Information-Theoretic Analysis for Generalization of Learning Algorithms

A Short Tutorial

uOttawa

Ziqiao Wang

University of Ottawa

School of Electrical Engineering and Computer Science

December 22, 2023

Contents

Preliminaries on Information Theory

Basic Measures

- Entropy: $H(X) = \mathbb{E}_{P_X}\left[\log \frac{1}{P(X)}\right]$, $H(X, Y) = \mathbb{E}_{P_{X,Y}}\left[\log \frac{1}{P(X,Y)}\right]$, $H(X|Y) = \mathbb{E}_{P_{X,Y}}\left[\log \frac{1}{P(X|Y)}\right]$
  - For discrete X, $H(X) \geq 0$
  - $H(X, Y) = H(X|Y) + H(Y)$
  - Conditioning reduces entropy: $H(X|Y) \leq H(X)$
  - For discrete X, $H(X) \leq \log|\mathcal{X}|$
- Relative Entropy: $\mathrm{D_{KL}}\left(Q||P\right) = \mathbb{E}_Q\left[\log \frac{Q(X)}{P(X)}\right]$
  - $\mathrm{D_{KL}}\left(Q||P\right) \geq 0$ with equality holds iff $Q = P$.
  - Usually $\mathrm{D_{KL}}\left(Q||P\right) \neq \mathrm{D_{KL}}\left(P||Q\right)$

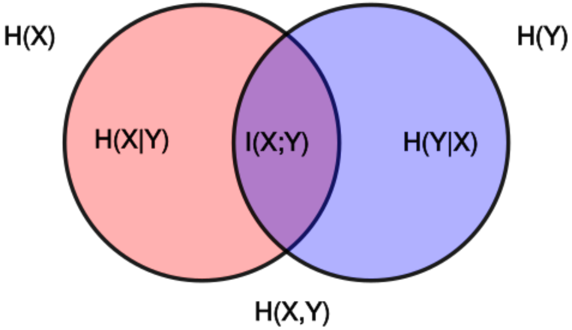
Basic Measures

- Mutual Information: $I(X;Y) = \mathbb{E}_{P_{X,Y}}\left[\log \frac{P(X,Y)}{P(X)P(Y)}\right] = \mathrm{D_{KL}}\left(P_{X,Y} || P_X P_Y\right)$.
  - $I(X;Y) \geq 0$ with equality holds iff $X \perp\!\!\!\perp Y$.
  - $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$.
  - $I(X;Y) = I(Y;X)$
  - $I(X;Y) = \mathbb{E}_{P_{X,Y}}\left[\log \frac{P(X|Y)}{P(X)}\right] = \mathbb{E}_{P_Y}\left[\mathrm{D_{KL}}\left(P_{X|Y} || P_X\right)\right]$
- Conditional Mutual Information and Disintegrated Mutual Information:

  $I(X;Y|Z) = \mathbb{E}_{P_{X,Y,Z}}\left[\log \frac{P(X,Y|Z)}{P(X|Z)P(Y|Z)}\right] = H(X|Z) - H(X|Y,Z)$

  $I^z(X;Y) = \mathbb{E}_{P_{X,Y|Z=z}}\left[\log \frac{P(X,Y|Z=z)}{P(X|Z=z)P(Y|Z=z)}\right]$
  - $\mathbb{E}_Z\left[I^Z(X;Y)\right] = I(X;Y|Z)$

Basic Measures



Venn diagram. Credit: https://en.wikipedia.org/wiki/Mutual_information

Useful Properties

- Chain-rule:
  - $H(X_1, X_2, \dots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \dots, X_1)$
  - $I(X_1, X_2, \dots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, \dots, X_1)$
  - $\mathrm{D_{KL}}\left(Q_{X,Y} || P_{X,Y}\right) = \mathrm{D_{KL}}\left(Q_X || P_X\right) + \mathrm{D_{KL}}\left(Q_{Y|X} || P_{Y|X}\right)$
- Data-processing inequality (DPI):
  If $X - Y - Z$ forms a Markov chain (i.e. $P_{X,Z|Y} = P_{X|Y} P_{Z|Y}$), then

  $$I(X; Y) \geq I(X; Z)$$

  e.g., $(X, Y) - f(X, Y) - Z$ is a Markov chain : $I(X, Y; Z) \leq I(f(X, Y); Z)$
- Other useful stuff: Fano's inequality, KL divergence between two Gaussian, Gaussian distribution maximizes the entropy over all distributions with the same variance, ...
- Textbook for beginners: *Thomas M. Cover and Joy A. Thomas. Elements of Information Theory, Wiley-Interscience, 2006.*

Lemma 1 (Variational Representation of Mutual Information)

For two random variables X and Y, we have

$$I(X;Y) = \inf_{Q} \mathbb{E}_{P_Y}\left[\mathrm{D}_{\mathrm{KL}}(P_{X|Y}||Q)\right],$$

where the infimum is achieved at $Q = P_X$.

Note that $I(X;Y) = \mathbb{E}_{P_Y}\left[\mathrm{D}_{\mathrm{KL}}\left(P_{X|Y}||P_X\right)\right]$

Lemma 2 (Donsker and Varadhan's variational formula)

For any measurable function $f : \Theta \to \mathbb{R}$, we have

$$\mathrm{D}_{\mathrm{KL}}(Q||P) = \sup_f \mathbb{E}_{\theta\sim Q}\left[f(\theta)\right] - \log \mathbb{E}_{\theta\sim P}\left[\exp f(\theta)\right].$$

proof. Define the density of the Gibbs measure P_f: $P_f(\theta) = \frac{e^{f(\theta)}}{\mathbb{E}_{\theta\sim P}\left[e^{f(\theta)}\right]} P(\theta)$.

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(Q||P_f) = \mathbb{E}_Q\left[\log \frac{Q}{P_f}\right] &= \mathbb{E}_Q\left[\log Q\right] - \mathbb{E}_Q\left[\log \frac{e^{f(\theta)}}{\mathbb{E}_P\left[e^{f(\theta)}\right]} P\right] \\
&= \mathbb{E}_Q\left[\log Q\right] - \mathbb{E}_Q\left[f(\theta)\right] - \mathbb{E}_Q\left[\log P\right] + \log \mathbb{E}_P\left[e^{f(\theta)}\right] \\
&= \mathrm{D}_{\mathrm{KL}}(Q||P) - \mathbb{E}_{\theta\sim Q}\left[f(\theta)\right] - \log \mathbb{E}_{\theta\sim P}\left[\exp f(\theta)\right] \\
&\geq 0
\end{aligned}$$

Second Textbook

- *Polyanskiy, Y. and Wu, Y.. Information Theory: From Coding to Learning, Cambridge University Press, 2023 (book draft).*

Background on Information-Theoretic Generalization Bounds

Generalization

- A learning algorithm $\mathcal{A} : S \to W$ i.e. mapping training sample S to a hypothesis W.
- Gen. err. $=$ $\mathbb{E}$ [Test err. $-$ Train err.] $\leq$ Gen. bound.

Generalization

- A learning algorithm $\mathcal{A} : S \to W$ i.e. mapping training sample S to a hypothesis W.
- Gen. err. $= \mathbb{E}\left[\text{Test err.} - \text{Train err.}\right] \leq$ Gen. bound.

