F; U|\widetilde{S}) + 16}{n\delta}}$.
  - $F = \{f_W(\widetilde{Z}_i^0), f_W(\widetilde{Z}_i^1)\}_{i=1}^n$.
- Evaluated CMI [Hellström and Durisi, 2022]:
  - $\Delta(W, \widetilde{S}) \leq \sqrt{\frac{2\imath(R; U|\widetilde{S}) + 2\log\left(\frac{\sqrt{n}}{\delta}\right)}{n-1}}$.
  - $R = \{L_i^0, L_i^1\}_{i=1}^n$, $\imath$: information density.

Existing Problem

- Only applies to bounded loss functions
  - Many loss functions (square loss, cross-entropy) are unbounded, and thus are not covered by existing results.

- Computational intractability
  - $I(F; U|\widetilde{S})$ contains high-dimensional variables.
  - $\imath(R; U|\widetilde{S})$ cannot be estimated empirically.

Our Results - Loss Entropy

Theorem 3

For any $\lambda \in (0,1)$, with probability at least $1-\delta$ over the draw of S, U and W:

$$\Delta(W, \widetilde{S}) \leq \sqrt{\frac{2\sum_{i=1}^{n}(\Delta L_i)^2}{n}} \sqrt{\frac{H_{1-\lambda}(R_\Delta) + \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right) + \log\left(\frac{2}{\delta}\right)}{n}}.$$

- $\Delta L_i = L_i^1 - L_i^0$, $R_\Delta = \{\Delta L_i\}_{i=1}^n$.
- When $\lambda \to 0$, Rényi's entropy satisfies subadditivity: $H(R_\Delta) \leq \sum_{i=1}^n H(\Delta L_i) \leq \sum_{i=1}^n H(L_i^0) + H(L_i^1)$.

Fast-rate Bounds for Bounded and Interpolating Case

Theorem 4

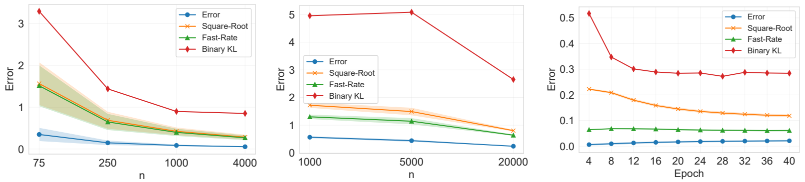
Assume $\ell(\cdot,\cdot) \in [0,\kappa]$ and $L_{\widetilde{S}_U}(W) = 0$. Then for any $\lambda \in (0,1)$, with probability at least $1-\delta$ over the draw of S, U and W:

$$\Delta(W,\widetilde{S}) \leq 2\kappa \frac{H_{1-\lambda}(R) + \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right) + \log\left(\frac{4}{\delta}\right)}{n\log 2}.$$

- Convergence rate: $O(1/\sqrt{n}) \to O(1/n)$.
- Similarly, by the subadditivity of entropy: $H(R) \leq \sum_{i=1}^{n} H(L_i^0, L_i^1) \leq \sum_{i=1}^{n} H(L_i^0) + H(L_i^1)$.

Experimental Results

Comparison between generalization bounds in 3 learning settings:
1. MNIST (Adam), 2. CIFAR10 (SGD), 3. MNIST (SGLD).



- Binary-KL: lower-bound of the currently tightest high-probability information-theoretic bound in the literature.

Main Idea

R is the "bottleneck" of information flow from W to Δ:

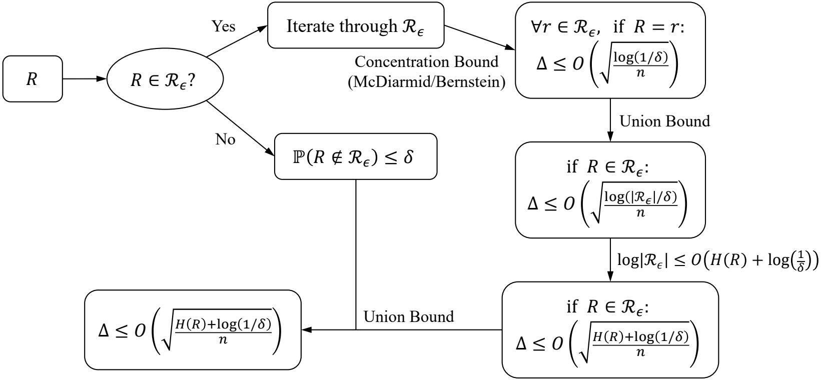
- Markov chain: $W \to R \to \Delta$.

Exhaustively explore every $R \in \mathcal{R}$ to decouple W and Δ.

Assume R is discrete, define the typical subset:

- $\mathcal{R}_\epsilon = \{r \in \mathcal{R} : -\log \mathbb{P}(R = r) - H(R) \leq \epsilon\}$.
- $\mathbb{P}(R \notin \mathcal{R}_\epsilon) \leq \delta$.
- $\log |\mathcal{R}_\epsilon| \leq O\left(H(R) + \log\left(\frac{1}{\delta}\right)\right)$.

Proof Sketch



Discretizing Continuous Losses

Most loss functions are continuous (square loss, cross-entropy), and require discretization before evaluating the bounds.

- Select bin size $b > 0$.
- Rounding function: $\phi_b(L) = b \times \arg\min_{i \in \mathbb{N}} |ib - L|$.
- Discretized loss: $\hat{L} = \phi_b(L + \xi)$, $\xi \sim \mathrm{Unif}([-\frac{b}{2}, \frac{b}{2}])$.

Lemma 5

Given test losses $L_1, \cdots, L_n$, with probability at least $1 - \delta$,

$$\frac{1}{n}\sum_{i=1}^{n} L_i - \frac{1}{n}\sum_{i=1}^{n} \hat{L}_i \leq b\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}}.$$

Future Works

- How to control test loss entropy?
- Reduce the required number of validation samples.
  - Leave-one-out settings and beyond.
- Other types of complexity measures.
  - Mutual information, Maximal leakage.

Thank You

[Hardt et al., 2016] Hardt, M., Recht, B., and Singer, Y. (2016).
Train faster, generalize better: Stability of stochastic gradient descent.
In *International Conference on Machine Learning*.

[Harutyunyan et al., 2021] Harutyunyan, H., Raginsky, M., Ver Steeg, G., and Galstyan, A. (2021).
Information-theoretic generalization bounds for black-box learning algorithms.
Advances in Neural Information Processing Systems.

[Hellström and Durisi, 2022] Hellström, F. and Durisi, G. (2022).
A new family of generalization bounds using samplewise evaluated cmi.
Advances in Neural Information Processing Systems.

[Kawaguchi et al., 2023] Kawaguchi, K., Deng, Z., Ji, X., and Huang, J. (2023).
How does information bottleneck help deep learning?
In *International Conference on Machine Learning*.

[Shwartz-Ziv et al., 2018] Shwartz-Ziv, R., Painsky, A., and Tishby, N. (2018).
Representation compression and generalization in deep neural networks.

[Steinke and Zakynthinou, 2020] Steinke, T. and Zakynthinou, L. (2020).
Reasoning about generalization via conditional mutual information.
In *Conference on Learning Theory*.

[Xu and Raginsky, 2017] Xu, A. and Raginsky, M. (2017).
Information-theoretic analysis of generalization capability of learning algorithms.
Advances in neural information processing systems.