Generalization Bounds via Conditional f-Information

Ziqiao Wang [1] Yongyi Mao [2]

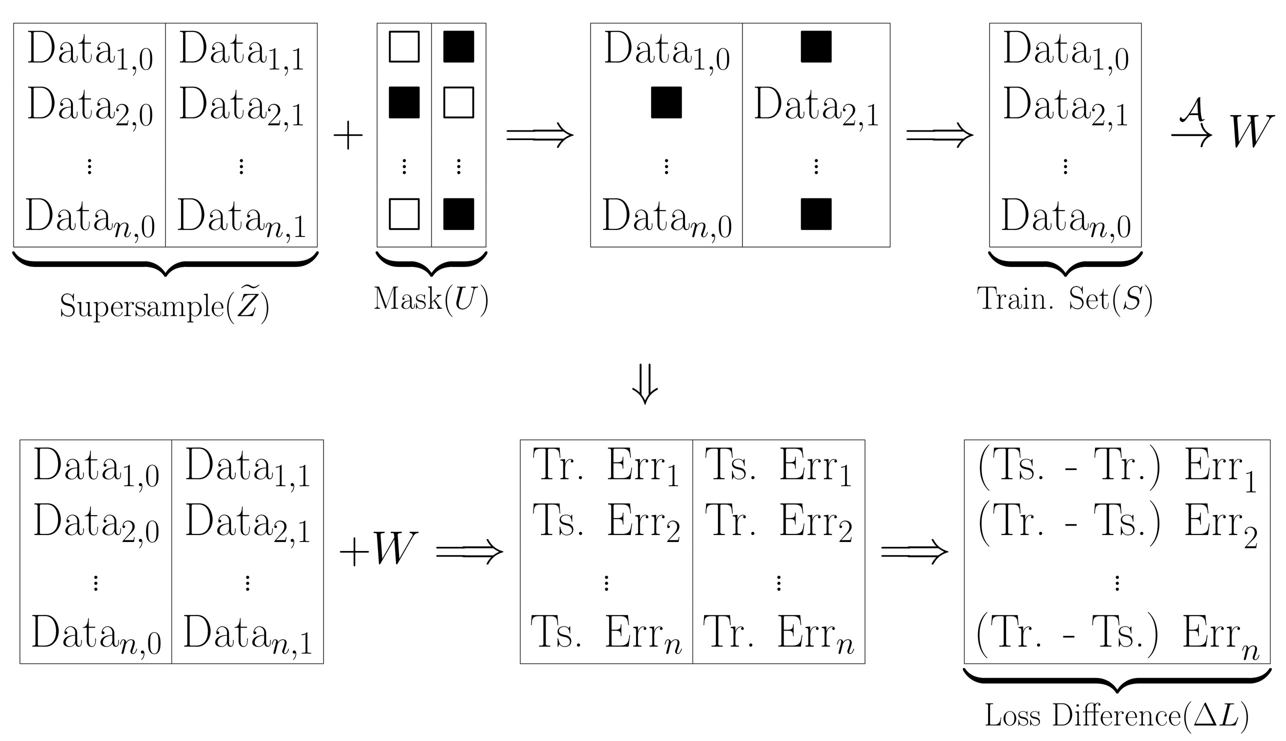
[1]Tongji University [2]University of Ottawa

Generalization

- Learning algorithm $\mathcal{A}: S \to W$ i.e. mapping a training sample to a hypothesis.
- Expected Gen. Err. $= \mathbb{E}$ [Test Err. − Train Err.] $\leq$ Gen. Bound.