Formal Notations:

- Training dataset: $S = \{Z_i\}_{i=1}^n \in \mathcal{Z}$, drawn i.i.d. from μ
- Hypothesis space: $\mathcal{W} \subseteq \mathbb{R}^d$
- 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: $L_\mu(w) \triangleq \mathbb{E}_{Z\sim\mu}[\ell(w, Z)]$
  - Empirical risk: $L_S(w) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(w, Z_i)$
  - Expected generalization error: $\mathcal{E}_\mu(\mathcal{A}) \triangleq \mathbb{E}_{W,S}[L_\mu(W) - L_S(W)]$

"First" MI Bound

Before Xu's bound:

- *Russo, D. and Zou, J.. Controlling bias in adaptive data analysis using information theory. AISTATS 2016.*
  Russo, D., and Zou, J. How much does your data exploration overfit? Controlling bias via information usage. TIT 2019.
- *Raginsky, M. et al. Information-theoretic analysis of stability and bias of learning algorithms. ITW 2016.*

Lemma 3 (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

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

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

- **Step 1: Finding the target.**

$$\begin{aligned}\mathcal{E} = \mathbb{E}_{S,W}\left[L_\mu(W) - L_S(W)\right] &= \mathbb{E}_{S,W}\left[\mathbb{E}_{S'}\left[L_{S'}(W)\right]\right] - \mathbb{E}_{S,W}\left[L_S(W)\right] \\ &= \mathbb{E}_{P_W \otimes P_{S'}}\left[L_{S'}(W)\right] - \mathbb{E}_{P_{W,S}}\left[L_S(W)\right]\end{aligned}$$

- **Step 2: Selecting the measurable function f.**

Recall DV Lemma:

$$\begin{aligned} I(W,S) &= \mathrm{D}_{\mathrm{KL}}(P_{W,S} || P_W \otimes P_{S'}) \\ &\geq \sup_f \mathbb{E}_{(W,S)\sim P_{W,S}}\left[f(W,S)\right] - \log \mathbb{E}_{(W,S')\sim P_W \otimes P_{S'}}\left[\exp f(W,S')\right]\end{aligned}$$

Let $f(W,S) = tL_S(W)$ for some $t > 0$.

- **Step 3: Bounding the CGF.**
  If $\ell(w, Z)$ is R-SubGaussian, $f(w, S') = L_{S'}(w)$ is $R/\sqrt{n}$-SubGaussian:

$$\log \mathbb{E}_{W,S'} \left[\exp \lambda(L_{S'} - \mathbb{E}\left[L_{S'}\right])\right] \leq t^2 R^2/2n$$

  Thus, $\log \mathbb{E}_{W,S'} \left[\exp tL_{S'}(W)\right] \leq t\mathbb{E}_{W,S'} \left[L_{S'}(W)\right] + t^2 R^2/2n$.
- **Step 4: Optimizing the bound.**

$$I(W, S) \geq \sup_{t>0} t \left(\mathbb{E}_{(W,S)\sim P_{W,S}} \left[L_S(W)\right] - \mathbb{E}_{(W,S')\sim P_W \otimes P_{S'}} \left[L_{S'}(W)\right]\right) - t^2 R^2/2n$$

$\implies \mathbb{E}_{(W,S)\sim P_{W,S}} \left[L_S(W)\right] - \mathbb{E}_{(W,S')\sim P_W \otimes P_{S'}} \left[L_{S'}(W)\right] \leq \inf_t \frac{I(W,S)}{t} + \frac{tR^2}{2n} = \sqrt{\frac{2R^2}{n} I(W, S)}$

$\implies |\mathcal{E}| \leq \sqrt{\frac{2R^2}{n} I(W, S)}$

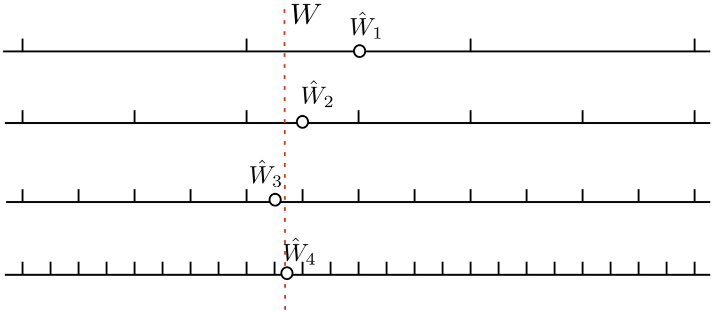
Limitations of Xu's bound

$I(W;S) \to \infty$ e.g., $\mathcal{A}$ is deterministic $\Longrightarrow I(W;S) = H(W) - H(W|S) = H(W)$.

Some previous efforts:

- **Chaining Method**: $3\sqrt{2}\sum_{k=k_1}^{\infty} 2^{-k}\sqrt{I([W]_k;Z_i)}$
  Asadi, A. et al. Chaining mutual information and tightening generalization bounds. NeurIPS 2018.
- **Individual Technique/Sample-Wise Bound**: $\frac{1}{n}\sum_{i=1}^{n}\sqrt{I(W;Z_i)}$
  Bu, Y. et al. Tightening Mutual Information Based Bounds on Generalization Error. ISIT 2019.
- **Random Subset Technique**: $\mathbb{E}\sqrt{\frac{1}{n-m}I^{S_J}(W;S_J^c)}$
  Negrea, J. et al. Information-theoretic generalization bounds for SGLD via data-dependent estimates. NeurIPS 2019.
- Solved by **CMI**: $\sqrt{\frac{1}{n}I(W;U|\widetilde{Z})} \leq \mathcal{O}(1)$
  Steinke, T. and Zakynthinou, L.. Reasoning about generalization via conditional mutual information. COLT 2020.

Idea:



Quantization of W. Credit: Zhou R, et al. Stochastic Chaining and Strengthened Information-Theoretic Generalization Bounds ISIT 2022.

Chaining Method

- **Step 1: Finding the target.** For any integers k_1 and k_0 such that $k_1 > k_0$, let $\mathcal{E}(W) = L_\mu(W) - L_S(W)$, we have

$$\mathcal{E}(W) = \mathcal{E}([W]_{k_0}) + \sum_{k=k_0+1}^{k_1} (\mathcal{E}([W]_k) - \mathcal{E}([W]_{k-1})) + \mathcal{E}(W) - \mathcal{E}([W]_{k_1}).$$

We require $\mathbb{E}[\mathcal{E}([W]_{k_0})] = 0$ and $\lim_{k_1 \to \infty} \mathcal{E}([W]_{k_1}) = \mathcal{E}(W)$.
Let $k_1 \to \infty$ and taking expectation over $(S, W) \sim P_{S,W}$ for both sides of the equation above, we have

$$\mathcal{E} = \sum_{k=k_0+1}^{\infty} \mathbb{E}_{S,[W]_k,[W]_{k-1}} [(\mathcal{E}([W]_k) - \mathcal{E}([W]_{k-1}))]. \tag{1}$$

