An Information-Theoretic Lower Bound on the Generalization Error of Autoencoders

Shyam Venkatasubramanian[1,*], Sean Moushegian[1,*], Ahmed Aloui[1,*], Vahid Tarokh[1]

[1]Duke University [*]Equal contribution

Background: Autoencoders

- Autoencoders, originating in the 1980s, have evolved alongside deep neural networks.
- Let X be a random variable on $\mathcal{X} \subseteq \mathbb{R}^d$, with training dataset $\mathcal{I} = \{x^{(i)}\}_{i=1}^N$, where $x^{(i)} \sim \mathcal{D}$.
- An autoencoder $\mathcal{A}_\theta : \mathbb{R}^d \to \mathbb{R}^d$ consists of an encoder $f_\theta : \mathbb{R}^d \to \mathbb{R}^l$ and decoder $g_\theta : \mathbb{R}^l \to \mathbb{R}^d$:

$$\hat{x} = \mathcal{A}_\theta(x) = g_\theta(f_\theta(x)), \quad \text{where:} \quad l < d.$$

- It is trained to approximate the identity map $\zeta(x) = x$, but the bottleneck $l < d$ forces learning an approximation $\hat{\zeta}(x) = \mathcal{A}_\theta(x)$ in expectation over $\mathcal{D}$.
- Training minimizes empirical reconstruction MSE:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left\| x^{(i)} - \mathcal{A}_\theta(x^{(i)}) \right\|^2 = \frac{1}{N} \sum_{i=1}^{N} \left\| x^{(i)} - g_\theta(f_\theta(x^{(i)})) \right\|^2.$$

- Accordingly, autoencoders can be viewed as a method of nonlinear, trainable dimensionality reduction: the latent code retains features that best explain the information in $\mathcal{I}$.

Compression, Approximation, and Generalization

- Learning low-dimensional representations is closely tied to **compression**, motivating work connecting information theory and deep autoencoders.
- Cybenko's Universal Approximation Theorem establishes MLPs are **expressive**:
  - With one hidden layer and n perceptrons, an MLP ψ can approximate any continuous target ξ as $n \to \infty$. Practical constraints enforce finite n, so perfect approximation is not guaranteed.
- We study **generalization error** (population risk), not the generalization gap:
  - Upper bounds exist but are often worst-case and probabilistic.
  - Lower bounds reveal **fundamental limits** on what is achievable; few results address this.
- Our main viewpoint is an inverse analogue of Cybenko's theorem for autoencoders:
  - For a broad class of autoencoders with finite perceptrons per layer, we prove that exact equality between the target ζ and the autoencoder $\mathcal{A}$ is **not achievable**.
  - We derive an information-theoretic **lower bound** on generalization error that depends on the architectural characteristics of $\mathcal{A}$ and the differential entropy of the data distribution $\mathcal{D}$.
  - This bound **cannot be violated** even with infinite data and perfect optimization.

Motivation

- A lower bound on the generalization error provides actionable insights for deep learning practice.
- **Neural architecture search.**
  - A common objective is to find the smallest architecture that attains a performance threshold.
  - Training many candidate architectures to measure performance is computationally expensive.
  - A generalization lower bound can **shrink the architecture search space before any training** by certifying that some architectures cannot achieve a desired threshold.
- **Overfitting detection.**
  - Overfitting is a serious problem, and many characterizations are empirical.
  - A lower bound yields an objective condition with theoretical guarantees:
    - if the training loss descends below the generalization lower bound, the fit is **theoretically too good to be true** and will not generalize.
  - We further hypothesize this can suggest overfitting in regression and classification tasks.
- We present empirical results for both neural architecture search and overfitting detection.