Supersample Setting in CMI Framework

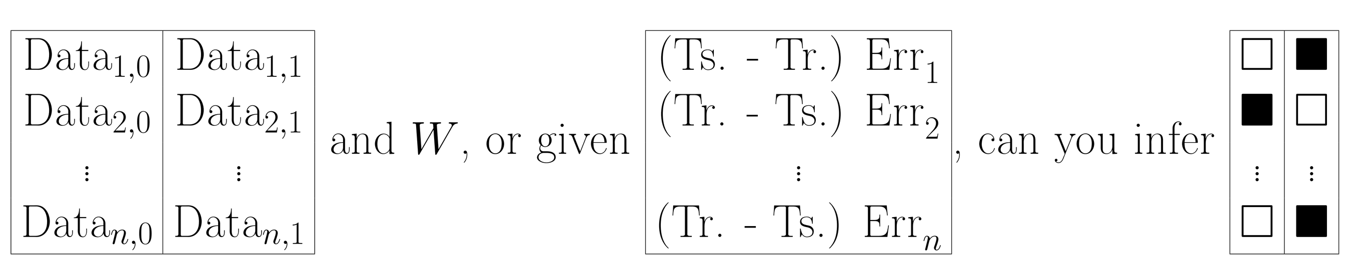
Supersample construction in CMI [3]:



- Data drawn i.i.d. from μ, $U = \{U_i\}_{i=1}^n \overset{i.i.d.}{\sim} \text{Unif}(\{0,1\}^n$.
- CMI Bound [3, 4]: Membership Inference of Train. Set 🧐

$$\text{Gen. Err.} \leq \mathcal{O}\left(\sqrt{\frac{I(\Delta L; U)}{n}}\right) \leq \mathcal{O}\left(\sqrt{\frac{I(W; U|\text{Supersample})}{n}}\right).$$

$\Downarrow$



Main Contributions

- We present a generic approach to derive generalization bounds based on conditional f-information, a natural extension from MI to other f-divergence-based dependence measures.
- For the MI case, our bound recovers many previous CMI bounds and implies some novel fast-rate bounds.
- We present several other f-information-based bounds, including the looser measure, χ^2-information and tighter measures, squared Hellinger-information and Jensen-Shannon-information.

Background on f-Divergence

- (f-Divergence) Let P and Q be two distributions on Θ. Let $\phi: \mathbb{R}_+ \to \mathbb{R}$ be a convex function with $\phi(1) = 0$. If $P \ll Q$, then $\mathrm{D}_\phi(P||Q) \triangleq \mathbb{E}_Q\left[\phi\left(\frac{dP}{dQ}\right)\right]$, e.g., Total variation, KL, χ^2, squared Hellinger, Jeffreys, Jensen-Shannon, etc.
- Variational Representation of f-divergence.

$$\mathrm{D}_\phi(P||Q) = \sup_{g\in\mathcal{G}} \mathbb{E}_{\theta\sim P}[g(\theta)] - \mathbb{E}_{\theta\sim Q}[\phi^*(g(\theta))]. \quad (1)$$

- Let $I_\phi(X;Y) \triangleq \mathrm{D}_\phi(P_{X,Y}||P_XP_Y)$ be the f-information

Conditional f-Information Bounds

Recall variational representation:

$$I_\phi(P||Q) = \sup_{g\in\mathcal{G}} \mathbb{E}_{P_{X,Y}}[g(X,Y)] - \mathbb{E}_{P_XP_{Y'}}[\phi^*(g(X,Y'))].$$

Lemma 1 (informal): Variational Formula of f-Information

Let $g = \phi^{*-1} \circ (tf)$ and if $\mathbb{E}_{X,Y'}[f(X,Y')] = 0$, then

$$\sup_t \mathbb{E}_{X,Y}\left[\phi^{*-1}(tf(X,Y))\right] \leq I_\phi(X;Y).$$

Mutual Information (KL-based) Generalization Bounds

- KL divergence $\Longrightarrow \begin{cases} \phi(x) = x\log x + x - 1 \\ \phi^*(y) = e^y - 1 \\ \phi^{*-1}(z) = \log(1+z) \end{cases}$
- "Oracle" Bound: Assume the loss difference $\in [-1, 1]$

$$|\text{Gen. Err.}| \leq \frac{1}{n}\sum_{i=1}^n \sqrt{2\left(\mathbb{E}\left[\Delta L_i^2\right] + |\mathbb{E}\left[G_i\right]|\right) I(\Delta L_i; U_i)},$$

where $G_i \triangleq (-1)^{U_i}\Delta L_i$.