- **Step 2: Selecting f, Q and P.**

$f = t \cdot (\mathcal{E}([W]_k) - \mathcal{E}([W]_{k-1}))\,, \quad Q = P_{S,[W]_k,[W]_{k-1}}, \quad P = P_S \otimes P_{[W]_k,[W]_{k-1}}$

- **Step 3: Bounding the CGF.**
  $\mathcal{E}([W]_k) - \mathcal{E}([W]_{k-1})$ is $d^2([W]_k, [W]_{k-1})$-subGaussian:

$$\mathrm{CGF} = \log \mathbb{E}_{S'}\left[\mathbb{E}_{[W]_k,[W]_{k-1}}\left[e^{t(\mathcal{E}([W]_k)-\mathcal{E}([W]_{k-1}))}\right]\right] \leq \frac{t^2\mathbb{E}\left[d^2([W]_k,[W]_{k-1})\right]}{2}$$

- **Step 4: Optimizing the bound.**

$$\mathcal{E} \leq \sum_{k=k_0+1}^{\infty} \sqrt{2\mathbb{E}_{[W]_k,[W]_{k-1}}\left[d^2([W]_k,[W]_{k-1})\right] I([W]_k,[W]_{k-1};S)}.$$

Notice that $S - W - [W]_k - [W]_{k-1}$ is a Markov chain, so
$I([W]_k,[W]_{k-1};S) = I([W]_k;S) + I([W]_k,[W]_{k-1};S|[W]_k) = I([W]_k,[W]_{k-1};S)$.

Special case: 2^{-k}-partition, $d([W]_k, W) \leq 2^{-k}$, then
$d([W]_k, W) + d([W]_{k-1}, W) \leq 2^{-k} + 2^{-(k-1)} = 3 \times 2^{-k}$.

Individual Technique

- **Step 1: Finding the target.**

$$\mathbb{E}_{W,S}[L_\mu(W) - L_S(W)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{W,Z_i}\left[\mathbb{E}_{Z'}\left[\ell(W,Z')\right] - \ell(W,Z_i)\right]$$

- **Step 2: Selecting f, Q and P.**

$$f = t \cdot \left(\mathbb{E}_{Z'}\left[\ell(W,Z')\right] - \ell(W,Z_i)\right), \quad Q = P_{W,Z_i}, \quad P = P_{Z'} \otimes P_W$$

- **Step 3: Bounding the CGF.**

$\ell(W,Z_i')$ is R-subGaussian:$\log \mathbb{E}_{Z'}\left[\mathbb{E}_W\left[e^{t(\mathbb{E}_{Z'}[\ell(W,Z')]-\ell(W,Z_i))}\right]\right] \leq \frac{t^2R^2}{2}$

- **Step 4: Optimizing the bound.**

$$\mathcal{E} \preceq \frac{1}{n}\sum_{i=1}^{n}\sqrt{I(W;Z_i)} \leq \sqrt{\frac{I(W;S)}{n}}$$

Random Subset Technique

- **Step 1: Finding the target.**
  Let $J \subseteq [n]$, $|J| = m$,

$$\begin{aligned}\mathbb{E}_{W,S}[L_\mu(W) - L_S(W)] =& \mathbb{E}_{W,S}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}_{Z'}\left[\ell(W,Z')\right] - \ell(W,Z_i)\right)\right] \\ =& \mathbb{E}_J\left[\mathbb{E}_{W,S_J}\left[\frac{1}{m}\sum_{i=1}^{m}\left(\mathbb{E}_{Z'}\left[\ell(W,Z')\right] - \ell(W,S^c_{Ji})\right)\right]\right]\end{aligned}$$

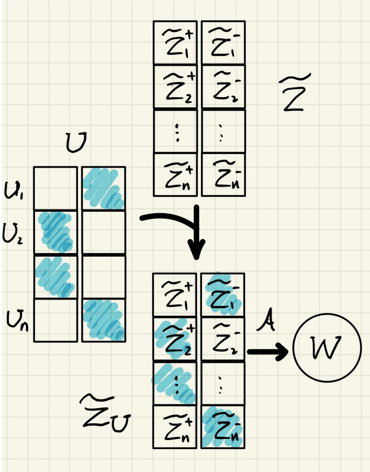
- **Step 2: Selecting f, Q and P.**

$$f = t\cdot\left(\frac{1}{m}\sum_{i=1}^{m}\left(\mathbb{E}_{Z'}\left[\ell(W,Z')\right] - \ell(W,S^c_{Ji})\right)\right), \quad Q = P_{W,S^c_J|S,J}, \quad P = P_{S^c_J}\otimes P_{W'|S_J}$$

$\Longrightarrow$ Data-Dependent Prior of W

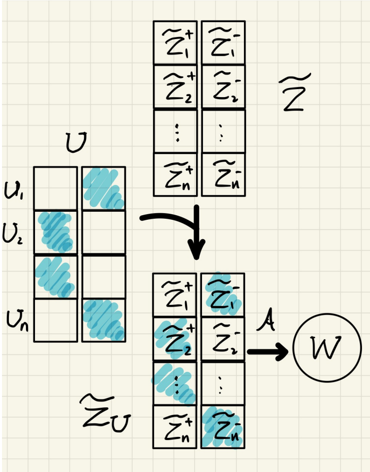
- $\mathcal{E} \precsim \mathbb{E}\sqrt{\frac{I^{S_J}(W;S^c_J)}{n-m}}$; Individual Technique is a special case for $m = n - 1$.

CMI:Supersample Setting



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

CMI:Supersample Setting



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

Lemma 4 (Steinke and Zakynthinou [2020])

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

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

CMI: Proof

- **Step 1: Finding the target.**

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

- **Step 2: Selecting f, Q and P.**

$$f = t\cdot\frac{1}{n}\sum_{i=1}^{n}(-1)^{U_i}\left(\ell(W,\tilde{z}_i^-) - \ell(W,\tilde{z}_i^+)\right), \quad Q = P_{W,U|\tilde{z}}, \quad P = P_{U'}\otimes P_{W|\tilde{z}}$$

- **Step 3: Bounding the CGF.**

$(-1)^{U_i}\left(\ell(w,\tilde{z}_i^-) - \ell(w,\tilde{z}_i^+)\right)$ is $\left|\ell(w,\tilde{z}_i^-) - \ell(w,\tilde{z}_i^+)\right|^2$-subGaussian:

$$\log\mathbb{E}_{W|\tilde{z}}\left[\mathbb{E}_{U'}\left[e^{t\frac{1}{n}\sum_{i=1}^{n}(-1)^{U_i}\left(\ell(W,\tilde{z}_i^-)-\ell(W,\tilde{z}_i^+)\right)}\right]\right] \leq \frac{t^2}{2n}$$

- **Step 4: Optimizing the bound.**

$$\mathcal{E} \preceq \sqrt{\frac{I(W;U|\widetilde{Z})}{n}} \leq \sqrt{\frac{I(W;S)}{n}}.$$

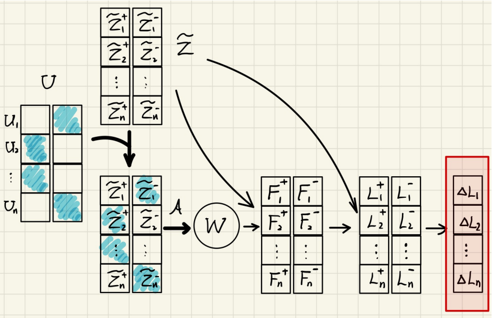
Further Development

- **Random Subset CMI**: *Haghifam, M. et al. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. NeurIPS 2020.*
- **Individual CMI**: *Rodríguez-Gálvez, B. et al. On random subset generalization error bounds and the stochastic gradient Langevin dynamics algorithm. ITW 2020.*
  Zhou R, et al. Individually conditional individual mutual information bound on generalization error. TIT 2022.
- **Stochastic Chaining IOMI/CMI**: *Zhou R, et al. Stochastic Chaining and Strengthened Information-Theoretic Generalization Bounds ISIT 2022.*
- **Leave-One-Out CMI**: *Haghifam, M. et al. Understanding Generalization via Leave-One-Out Conditional Mutual Information. ISIT 2022.*
  Rammal, M. R. et al. On leave-one-out conditional mutual information for generalization. NeurIPS 2022.

