Which Algorithms Have Tight Generalization Bounds?

NeurIPS 2025 (Spotlight)

Michael Gastpar Ido Nachum Jonathan Shafer
Thomas Weinberger

November 2025

Generalization Bounds (GBs)

Given learning algorithm $\mathcal{A}$, sample $S \sim \mathcal{D}$:

$$L_{\mathcal{D}}(\mathcal{A}(S)) < b(S)$$

- High probability upper bounds on population risk $L_{\mathcal{D}}(\mathcal{A}(S))$.
- Typically, $b(S) = L_S(\mathcal{A}(S)) + C(\mathcal{A}(S), S)$
- VC bounds, Rademacher, norm/margin-based, sharpness-based, ...

We investigate the **on-average tightness** of **algorithm-dependent** bounds across fixed distribution families $\mathbb{D}$.

Formally: $(\mathcal{A}, \mathbb{D})$ is <u>estimable on average</u> if $\exists \mathcal{E}(S)$ such that

$$\mathbb{P}_{\mathcal{D} \sim \mathsf{U}(\mathbb{D}), S \sim \mathcal{D}^m}\left[\left|\mathcal{E}(S) - L_{\mathcal{D}}(A(S))\right| \leq \varepsilon\right] \geq 1 - \delta.$$

Picking $b(S) = \mathcal{E}(S) + \varepsilon$, where $\mathcal{E}$ viable estimator implies:

$$\mathbb{P}_{\mathcal{D} \sim \mathsf{U}(\mathbb{D}), S \sim \mathcal{D}^m}[b(S) - 2\varepsilon \leq L_{\mathcal{D}}(A(S)) \leq b(S)] \geq 1 - \delta.$$

Important: learnability $\not\Leftrightarrow$ existence of tight bounds

- Example: $\exists$ accurate learner for $\mathbb{D} \not\Rightarrow \exists$ tight bound which can verify learning success across $\mathbb{D}$ for any $\mathcal{A}$!
- Above: example 1.4 in our paper.
- See the thorough discussion in our paper (many more examples).

GBs are Typically Vacuous, **Empirically**

From "Fantastic Generalization Measures and Where to Find Them" (Jiang et al., ICLR 2020):

"...almost all existing bounds are vacuous on current deep learning tasks (combination of models and datasets), and therefore, one cannot rely on their proof as an evidence on the causal relationship between a complexity measure and generalization currently."

Follow-up work: "In Search of Robust Measures of Generalization" (Dziugaite et al., NeurIPS 2020):

- Similar findings for other measures of correlation.

GBs are Typically Vacuous, **Theoretically**

Our previous work, "Fantastic Generalization Measures are Nowhere to be Found" (ICLR 2024):

- Overparametrized setting (no uniform convergence)
- There can be a trade-off between an algorithm's learning performance and the ability to verify successful learning (with a bound).
- Lower bound quantified via total-variation term that depends on $(\mathcal{A}, \mathbb{D})$.

Motivation

Goal 1: devise lower bounds on the tightness of generalization bounds in overparametrized settings that...

- ...make the dependence on the implicit bias of the algorithm explicit.
- ...make the dependence on the complexity of $\mathbb{D}$ more explicit.

Goal 2: conversely, which conditions are sufficient for the existence of tight generalization bounds?

Algorithmic and Distributional Setting

Let $\mathcal{A} : S \mapsto \mathcal{H}$, m sample size.
Let $\mathcal{F} \subseteq \mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ be a hypothesis class.

Our notion of **inductive bias** towards a particular set of hypotheses $\mathcal{F}$:

Definition

Let $\mathcal{F}_S$ denote the hypotheses in $\mathcal{F}$ that interpolate S. We say that a learning rule $\mathcal{A}$ is $\mathcal{F}$-interpolating if $\mathcal{A}(S) \in \mathcal{F}_S$ for every sample S such that $\mathcal{F}_S \neq \varnothing$.

Our notion of **complexity** (together with card($\mathcal{F}$)) of a distribution family $\mathbb{D}$ labelled by $\mathcal{F}$:

Definition

We say that $\mathcal{F}$ is ε-orthogonal, if every distinct $f, g \in \mathcal{F}$ satisfy

$$\left|\mathbb{E}_{x\sim \text{Uniform}(\mathcal{X})}[f(x)g(x)]\right| \leq \varepsilon.$$

Lower Bound 1

Theorem

Let $\mathcal{H}$ have VC dimension $\mathrm{VC}(\mathcal{H}) = d > d_0$, and let $m \leq \sqrt{d}/10$. Then there exists a subset $\mathcal{F} \subseteq \mathcal{H}$ and a collection $\mathbb{D}$ of $\mathcal{F}$-realizable distributions such that for any $\mathcal{F}$-interpolating learning rule A and for any $\mathcal{E}(S)$,

$$\mathbb{P}_{\substack{\mathcal{D}\sim \mathit{Uniform}(\mathbb{D}) \\ S\sim\mathcal{D}^m}}\left(\left|\mathcal{E}(S) - L_{\mathcal{D}}(\mathcal{A}(S))\right| \geq \frac{1}{4}\right) \geq \frac{1}{6}.$$

Lower Bound 2

Theorem

Assume that $\mathcal{A}$ is $\mathcal{F}$-interpolating for a nearly orthogonal set $\mathcal{F}$ consisting of roughly 2^m functions. Then there exists a collection $\mathbb{D}$ of $\mathcal{F}$-realizable distributions such that for any $\mathcal{E}(S)$,

$$\mathbb{P}_{\substack{\mathcal{D}\sim Uniform(\mathbb{D}) \\ S\sim\mathcal{D}^m}}\left(\left|\mathcal{E}(S) - L_{\mathcal{D}}(\mathcal{A}(S))\right| \geq \frac{1}{4} - o(1)\right) \geq 0.16.$$

Upper Bound

Theorem

Let $\mathcal{A}$ be an algorithm that is sufficiently stable with respect to $\mathbb{D}$ (in the sense of remove-multiple loss stability). Then, there exists a tight generalization bound for $(\mathcal{A}, \mathbb{D})$.

Open Questions

Question 1: lower bounds in the $\omega(\sqrt{d}) = m = o(d)$ regime?

Question 2: is there a single quantity that is both sufficient and necessary for the existence of a tight generalization bound?