$\Downarrow$

- Existing Bounds $\begin{cases} \mathcal{O}\left(\frac{1}{n}\sum_{i=1}^n I(\Delta L_i; U_i)\right) & \text{realizable setting,} \\ \mathcal{O}\left(\frac{1}{n}\sum_{i=1}^n \sqrt{I(\Delta L_i; U_i)}\right) & \text{otherwise.} \end{cases}$
- New Bounds $\begin{cases} \frac{1}{n}\sum_{i=1}^n \left(2I(\Delta L_i; U_i) + 2\sqrt{2\text{Var}\left(\text{Single Col. Err.}\right) I(\Delta L_i; U_i)}\right) \\ \frac{1}{n}\sum_{i=1}^n \left(\sqrt{2\mathbb{E}\left[\Delta L_i^2\right] I(\Delta L_i; U_i)} + \sqrt{2\mathbb{E}_{U_i}\left[\mathrm{D_{TV}}\left(P_{\Delta L_i|U_i}, P_{\Delta L_i}\right)\right] I(\Delta L_i; U_i)}\right) \end{cases}$

Other f–Information-based Generalization Bounds

Table 1. Generalization Bounds for χ^2-divergence, Squared Hellinger (SH) Distance and Jensen-Shannon (JS) divergence.

Div.	$\phi(x)$	$\phi^*(y)$	$\phi^{*-1}(z)$	Oracle Bound
χ^2	$(x-1)^2$	$\frac{y^2}{4} + y$	$2(\sqrt{z+1} - 1)$	$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\widetilde{Z}_i}\sqrt{2\left(\mathbb{E}\left[\Delta L_i^2\|\widetilde{Z}_i\right] + \left\|\mathbb{E}\left[G_i\|\widetilde{Z}_i\right]\right\|\right) I_{\chi^2}^{\widetilde{Z}_i}(\Delta L_i; U_i)}$
SH	$(\sqrt{x} - 1)^2$	$\frac{y}{1-y}$	$\frac{z}{1+z}$	$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\widetilde{Z}_i}\sqrt{\left(4\mathbb{E}\left[\Delta L_i^2\|\widetilde{Z}_i\right] + 2\left\|\mathbb{E}\left[G_i\|\widetilde{Z}_i\right]\right\|\right) I_{\mathrm{H}^2}^{\widetilde{Z}_i}(\Delta L_i; U_i)}$
JS	$x\log\frac{2x}{1+x} + \log\frac{2}{1+x}$	$-\log(2 - e^y)$	$\log(2 - e^{-z})$	$\frac{1}{n}\sum_{i=1}^n 2\mathbb{E}_{\widetilde{Z}_i}\sqrt{\left(4\mathbb{E}\left[\Delta L_i^2\|\widetilde{Z}_i\right] + \left\|\mathbb{E}\left[G_i\|\widetilde{Z}_i\right]\right\|\right) I_{\mathrm{JS}}^{\widetilde{Z}_i}(\Delta L_i; U_i)}$