Information-Theoretic Generalization Bounds for Black-Box Algorithms

Tighter CMI Bound

- *Wang, Z., and Mao, Y.. Tighter Information-Theoretic Generalization Bounds from Supersamples. ICML 2023.*
- Main Contribution: New **Conditional Mutual Information (CMI)** bounds which are **either theoretically or empirically tighter** than previous CMI bounds for the **same supersample** setting.



- $F_i^+ := f_W(\widetilde{X}_i^+)$, $F_i^- := f_W(\widetilde{X}_i^-)$, $F_i := (F_i^+, F_i^-)$

  $\Rightarrow$ **f-CMI Bound**: $|\mathcal{E}| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{I(F_i; U_i|\widetilde{Z})}$

  [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**: $|\mathcal{E}| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{I(L_i; U_i|\widetilde{Z})}$

  [Hellström and Durisi, 2022]

- This paper: $\Delta L_i := L_i^- - L_i^+$

  $\Rightarrow$ **ld-CMI**: $I(\Delta L_i; U_i|\widetilde{Z})$

Revist CMI Proof

- **Step 1: Finding the target.**

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

- **Step 2: Selecting f, Q and P.**

$$\begin{aligned}f =& t\cdot\frac{1}{n}\sum_{i=1}^{n}(-1)^{U_i}\left(\ell(W,\tilde{z}_i^-)-\ell(W,\tilde{z}_i^+)\right)\\ =& (-1)^{U_i}\Delta L_i\end{aligned}$$

$$\begin{aligned}Q =& P_{W,U|\tilde{z}} = P_{\Delta L_i,U_i|\tilde{z}} \text{ or } P_{\Delta L_i,U_i}\\ P =& P_{U'}\otimes P_{W|\tilde{z}} = P_{U'}\otimes P_{\Delta L_i|\tilde{z}} \text{ or } P_{U'}\otimes P_{\Delta L_i}\end{aligned}$$

CMI, f-CMI and e-CMI

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

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

Generalization Bounds via Loss Difference

Theorem 1

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

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

$$|\mathcal{E}| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2I(\Delta L_i; U_i)}. \tag{3}$$

Theorem 1

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

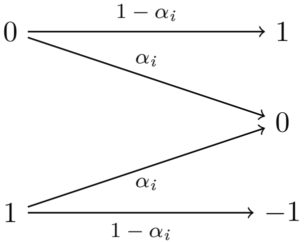
$$|\mathcal{E}| \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\widetilde{Z}}\sqrt{2I^{\widetilde{Z}}(\Delta L_i; U_i)} \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2I(\Delta L_i; U_i|\widetilde{Z})}, \tag{2}$$

$$|\mathcal{E}| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2I(\Delta L_i; U_i)}. \tag{3}$$

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



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

Theorem 2

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

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

Key observation:

$\mathcal{E} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{W,U_i,\widetilde{Z}}\left[(-1)^{U_i}\left(\ell(W,\widetilde{Z}_i^+) - \ell(W,\widetilde{Z}_i^-)\right)\right] = \frac{2}{n}\sum_{i=1}^n \mathbb{E}_{L_i^+,\varepsilon_i}\left[\varepsilon_i L_i^+\right]$,

where $\varepsilon_i = (-1)^{\overline{U}_i}$ is the Rademacher variable.

Key observation:

$\mathcal{E} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{W,U_i,\widetilde{Z}}\left[(-1)^{U_i}\left(\ell(W,\widetilde{Z}_i^{+}) - \ell(W,\widetilde{Z}_i^{-})\right)\right] = \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_{L_i^{+},\varepsilon_i}\left[\varepsilon_i L_i^{+}\right]$,

where $\varepsilon_i = (-1)^{\overline{U}_i}$ is the Rademacher variable.

Lemma 5

Consider the <u>*weighted generalization error,* $\mathcal{E}_{C_1} \triangleq L_\mu - (1+C_1)L_n$</u>. *We have*

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

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 3

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

$$L_\mu \leq (1 + C_1) L_n + \sum_{i=1}^{n} \frac{I(L_i^+; U_i)}{C_2 n}, \tag{4}$$

$$L_\mu \leq L_n + \sum_{i=1}^{n} \frac{4I(L_i^+; U_i)}{n} + 4\sqrt{\sum_{i=1}^{n} \frac{L_n I(L_i^+; U_i)}{n}}. \tag{5}$$

Theorem 3

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

$$L_\mu \leq (1+C_1)L_n + \sum_{i=1}^{n} \frac{I(L_i^+;U_i)}{C_2 n}, \tag{4}$$

$$L_\mu \leq L_n + \sum_{i=1}^{n} \frac{4I(L_i^+;U_i)}{n} + 4\sqrt{\sum_{i=1}^{n} \frac{L_n I(L_i^+;U_i)}{n}}. \tag{5}$$

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

Sharpness Based MI Bound

Theorem 4

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

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

Theorem 4

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

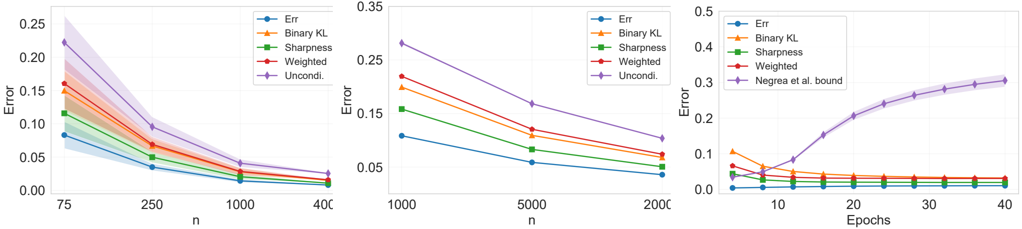
$$F_i(\lambda) \triangleq \mathbb{E}_{W,Z_i}\left[\ell(W,Z_i) - (1+\lambda)\mathbb{E}_{W|Z_i}\ell(W,Z_i)\right]^2.$$

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

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

- $L_n = 0 \to F(\lambda) = 0$, but $L_n = 0 \not\leftarrow F(\lambda) = 0$;
- For any fixed C_1 and C_2, Eq. (6) is tighter than Eq. (4).