Background: Information Theory and Form of Bound

- Consider the Markov chain $X \mapsto Y \mapsto \hat{X}$, with $X, \hat{X} \in \mathcal{X}$ and $Y \in \mathcal{Y}$. We recall information-theoretic inequalities that lower bound reconstruction error.
- **Fano's Inequality** (discrete support): $\mathbb{P}[\hat{X} \neq X] \geq \frac{H(X|Y)-1}{\log|\mathcal{X}|}$, where $|\mathcal{X}|$ is the cardinality of the support of X, and $H(X|Y)$ is conditional discrete entropy (in bits).
- **EEDE Inequality** (continuous support): $\mathbb{E}\left[(\hat{X}-X)^2\right] \geq \mathrm{Var}(X|Y) \geq \frac{1}{2\pi e}\exp\big(2h(X|Y)\big)$, where $h(X|Y)$ is conditional differential entropy (in nats).
- Using these inequalities, we seek an information-theoretic lower bound on the **generalization mean squared error** of an autoencoder:

$$\mathbb{E}\big[\|X - g(f(X))\|^2\big] \geq \mathcal{F}(d, l, K, h_{\mathcal{D}}).$$

- Parameters in the bound:
  - d: input dimension, l: latent dimension, $h_{\mathcal{D}}$: differential entropy of $\mathcal{D}$,
  - $K = \sup_\theta K_\theta$, where K_θ is the Lipschitz constant of the decoder w.r.t. latent input.

Lower Bound From Injective Noise

- Setup and assumptions:
  - $\mathcal{D}$ is supported on $\mathcal{X} = [0,1]^d$ (bounded inputs; common for images).
  - Autoencoders are sigmoidal (following Cybenko's universal approximation theorem).
- Choice of information-theoretic inequality:
  - Many datasets are discrete (e.g., 256 pixel levels), so Fano's inequality might seem natural.
  - In high dimensions, $|\mathcal{X}|$ is enormous, making Fano's bound unusable.
  - We view samples as continuous (discretization is numerical convenience), enabling EEDE:

$$\mathbb{E}\left[(\hat{X} - X)^2\right] \;\geq\; \mathrm{Var}(X|Y) \;\geq\; \frac{1}{2\pi e}\, e^{2h(X|Y)}.$$

- Fundamental obstacle in applying EEDE directly:
  - The encoder is deterministic ($Y = f(X)$), so $h(Y|X) = -\infty$.
  - Using $h(X|Y) = h(X) - h(Y) + h(Y|X)$ yields a trivial bound: $\mathbb{E}\left[\|X - g(f(X))\|^2\right] \geq 0$.

Lower Bound from Injective Noise (continued)

- We inject noise into the latent space **solely as a proof technique**:

$$Z \sim \mathcal{N}(0_l, \sigma^2 I_l), \quad \text{whereby:} \quad g(f(X)) \implies g(f(X) + Z).$$

- Expanding the squared ℓ_2 norm gives:

$$\underbrace{\mathbb{E}\left[\|g(f(X)) - X\|^2\right]}_{\text{Term 1}} = \underbrace{\mathbb{E}\left[\|g(f(X)) - g(f(X)+Z)\|^2\right]}_{\text{Term 2}} + \underbrace{\mathbb{E}\left[\|X - g(f(X)+Z)\|^2\right]}_{\text{Term 3}} - \underbrace{\mathbb{E}\left[2[g(f(X)) - g(f(X)+Z)]^T[X - g(f(X)+Z)]\right]}_{\text{Term 4}}.$$

  - **Term 1:** the MSE of the original, unperturbed *noiseless* autoencoder.
  - **Term 2:** the difference in reconstruction between the noisy and noiseless decoders.
  - **Term 3:** the MSE of the *noise-injected* autoencoder, where reconstruction is formed from the perturbed latent representation $f(X) + Z$.
  - **Term 4:** a cross-term coupling the decoder perturbation $g(f(X)) - g(f(X) + Z)$ with the noisy reconstruction residual $X - g(f(X) + Z)$.

Lower Bound From Injective Noise (continued)

- **Term 2 lower bound**:

$$\mathbb{E}\left[\|g(f(X)) - g(f(X)+Z)\|^2\right] \;\geq\; 0.$$

- **Term 3 lower bound**:

$$\mathbb{E}\left[\|X - g(f(X)+Z)\|^2\right] \geq \frac{d}{2\pi e}\,\exp\left(\frac{2}{d}\left[h_{\mathcal{D}} - \frac{l}{2}\log\left(2\pi e\left(\frac{1}{4}+\sigma^2\right)\right) + \frac{l}{2}\log(2\pi e\sigma^2)\right]\right).$$

- **Term 4 upper bound**:

$$\mathbb{E}\left[2[g(f(X)) - g(f(X)+Z)]^T[X - g(f(X)+Z)]\right] \leq 2K\sqrt{ld\sigma^2},$$

where K is the upper bound on the Lipschitz constant of g.

