Sample-Conditioned Hypothesis Stability Sharpens Information-Theoretic Generalization Bounds

Ziqiao Wang[1] Yongyi Mao[1]

[1]University of Ottawa

November 12, 2023

Motivation & Contribution

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

Motivation & Contribution

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

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]

Motivation & Contribution

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

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]

Motivation & Contribution

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

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]

Motivation & Contribution

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

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

Novel Construction

By using Donsker-Varadhan (DV) lemma:

$$\text{Gen. Err.} \leq \inf_{t>0} \frac{\text{IOMI or CMI} + \text{CGF}}{t}. \tag{1}$$

Let f_{DV} be so-called DV auxiliary function, then

$$\mathrm{CGF} = \log \mathbb{E}\left[\exp\left(t \cdot f_{\mathrm{DV}}\right)\right]. \tag{2}$$

Novel Construction

By using Donsker-Varadhan (DV) lemma:

$$\text{Gen. Err.} \leq \inf_{t>0} \frac{\text{IOMI or CMI} + \text{CGF}}{t}. \tag{1}$$

Let f_{DV} be so-called DV auxiliary function, then

$$\mathrm{CGF} = \log \mathbb{E}\left[\exp\left(t \cdot f_{\mathrm{DV}}\right)\right]. \tag{2}$$

Let $\ell(w, z)$ be the loss of hypothesis w evaluated at data z, $U \sim \mathrm{Bern}(\frac{1}{2})$.

- 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 using Donsker-Varadhan (DV) lemma:

$$\text{Gen. Err.} \leq \inf_{t>0} \frac{\text{IOMI or CMI} + \text{CGF}}{t}. \tag{1}$$

Let f_{DV} be so-called DV auxiliary function, then

$$\mathrm{CGF} = \log \mathbb{E}\left[\exp\left(t \cdot f_{\mathrm{DV}}\right)\right]. \tag{2}$$

Let $\ell(w, z)$ be the loss of hypothesis w evaluated at data z, $U \sim \mathrm{Bern}(\frac{1}{2})$.

- 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

Novel Construction

$$\text{Gen. Err.} \leq \inf_{t>0} \frac{\text{IOMI or CMI} + \text{CGF}}{t}.$$

For example,

- Previous CMI: $f_{\mathrm{DV}} = (-1)^U \left(\ell(W, Z_1) - \ell(W, Z_2)\right)$
  $\implies \mathrm{CGF} \leq \frac{t^2\alpha^2}{2}$, where $\alpha = \sup_{w,z_1,z_2} |\ell(w, z_1) - \ell(w, z_2)|$
  $\implies \text{Gen. Err.} \leq \inf_{t>0} \frac{\text{CMI}+\text{CGF}}{t} \lesssim \alpha\sqrt{\text{CMI}}$.

Novel Construction

$$\text{Gen. Err.} \leq \inf_{t>0} \frac{\text{IOMI or CMI} + \text{CGF}}{t}.$$

For example,

- Previous CMI: $f_{\mathrm{DV}} = (-1)^{U} \left(\ell(W, Z_1) - \ell(W, Z_2)\right)$
  $\implies \mathrm{CGF} \leq \frac{t^2\alpha^2}{2}$, where $\alpha = \sup_{w,z_1,z_2} |\ell(w, z_1) - \ell(w, z_2)|$
  $\implies$ Gen. Err. $\leq \inf_{t>0} \frac{\text{CMI}+\text{CGF}}{t} \precsim \alpha\sqrt{\text{CMI}}$.
- New CMI in this paper: $f_{\mathrm{DV}} = (-1)^{U} \left(\ell(W, Z) - \ell(W^{-i}, Z)\right)$
  $\implies \mathrm{CGF} \leq \frac{t^2\beta^2}{2}$, where $\beta = \sup_{w,w^{-i},z} \left|\ell(w, z) - \ell(w^{-i}, z)\right|$
  $\implies$ Gen. Err. $\leq \inf_{t>0} \frac{\text{New CMI}+\text{CGF}}{t} \precsim \beta\sqrt{\text{New CMI}}$.

Novel Construction

$$\text{Gen. Err.} \leq \inf_{t>0} \frac{\text{IOMI or CMI} + \text{CGF}}{t}.$$

For example,

- Previous CMI: $f_{\mathrm{DV}} = (-1)^U \left(\ell(W, Z_1) - \ell(W, Z_2)\right)$
  $\Longrightarrow \mathrm{CGF} \leq \frac{t^2\alpha^2}{2}$, where $\alpha = \sup_{w,z_1,z_2} |\ell(w, z_1) - \ell(w, z_2)|$
  $\Longrightarrow \text{Gen. Err.} \leq \inf_{t>0} \frac{\text{CMI}+\text{CGF}}{t} \precsim \alpha\sqrt{\text{CMI}}$.
- New CMI in this paper: $f_{\mathrm{DV}} = (-1)^U \left(\ell(W, Z) - \ell(W^{-i}, Z)\right)$
  $\Longrightarrow \mathrm{CGF} \leq \frac{t^2\beta^2}{2}$, where $\beta = \sup_{w,w^{-i},z} \left|\ell(w, z) - \ell(w^{-i}, z)\right|$
  $\Longrightarrow \text{Gen. Err.} \leq \inf_{t>0} \frac{\text{New CMI}+\text{CGF}}{t} \precsim \beta\sqrt{\text{New CMI}}$.
  $\Longrightarrow$ β **is the uniform stability parameter!**

SCO

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

$$\text{Gen. Err.} \leq \mathcal{O}(1/\sqrt{n}).$$

- Previous IOMI or CMI bound in these examples:
  $\alpha = \mathcal{O}(1)$ (=Lip. Para.$\times$Diam. of data space)
  and IOMI$\geq$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.

Main Result

Key observation: train. data $Z_i \xrightarrow{\mathcal{A}} W$; test. data $Z_i' \xrightarrow{\mathcal{A}} W^{-i}$.

$$
\begin{aligned}
\mathbb{E}\left[\text{Test err.} - \text{Train err.}\right] =& \mathbb{E}\left[\ell(W, Z_i') - \ell(W, Z_i)\right] \\
=& \mathbb{E}\left[\ell(W^{-i}, Z_i) - \ell(W, Z_i)\right] \\
=& \mathbb{E}\left[\ell(W, Z_i') - \ell(W^{-i}, Z_i')\right]
\end{aligned}
$$

Theorem (Informal.)

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

Main Result

Key observation: train. data $Z_i \xrightarrow{\mathcal{A}} W$; test. data $Z_i' \xrightarrow{\mathcal{A}} W^{-i}$.

$$
\begin{aligned}
\mathbb{E}\left[\text{Test err.} - \text{Train err.}\right] =& \mathbb{E}\left[\ell(W, Z_i') - \ell(W, Z_i)\right] \\
=& \mathbb{E}\left[\ell(W^{-i}, Z_i) - \ell(W, Z_i)\right] \\
=& \mathbb{E}\left[\ell(W, Z_i') - \ell(W^{-i}, Z_i')\right]
\end{aligned}
$$

Theorem (Informal.)

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

- β is necessarily *uniform stability* parameter, e.g., sample-conditioned hypothesis (SCH) stability.
- More bounds, e.g., fast-rate bounds and second-moment bounds.
- More examples, e.g., our bounds can also improve stability-based bounds.
- More results refer to our paper.

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.

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.

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.

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.

References III

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.

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.

Thank You!