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: $C_1 F(\lambda) + \sum_{i=1}^{n}\frac{I(L_i^+;U_i)}{C_2 n}$.

Information-Theoretic Bounds in Stochastic Convex Optimization

Limitations in SCO

Limitations of Information-Theoretic (IT) bounds:

- Original input-output mutual information (IOMI) (e.g., $I(W;S)$ [Xu and Raginsky, 2017]) based bound can $\to \infty$
  $\Longrightarrow$ solved by conditional mutual information (CMI) $I(W;U|\widetilde{Z})$ [Steinke and Zakynthinou, 2020]
- Slow convergence rate, e.g., $\mathcal{O}(1/\sqrt{n})$
  $\Longrightarrow$ mitigated by [Haghifam et al., 2021, Hellström and Durisi, 2021, 2022, Wang and Mao, 2023, Wu et al., 2023, Zhou et al., 2023]

Limitations in SCO

Limitations of Information-Theoretic (IT) bounds:

- Original input-output mutual information (IOMI) (e.g., $I(W;S)$ [Xu and Raginsky, 2017]) based bound can $\to \infty$
  $\Longrightarrow$ solved by conditional mutual information (CMI) $I(W;U|\widetilde{Z})$ [Steinke and Zakynthinou, 2020]
- Slow convergence rate, e.g., $\mathcal{O}(1/\sqrt{n})$
  $\Longrightarrow$ mitigated by [Haghifam et al., 2021, Hellström and Durisi, 2021, 2022, Wang and Mao, 2023, Wu et al., 2023, Zhou et al., 2023]
- Non-vanishing in Stochastic Convex Optimization (SCO) problems for (nearly) all previous IT bounds![Haghifam et al., 2023]

Limitations in SCO

Limitations of Information-Theoretic (IT) bounds:

- Original input-output mutual information (IOMI) (e.g., $I(W;S)$ [Xu and Raginsky, 2017]) based bound can $\to \infty$
  $\Longrightarrow$ solved by conditional mutual information (CMI) $I(W;U|\widetilde{Z})$ [Steinke and Zakynthinou, 2020]
- Slow convergence rate, e.g., $\mathcal{O}(1/\sqrt{n})$
  $\Longrightarrow$ mitigated by [Haghifam et al., 2021, Hellström and Durisi, 2021, 2022, Wang and Mao, 2023, Wu et al., 2023, Zhou et al., 2023]
- Non-vanishing in Stochastic Convex Optimization (SCO) problems for (nearly) all previous IT bounds![Haghifam et al., 2023]

  Wang, Z. and Mao, Y.. Sample-Conditioned Hypothesis Stability Sharpens Information-Theoretic Generalization Bounds. NeurIPS 2023.

  Our contribution: **Incorporating stability-based analysis into IT framework which improves both stability-based bounds and IT bounds**.

SCO Counterexample

- Z is one-hot vector in $\mathbb{R}^d$
- Loss: $-\langle w, z \rangle$; ERM solution $W = \frac{1}{n}\sum_{i=1}^{n} Z_i$
- Birthday Paradox Problem: For a large 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 (independent of n).
- If $d \geq \frac{2n-1}{1-c^{1/(2n-1)}}$, then $P(E_0) \geq c \geq \left(1 - \frac{2n-1}{d}\right)^{2n-1}$, e.g., $d = 2n^2 \Longrightarrow c \geq 0.1$.
- Let $d = 2n^2$, $I(W; U_i|\widetilde{Z}_i) = \log 2 - H(U_i|W, \widetilde{Z}_i) \geq 0.1 \cdot \log 2$.
- **CMI bound is non-vanishing but $\mathcal{E} \leq \mathcal{O}(1/\sqrt{n})$.**

Algorithmic Stability

$$\begin{array}{c|c|cc|c|cc} Z_1,\ldots, & Z_i, & \ldots, Z_n & \xrightarrow{\mathcal{A}} & W & \Rightarrow & \ell\left(W, Z\right) \\ Z_1,\ldots, & Z_i', & \ldots, Z_n & \xrightarrow{\mathcal{A}} & W^{-i} & \Rightarrow & \ell\left(W^{-i}, Z\right) \end{array}$$

$\mathcal{A}$ is Stable $\iff$ Loss of $\left(W^{-i}, Z\right)$ is close to Loss of $\left(W, Z\right)$.

- Uniform Stability [Bousquet and Elisseeff, 2002]:
  $\sup_{W, W^{-i}, Z} \left|\ell(W, Z) - \ell(W^{-i}, Z)\right| \leq$ Unif. Stability Param.
- Sample-Conditioned Hypothesis (SCH) Stability in our paper
  $\mathbb{E}_{W, W^{-i}}\left[\sup_Z \left|\ell(W, Z) - \ell(W^{-i}, Z)\right|\right] \leq$ SCH Stability Param.,
  where Z can be either Z_i or Z_i'.

Novel Construction

By DV lemma: $\mathcal{E} \leq \inf_{\mathrm{t}>0} \frac{\text{IOMI or CMI+CGF}}{\mathrm{t}}$.
where

$$\text{CGF} = \log \mathbb{E}\left[\exp\left(t \cdot f_{\text{DV}}\right)\right].$$

Novel Construction

By DV lemma: $\mathcal{E} \leq \inf_{\mathrm{t}>0} \frac{\text{IOMI or CMI+CGF}}{\mathrm{t}}$.
where

$$\mathrm{CGF} = \log \mathbb{E}\left[\exp\left(t \cdot f_{\mathrm{DV}}\right)\right].$$

- Previous works:
  $f_{\mathrm{DV}} = \ell(W, Z')$ e.g., [Bu et al., 2019]
  $f_{\mathrm{DV}} = \ell(W, Z') - \mathbb{E}_{Z'}\left[\ell(W, Z')\right]$ e.g., [Wu et al., 2023]
  $f_{\mathrm{DV}} = (-1)^{U}\left(\ell(W, Z_1) - \ell(W, Z_2)\right)$ e.g., [Steinke and Zakynthinou, 2020]

Novel Construction

By DV lemma: $\mathcal{E} \leq \inf_{t>0} \frac{\text{IOMI or CMI+CGF}}{t}$.
where

$$\mathrm{CGF} = \log \mathbb{E}\left[\exp\left(t \cdot f_{\mathrm{DV}}\right)\right].$$

- Previous works:
  $f_{\mathrm{DV}} = \ell(W, Z')$ e.g., [Bu et al., 2019]
  $f_{\mathrm{DV}} = \ell(W, Z') - \mathbb{E}_{Z'}\left[\ell(W, Z')\right]$ e.g., [Wu et al., 2023]
  $f_{\mathrm{DV}} = (-1)^{U}\left(\ell(W, Z_1) - \ell(W, Z_2)\right)$ e.g., [Steinke and Zakynthinou, 2020]
- This paper: let W^{-i} be obtained by replacing one data in S,
  $f_{\mathrm{DV}} = \ell(W, Z') - \mathbb{E}_{W^{-i}|W}\left[\ell(W^{-i}, Z')\right] \Longrightarrow$ IOMI
  $f_{\mathrm{DV}} = (-1)^{U}\left(\ell(W, Z) - \ell(W^{-i}, Z)\right) \Longrightarrow$ New CMI