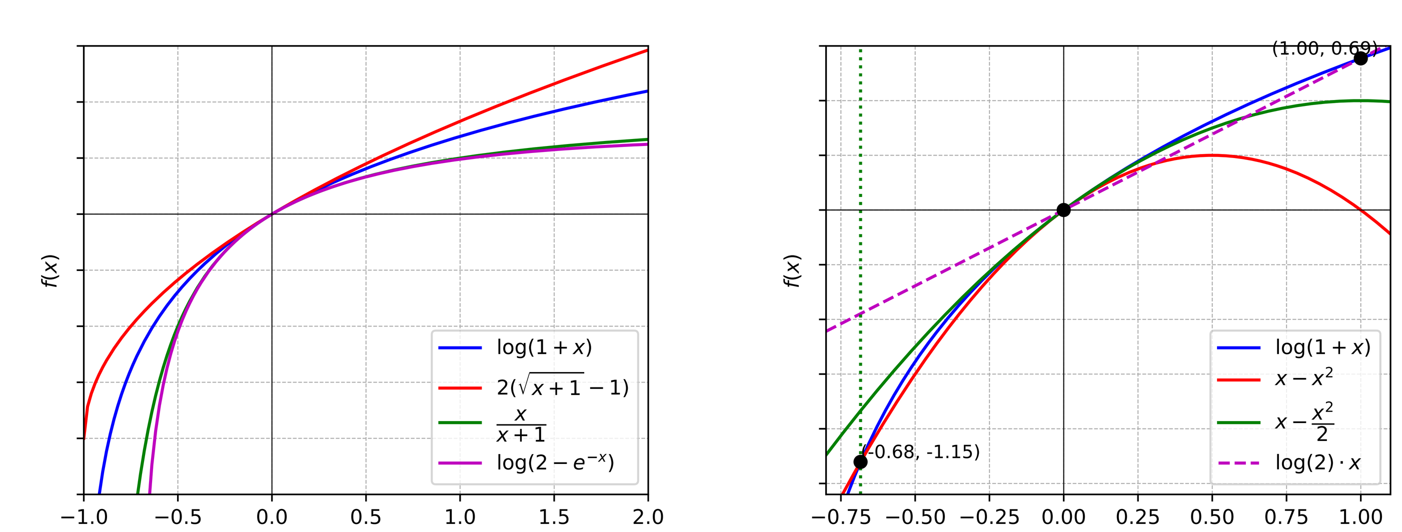


Figure 1. Comparison of different ϕ^{*-1} (Left) and examples of $x - ax^2$ for lower-bounding $\log(1+x)$ (Right).

Proof Sketch (KL)

- Step 1: Lemma 1 implies $I(\Delta L_i; U_i) \geq \sup_t \mathbb{E}\left[\log\left(1 + t(-1)^{U_i}\Delta L_i\right)\right]$.
- Step 2: Let $f(x) = \log(1+x) - x + ax^2$ and set $a = \frac{|\mathbb{E}[G_i]|}{2\mathbb{E}[G_i^2]} + \frac{1}{2}$.
  Ineq. (inspired by [1]): $f(x) \geq 0$ holds when $a \geq \frac{1}{2}$ and $|x| \leq 1 - \frac{1}{2a}$.
- Step 3: $\sup_{t>-1}\mathbb{E}\left[\log\left(1 + tG_i\right)\right] \geq \sup_{t\in[\frac{1}{2a}-1, 1-\frac{1}{2a}]}\mathbb{E}\left[tG_i - at^2G_i^2\right]$.
  The supremum is attained when $t^* = \frac{\mathbb{E}[G_i]}{2a\mathbb{E}[G_i^2]}$, which is achievable.
- Step 4: $I(\Delta L_i; U_i) \geq \sup_{t>-1}\mathbb{E}_{\Delta L_i, U_i}\left[\log\left(1 + t(-1)^{U_i}\Delta L_i\right)\right] \geq \frac{\mathbb{E}^2[G_i]}{4a\mathbb{E}[G_i^2]}$, which simplifies to

$$|\mathbb{E}[G_i]| \leq \sqrt{2\left(|\mathbb{E}\left[G_i\right]| + \mathbb{E}\left[G_i^2\right]\right) I(\Delta L_i; U_i)}.$$

Ineqs for SH & JS: $\begin{cases} \frac{x}{1+x} \geq x - ax^2 & \text{for } a \geq 1 \text{ and } x \in \left[\frac{1}{a} - 1, 1 - \frac{1}{a}\right], \\ \log(2 - e^{-x}) \geq x - ax^2 & \text{for } a \geq 4 \text{ and } x \in \left[-\frac{1}{2}, \frac{1}{2}\right]. \end{cases}$