Lower Bound From Injective Noise (continued)

- We can now substitute the bounds for Terms 2–4 into Term 1.
- For brevity, we define:

$$s = \sigma^2, \qquad \beta = \exp\Big(\frac{2h_{\mathcal{D}}}{d}\Big), \qquad \alpha = \frac{d}{2\pi e}, \qquad \gamma = 2K\sqrt{ld},$$

  where α, β, γ do not depend on s.

- **Theorem (main).** The lower bound on the generalization MSE of the noiseless autoencoder is:

$$\mathcal{F}(s, d, l, K, h_{\mathcal{D}}) \;=\; \alpha\beta\Big(\frac{s}{\frac{1}{4}+s}\Big)^{\frac{l}{d}} - \gamma\sqrt{s}, \qquad s \in [0, \infty).$$

- The inequality holds for any $s \geq 0$. Accordingly, we seek the maximizer:

$$s^* \in \arg\max_{s \geq 0} \; \mathcal{F}(s, d, l, K, h_{\mathcal{D}}),$$

  and study a numerical approximation $\hat{s}^*$ of s^* since $\frac{\partial \mathcal{F}}{\partial s}$ appears intractable.

Properties of the Lower Bound

- We study the shape of the bound function $\mathcal{F}$ with respect to s; under "practical" architectures:
  - when I is large, $\mathcal{F}$ is non-positive for all s (non-informative),
  - when I is small, $\mathcal{F}$ is positive for some s, ruling out perfect reconstruction.
- Implication: larger latent dimension retains more information about x, improving reconstruction.
- We define the maximum value achieved by the MSE lower bound as:

$$\mathcal{F}_\eta^* \;=\; \max_{s\in[0,\infty)} \mathcal{F}_\eta(s), \qquad \eta \;=\; (d, I, K, h_{\mathcal{D}}).$$

- For analytical purposes, assume $d, I \in \mathbb{Z}^+$, $K \in \mathbb{R}^+$, $h_{\mathcal{D}} \leq 0$, and $s \geq 0$.
- **Extremes**: $\mathcal{F}_\eta(0) = 0, \quad \lim_{s\to\infty} \mathcal{F}_\eta(s) = -\infty, \quad \mathcal{F}_\eta(s) \leq d.$
  - Since $\mathcal{F}_\eta$ is continuous in s, it is bounded above, so $\mathcal{F}_\eta^*$ exists and is finite.
- **Monotonicity for fixed $s \geq 0$**: $\frac{\partial \mathcal{F}_\eta(s)}{\partial K} \leq 0$, $\frac{\partial \mathcal{F}_\eta(s)}{\partial h_{\mathcal{D}}} \geq 0$, and $\mathcal{F}_\eta(s)$ is a decreasing function of I.
  - Increasing decoder expressivity decreases the bound; increasing distributional entropy increases the bound; decreasing latent dimension increases the bound.

Properties of the Lower Bound: The Practical Regime

- Two cases yield bounds that are not of interest:
  - If $\mathcal{D}$ is degenerate/countable ($h_{\mathcal{D}} = -\infty$), then even with $l = 1$ an injective encoder f can be learned (since Y has uncountable support for any $l \geq 1$), so no positive lower bound can hold.
  - If K is constrained very close to 0, the decoder lacks expressivity, yielding overly compressed latent representations and large reconstruction errors.
- We focus on the more common **practical regime**, defined by:

$$\textbf{Assumption 1: } h_{\mathcal{D}} > -\infty, \qquad \textbf{Assumption 2: } K \geq (\pi e \sqrt{2})^{-1} \approx 0.083.$$