New CMI: Neighboring-Hypothesis Matrix

$$\begin{array}{|c|c|}\hline \widetilde{Z}_{1,0} & \blacksquare \\ \widetilde{Z}_{2,0} & \blacksquare \\ \vdots & \vdots \\ \widetilde{Z}_{n,0} & \blacksquare \\ \hline \end{array} \xrightarrow{\mathcal{A}} W, \begin{array}{|c|c|}\hline \blacksquare & \widetilde{Z}_{1,1} \\ \widetilde{Z}_{2,0} & \blacksquare \\ \vdots & \vdots \\ \widetilde{Z}_{n,0} & \blacksquare \\ \hline \end{array} \xrightarrow{\mathcal{A}} \underbrace{W^{-1}, \cdots, W^{-n}}_{n \text{ new neighboring hypotheses}} \Longrightarrow \begin{array}{|c|c|}\hline W & W^{-1} \\ W & W^{-2} \\ \vdots & \vdots \\ W & W^{-n} \\ \hline \end{array}$$

The generalization game:

$$\begin{array}{|c|c|}\hline \widetilde{Z}_{1,0} & \widetilde{Z}_{1,1} \\ \widetilde{Z}_{2,0} & \widetilde{Z}_{2,1} \\ \vdots & \vdots \\ \widetilde{Z}_{n,0} & \widetilde{Z}_{n,1} \\ \hline \end{array} + \begin{array}{|c|c|}\hline \square & \blacksquare \\ \blacksquare & \square \\ \vdots & \vdots \\ \square & \blacksquare \\ \hline \end{array} \Longrightarrow \begin{array}{|c|c|}\hline \widetilde{Z}_{1,0} & \blacksquare \\ \blacksquare & \widetilde{Z}_{2,1} \\ \vdots & \vdots \\ \widetilde{Z}_{n,0} & \blacksquare \\ \hline \end{array} + \begin{array}{|c|c|}\hline W & W^{-1} \\ W & W^{-2} \\ \vdots & \vdots \\ W & W^{-n} \\ \hline \end{array} \Longrightarrow \begin{array}{|c|c|}\hline \text{Tr. Err}_1 & \text{Ts. Err}_1 \\ \text{Ts. Err}_2 & \text{Tr. Err}_2 \\ \vdots & \vdots \\ \text{Ts. Err}_n & \text{Tr. Err}_n \\ \hline \end{array}$$

$$\Longrightarrow \mathcal{E} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\text{Ts. Err}_i - \text{Tr. Err}_i\right] (\leq \text{Stability Param.})$$

New CMI: Neighboring-Hypothesis Matrix

Given $\begin{array}{|c|c|} W & W^{-1} \\ W & W^{-2} \\ \vdots & \vdots \\ W & W^{-n} \end{array}$ and $\begin{array}{|c|} \widetilde{Z}_{1,0} \\ \widetilde{Z}_{2,1} \\ \vdots \\ \widetilde{Z}_{n,0} \end{array}$, can you infer $\begin{array}{|c|c|} \square & \blacksquare \\ \blacksquare & \square \\ \vdots & \vdots \\ \square & \blacksquare \end{array}$ 🤔?

Theorem 5 (Informal.)

If $\mathcal{A}$ is β-stable, we have $\mathcal{E} \precsim \beta\sqrt{I(Z_U; U|W, W^{-i})} \leq \beta\sqrt{I(W; Z_i)}$

Revisit SCO Example

In SCO counterexamples given by Haghifam et al. [2023]:

$$\mathcal{E} \leq \mathcal{O}(1/\sqrt{n}).$$

- Previous IOMI or CMI bound in these examples:
  SubGaussian param. $R = \mathcal{O}(1)$ (=Lip. Para.×Diam. of hypo. space)
  and IOMI≥CMI= $\mathcal{O}(1)$.
  $\Longrightarrow$ IOMI bound $\geq$ CMI bound $\in \mathcal{O}(1) \Longrightarrow$ fail to explain the learnability.
- New CMI bound in these examples:
  $\beta = \mathcal{O}(1/\sqrt{n})$
  and New CMI= $\mathcal{O}(1)$.
  $\Longrightarrow$ **New CMI bound** $\in \mathcal{O}(1/\sqrt{n}) \Longrightarrow$ can explain the learnability.
- More bounds, e.g., fast-rate bounds and second-moment bounds.
- More examples, e.g., our bounds can also improve stability-based bounds.

CMI on Distribution-free Setting

CMI and VC-dim:

Theorem 6

Let $\mathcal{Z} = \mathcal{X} \times \{0,1\}$, and let $\mathcal{F} = \{f_w : \mathcal{X} \to \{0,1\} | w \in \mathcal{W}\}$ be a functional hypothesis class with finite VC dimension d. Let $n > d+1$, for any algorithm $\mathcal{A}$,

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

Proof Sketch.
For a given $\widetilde{Z}$, the number of distinct values of their predictions, denoted by k, by Sauer-Shelah lemma for $n > d+1$, $k \leq \sum_{i=1}^{d}\binom{n}{i} \leq \left(\frac{en}{d}\right)^d$.

$$I(F^+, F^-; U\big|\widetilde{Z}) \leq H(F^+, F^-\big|\widetilde{Z}) \leq H(F^+\big|\widetilde{Z}) + H(F^-\big|\widetilde{Z}) \leq 2d\log\left(\frac{en}{d}\right).$$

- *Steinke, T., and Zakynthinou, L.. Open problem: Information complexity of vc learning. COLT 2020.*
- *Hafez-Kolahi, H. et al. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. NeurIPS 2020.*
- *Haghifam, M. et al. Towards a unified information-theoretic framework for generalization. NeurIPS 2021.*
- LOO CMI: *Haghifam, M. et al. Understanding Generalization via Leave-One-Out Conditional Mutual Information. ISIT 2022.*
- f-CMI and e-CMI: *Harutyunyan, H. et al. Information-theoretic generalization bounds for black-box learning algorithms. NeurIPS 2021.*
  Hellström, F. and Durisi, G.. A new family of generalization bounds using samplewise evaluated CMI. NeurIPS 2022.
- *Bassily, R. et al. Learners that use little information. ALT 2018.*
- *Livni, R.. Information Theoretic Lower Bounds for Information Theoretic Upper Bounds. NeurIPS 2023.*

Information-Theoretic Generalization Bounds for SGD

Previous MI Bound

Lemma 6 (Xu and Raginsky [2017])

Assume the loss $\ell(w, Z)$ is R-subgaussian[2] for any $w \in \mathcal{W}$. The generalization error of $\mathcal{A}$ is bounded by

$$|\mathcal{E}| \leq \sqrt{\frac{2R^2}{n} I(W;S)},$$

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

$\Longrightarrow$ Distribution-dependent and Algorithm-dependent

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

Stochastic Gradient Langevin Dynamics (SGLD)

SGLD updates:

$$W_t \triangleq W_{t-1} - \lambda_t g(W_{t-1}, B_t) + N_t,$$

where

$$g(w, B_t) \triangleq \frac{1}{b} \sum_{z \in B_t} \nabla_w \ell(w, z),$$

- λ_t: learning rate
- b: batch size
- B_t denotes the batch used for the t^{th} update
- $N_t \sim \mathcal{N}(0, \sigma_t^2 \mathbf{I}_d)$

Assume SGLD outputs W_T as the learned model parameter.