Extension: Unbounded Case

- Key Idea: Truncation + Special f-divergence $\mathrm{D}_{\phi_\alpha}(P||Q) \triangleq \mathbb{E}_Q\left[\left(\frac{dP}{dQ} - 1\right)^\alpha\right]$ [2]

Lemma 2 (informal): Truncated Variational Formula

Let ε be a Rademacher variable, and $t \in (-b, b)$. If $\phi^*(0) = 0$, then

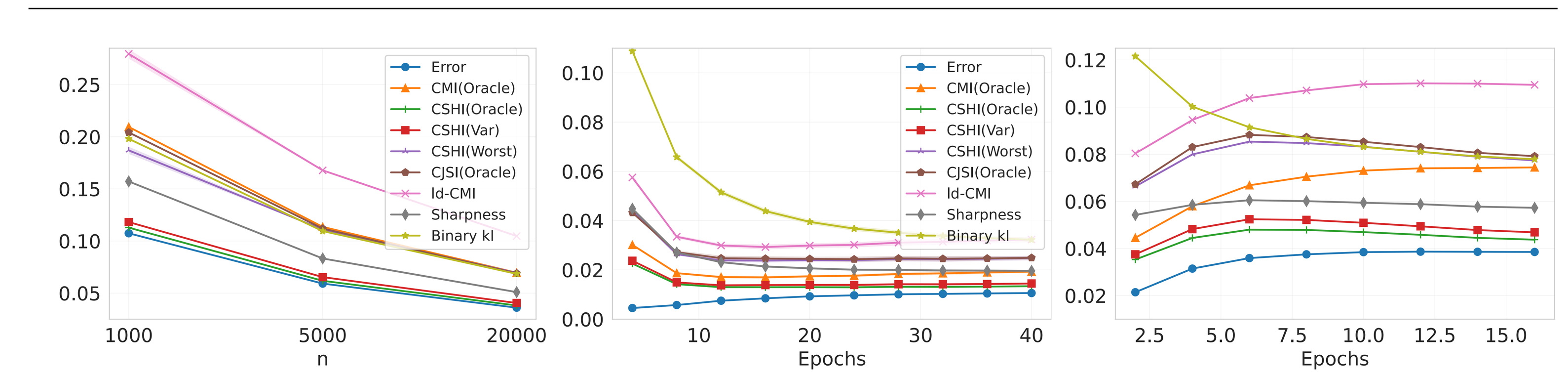
$$\sup_{t\in(-b,b)} \mathbb{E}_{X,\varepsilon}\left[\phi^{*-1}(t\varepsilon X) \cdot \mathbb{1}_{|X|\leq C}\right] \leq I_\phi(X;\varepsilon).$$

- Final Bound: For constants $C \geq 0$, $q, \alpha, \beta \geq 1$ s.t. $\frac{1}{\alpha} + \frac{1}{\beta} = 1$,

$$|\text{Gen. Err.}| \leq \inf_{C,q,\alpha,\beta}\frac{1}{n}\sum_{i=1}^n\left(\zeta_1\sqrt{I(\Delta L_i; U_i)} + \zeta_2\sqrt[\alpha]{I_{\phi_\alpha}(\Delta L_i; U_i))}\right),$$

where ζ_1 and ζ_2 are terms related to tail behavior, controlled by C, q and β.

Empirical Results



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

Figure 2. Comparison of bounds on MNIST ("4 vs 9") and CIFAR10. (a) Dynamics of generalization bounds as dataset size changes. (b-c) Dynamics of generalization bounds during SGLD training.

References

[1] Kyoungseok Jang, Kwang-Sung Jun, Ilja Kuzborskij, and Francesco Orabona. Tighter pac-bayes bounds through coin-betting. In The Thirty Sixth Annual Conference on Learning Theory, pages 2240–2264. PMLR, 2023.

[2] Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Dependence measures bounding the exploration bias for general measurements. In 2017 IEEE International Symposium on Information Theory (ISIT), pages 1475–1479. IEEE, 2017.

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

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