IT Bounds for SGLD

$$
\begin{aligned}
I(W_T;S) =& I\left(W_{T-1} - \lambda_T g(W_{T-1},B_T) + N_T; S\right) \\
\leq& I\left(W_{T-1}, -\lambda_T g(W_{T-1},B_T) + N_T; S\right) && (7) \\
=& I(W_{T-1};S) + I\left(-\lambda_T g(W_{T-1},B_T) + N_T; S | W_{T-1}\right) && (8) \\
&\vdots \\
\leq& \sum_{t=1}^{T} I\left(-\lambda_t g(W_{t-1},B_t) + N_t; S | W_{t-1}\right)
\end{aligned}
$$

$$
\begin{aligned}
& I\left(-\lambda_t g(W_{t-1},B_t) + N_t; S | W_{t-1}\right) \\
=& \mathbb{E}_{S,W_{t-1}}\left[\mathrm{D_{KL}}\left(Q_{-\lambda_t g(W_{t-1},B_t)+N_t|S,W_{t-1}} || P_{-\lambda_t g(W_{t-1},B'_t)+N_t|W_{t-1}}\right)\right] \\
\leq& \frac{d}{2}\mathbb{E}_{W_{t-1}} \log\left(\frac{\lambda_t^2 \mathbb{E}_S^{W_{t-1}} ||g - \mathbb{E}[g]||_2^2}{d\sigma_t^2} + 1\right).
\end{aligned}
$$

IT Bounds for SGLD

Theorem 7

Gen. err. of SGLD is upper bounded by

$$\mathcal{E} \precsim \sqrt{\frac{d}{n}\sum_{t=1}^{T}\mathbb{E}\log\left(\frac{\lambda_t^2\mathbb{E}\left|\left|g-\mathbb{E}\left[g\right]\right|\right|_2^2}{d\sigma_t^2}+1\right)}.$$

Other Works on SGLD

- *Bu, Y. et al. Tightening Mutual Information Based Bounds on Generalization Error. ISIT 2019.*
  Negrea, J. et al. Information-theoretic generalization bounds for SGLD via data-dependent estimates. NeurIPS 2019.
  Haghifam, M. et al. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. NeurIPS 2020.
  Rodríguez-Gálvez, B. et al. On random subset generalization error bounds and the stochastic gradient Langevin dynamics algorithm. ITW 2020.
- *Wang, Hao et al. Analyzing the generalization capability of sgld using properties of gaussian channels. NeurIPS 2021.*
- *Li, J. et al. On generalization error bounds of noisy gradient methods for non-convex learning. ICLR 2020.*

Other Works on SGLD (Cont.)

- *Mou, W.. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. COLT 2018.*
- *Banerjee, A. et al. Stability based generalization bounds for exponential family langevin dynamics. ICML 2022.*
- *Futami, F., and Fujisawa, M.. Time-Independent Information-Theoretic Generalization Bounds for SGLD. NeurIPS 2023.*

Stochastic Gradient Descent (SGD)

SGD updates:

$$W_t \triangleq W_{t-1} - \lambda_t g(W_{t-1}, B_t),$$

where

$$g(w, B_t) \triangleq \frac{1}{b} \sum_{z \in B_t} \nabla_w \ell(w, z),$$

- λ_t: learning rate
- b: batch size
- B_t denotes the batch used for the t^{th} update.

Assume SGD outputs W_T as the learned model parameter.

Stochastic Gradient Descent (SGD)

SGD updates:

$$W_t \triangleq W_{t-1} - \lambda_t g(W_{t-1}, B_t),$$

where

$$g(w, B_t) \triangleq \frac{1}{b} \sum_{z \in B_t} \nabla_w \ell(w, z),$$

- λ_t: learning rate
- b: batch size
- B_t denotes the batch used for the t^{th} update.

Assume SGD outputs W_T as the learned model parameter.

Difficulty of using MI based bound: $I(W_T; S) \rightarrow$ too large for SGD

Auxiliary Weight Process (only exists in the analysis)

Follow up the work of Neu et al. [2021], let $\{\sigma_t\}_{t=1}^{T}$ be a sequence of positive real numbers.

Define $\widetilde{W}_0 \triangleq W_0$, and $\widetilde{W}_t \triangleq \widetilde{W}_{t-1} - \lambda_t g(W_{t-1}, B_t) + N_t$, for $t > 0$, where $N_t \sim \mathcal{N}(0, \sigma_t^2 \mathbf{I}_d)$ is a Gaussian noise.

$$
\begin{array}{ccccccccccc}
 & & N_1 & & N_2 & & \cdots & & N_{T-1} & & N_T \\
 & & \downarrow & & \downarrow & & & & \downarrow & & \downarrow \\
\widetilde{W}_0 & \rightarrow & \widetilde{W}_1 & \rightarrow & \widetilde{W}_2 & \rightarrow & \cdots & \rightarrow & \widetilde{W}_{T-1} & \rightarrow & \widetilde{W}_T \\
\| & \nearrow & & \nearrow & & \nearrow & & & & \nearrow & \\
W_0 & \rightarrow & W_1 & \rightarrow & W_2 & \rightarrow & \cdots & \rightarrow & W_{T-1} & \rightarrow & W_T
\end{array}
$$

Let $\Delta_t = \sum_{\tau=1}^{t} N_\tau$. Notice that $\widetilde{W}_t = W_t + \Delta_t$.

Bounding by Auxiliary Weights

Denote this auxiliary weight process by $\mathcal{A}_{AWP}$. Let $\mathcal{A}_{SGD}$ be the original algorithm of SGD,

$$\begin{aligned}
\mathcal{E}_\mu\left(\mathcal{A}_{SGD}\right) =& \mathcal{E}_\mu\left(\mathcal{A}_{SGD}\right) + \mathcal{E}_\mu\left(\mathcal{A}_{AWP}\right) - \mathcal{E}_\mu\left(\mathcal{A}_{AWP}\right) \\
\leq & \underbrace{\mathcal{O}\left(\sqrt{\frac{I(\widetilde{W}_T;S)}{n}}\right)}_{\text{Lemma 3}} + \underbrace{\left|\mathcal{E}_\mu\left(\mathcal{A}_{SGD}\right) - \mathcal{E}_\mu\left(\mathcal{A}_{AWP}\right)\right|}_{\text{residual term}} \qquad (9) \\
\precsim & \text{ Trajectories of Gradient Variance/Dispersion} + \text{Sharpness.}
\end{aligned}$$

Main Result

Theorem 8 (Wang and Mao [2022])

The generalization error of SGD is upper bounded by

$$\mathcal{E} \precsim \sqrt[3]{\sum_{t=1}^{T} \frac{\mathbb{E}\left[\mathbb{V}_t(W_{t-1})\right]\mathbb{E}\left[\mathrm{Tr}\left(\mathrm{H}_{W_T}(Z)\right)\right]}{n}} \tag{10}$$

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

SDE updates: $W_t \triangleq W_{t-1} - \eta g(W_{t-1}, S) + \eta C_t^{1/2} N_t$, where

$$C_t \triangleq \frac{n-b}{b(n-1)} \left(\frac{1}{n} \sum_{i=1}^{n} \nabla \ell_i \nabla \ell_i^{\mathrm{T}} - G_t G_t^{\mathrm{T}} \right)$$

is the *gradient noise covariance* matrix.

Denote SDE approximation as $\mathcal{A}_{SDE}$,

$$\begin{aligned} \mathcal{E}_\mu \left(\mathcal{A}_{SGD}\right) =& \mathcal{E}_\mu \left(\mathcal{A}_{SGD}\right) + \mathcal{E}_\mu \left(\mathcal{A}_{SDE}\right) - \mathcal{E}_\mu \left(\mathcal{A}_{SDE}\right) \\ \leq & \underbrace{\mathcal{O}\left(\sqrt{\frac{I(W_{\mathrm{SDE}}; S)}{n}}\right)}_{\text{Lemma 3}} + \underbrace{\left|\mathcal{E}_\mu \left(\mathcal{A}_{SGD}\right) - \mathcal{E}_\mu \left(\mathcal{A}_{SDE}\right)\right|}_{\text{residual term}}, \end{aligned} \tag{11}$$

where W_{SDE} is the output hypothesis by $\mathcal{A}_{SDE}$.

$$\mathcal{E}_\mu\left(\mathcal{A}_{SGD}\right) \leq \underbrace{\mathcal{O}\left(\sqrt{\frac{I(W_{\mathrm{SDE}};S)}{n}}\right)}_{\text{Lemma 3}} + \underbrace{\left|\mathcal{E}_\mu\left(\mathcal{A}_{SGD}\right) - \mathcal{E}_\mu\left(\mathcal{A}_{SDE}\right)\right|}_{\text{residual term}}.$$

Empirical evidence from [Wu et al., 2020, Li et al., 2021] and suggests that the residual term is small.
$\Longrightarrow$ It is safe to investigate the generalization of SGD using the IT bounds of SDE directly.

Information-Theoretic Analysis Beyond Supervised Learning

Applying IT Analysis to Unsupervised Domain Adaptation

Wang, Z. and Mao Y.. Information-theoretic analysis of unsupervised domain adaptation. ICLR 2023.

- 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:*

$$\begin{aligned}\mathcal{E} \triangleq & \mathbb{E}_{W,S,S'_{X'}}\left[R_{\mu'}(W) - R_S(W)\right] \\ = & \mathbb{E}_{W,S,S'_{X'}}\left[L_{\mu'}(W) - L_{\mu}(W) + L_{\mu}(W) - L_S(W)\right]\end{aligned}$$

MI Bound for UDA

Theorem 9

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

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

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 10

Under the assumption of Theorem 9. Let the total iteration number be T, then

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

restrict the gradient norm $\Longrightarrow$ reduce $|\mathcal{E}|$.

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**

Other Works beyond Supervised Learning

- **Semi-supervised learning**: *He, H. et al. Information-theoretic characterization of the generalization error for iterative semisupervised learning. JMLR 2022.*
  Aminian, G. et al. An information-theoretical approach to semi-supervised learning under covariate-shift. AISTATS 2022.
- **Transfer Learning**: *Wu, X. et al. Informationtheoretic analysis for transfer learning. ISIT 2020.*
  Bu, Y. et al. Characterizing and understanding the generalization error of transfer learning with gibbs algorithm. AISTATS 2022.
- **Meta Learning**: *Jose, S. T. and Simeone, O. Information-theoretic generalization bounds for meta-learning and applications. Entropy 2021.*
  Chen, Q. et al. Generalization bounds for meta-learning: An information-theoretic analysis. NeurIPS 2021
  Hellström, F. and Durisi, G. Evaluated CMI bounds for meta learning: Tightness and expressiveness. NeurIPS 2022.

Other Works beyond Supervised Learning (Cont.)

- **Federated Learning**: *Yagli, S. et al. Information-theoretic bounds on the generalization error and privacy leakage in federated learning. SPAWC 2020.*
- **Transductive Learning**: *Tang, H. and Liu, Y. Information-Theoretic Generalization Bounds for Transductive Learning and its Applications. arXiv preprint arXiv:2311.04561, 2023.*
- **Quantum Learning**: *Caro, M. et al. Information-theoretic generalization bounds for learning from quantum data. arXiv preprint arXiv:2311.05529, 2023.*

More Related Works (beyond DV Lemma)

- *Wang, H. et al. An information-theoretic view of generalization via Wasserstein distance. ISIT 2019.*
- *Rodríguez-Gálvez, B. et al. Tighter expected generalization error bounds via Wasserstein distance. NeurIPS 2021.*
- *Lugosi, G., and Neu, G.. Generalization bounds via convex analysis. COLT 2022.*
- *Lugosi, G., and Neu, G.. Online-to-PAC Conversions: Generalization Bounds via Regret Analysis. arXiv preprint arXiv:2305.19674, 2023.*
- *Chu, Y., and Raginsky, M.. A unified framework for information-theoretic generalization bounds. NeurIPS 2023.*

References I

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.

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

Mahdi Haghifam, Gintare Karolina Dziugaite, Shay Moran, and Dan Roy. Towards a unified information-theoretic framework for generalization. *Advances in Neural Information Processing Systems*, 34:26370–26381, 2021.

Fredrik Hellström and Giuseppe Durisi. Fast-rate loss bounds via conditional information measures with applications to neural networks. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 952–957. IEEE, 2021.

Ziqiao Wang and Yongyi Mao. Tighter information-theoretic generalization bounds from supersamples. In *International Conference on Machine Learning*. PMLR, 2023.

Xuetong Wu, Jonathan H Manton, Uwe Aickelin, and Jingge Zhu. On the tightness of information-theoretic bounds on generalization error of learning algorithms. *arXiv preprint arXiv:2303.14658*, 2023.

Ruida Zhou, Chao Tian, and Tie Liu. Exactly tight information-theoretic generalization error bound for the quadratic gaussian problem. *arXiv preprint arXiv:2305.00876*, 2023.

Mahdi Haghifam, Borja Rodríguez-Gálvez, Ragnar Thobaben, Mikael Skoglund, Daniel M Roy, and Gintare Karolina Dziugaite. Limitations of information-theoretic generalization bounds for gradient descent methods in stochastic convex optimization. In *International Conference on Algorithmic Learning Theory*, pages 663–706. PMLR, 2023.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. Tightening mutual information based bounds on generalization error. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 587–591. IEEE, 2019.

Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel M Roy. Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory*. PMLR, 2021.

Ziqiao Wang and Yongyi Mao. On the generalization of models trained with SGD: Information-theoretic bounds and implications. In *International Conference on Learning Representations*, 2022.

Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the noisy gradient descent that generalizes as sgd. In *International Conference on Machine Learning*. PMLR, 2020.

Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling sgd with stochastic differential equations (sdes). *Advances in Neural Information Processing Systems*, 2021.

Thank You!