- Auxiliary function (used as a proof technique):

$$r(s) = \left(\frac{s}{s + \frac{1}{4}} \right)^{-\frac{l}{d}} \sqrt{s}.$$

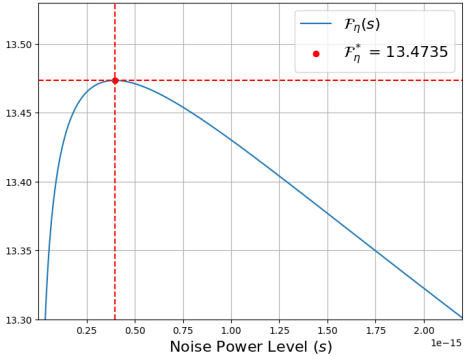
- Key equivalence:

$$\mathcal{F}_\eta(s) > 0 \iff \frac{\alpha\beta}{\gamma} > r(s), \qquad \text{and} \quad \frac{\alpha\beta}{\gamma} \text{ is constant in } s.$$

Properties of the Lower Bound: The Practical Regime (continued)

- **Lemma 1**. Under Assumption 1, if $l < \frac{d}{2}$, then $\exists s_0 > 0$ such that $\mathcal{F}_\eta(s_0) > 0$.
- **Lemma 2**. Under Assumption 2, if $l \geq \frac{d}{2}$, then $\frac{\alpha\beta}{\gamma} \leq \frac{1}{2}$.
- **Lemma 3**. Under Assumption 2, when $l \geq \frac{d}{2}$, $\mathcal{F}_\eta(s) \leq 0$ for all $s > 0$.

- **Principal result in the practical regime**. Under Assumptions 1 and 2: $l < \frac{d}{2} \iff \mathcal{F}_\eta^* > 0$.
- Interpretation: with finite $h_{\mathcal{D}}$ and sufficiently expressive g, the most informative bound is strictly positive precisely when the latent dimension is less than half the input dimension.
- Illustration of positiveness of $\mathcal{F}_\eta(s)$:
  - $d = 784,\ l = 10,\ h_{\mathcal{D}} = -300,\ K = 10^5$.



Properties of the Lower Bound: Most Informative Bound

- In the practical regime, the decoder (with moderate to high K) is sufficiently expressive.
  - Even a small amount of injective noise within the latent layer substantially perturbs the autoencoder reconstruction.
- Accordingly, $g(f(X)) - g(f(X) + Z)$ from Term 4 becomes very large as s increases, especially for large K, implying that s^* must be very small, many orders of magnitude less than one.
- Under the relatively mild assumption that $s^* \ll \frac{1}{4}$, we use: $\frac{s}{s+\frac{1}{4}} \sim 4s, \quad \text{as} \quad s \to 0^+$.
- **Theorem (most informative bound).** The maximizer s^* can be approximated by $\hat{s}^*$, and $\mathcal{F}_\eta^*$ can be approximated by $\hat{\mathcal{F}}_\eta^*$:

$$\hat{s}^* = \left(\frac{4^d \gamma d}{2 \cdot 4^l \alpha \beta l} \right)^{\frac{2d}{2l-d}}, \qquad \hat{\mathcal{F}}_\eta^* = \alpha\beta\left(4\hat{s}^*\right)^{\frac{l}{d}} - \gamma\left(\hat{s}^*\right)^{-\frac{1}{2}}.$$

- Since $\hat{\mathcal{F}}_\eta^*$ is a conservative estimate of $\mathcal{F}_\eta^*$, it follows that $\hat{\mathcal{F}}_\eta^*$ is a valid lower bound on the generalization MSE of $\mathcal{A}_\theta$.

Estimation of Bound Parameters

- The lower bound on the noiseless autoencoder depends on $(d, l, K, h_{\mathcal{D}})$.
  - d, l are known. What about K (decoder Lipschitz constant) and $h_{\mathcal{D}}$ (differential entropy)?
  - We focus on estimation of K and $h_{\mathcal{D}}$ in the regime $l < d/2$.
- **Bounding the Lipschitz constant.**
  - Since g is a sigmoidal MLP, it is Lipschitz for K_θ. We estimate an architecture-dependent *upper bound* K such that $K_\theta \leq K$ for any learnable g of the fixed architecture.
  - Overestimating K_θ underestimates $\mathcal{F}_\eta$, preserving reliability.
  - **Theorem (lower bound insensitivity)**.

$$\left|\frac{\partial \mathcal{F}_\eta^*}{\partial K_\theta}\right| \leq 2\sqrt{lds^*}, \qquad \lim_{K_\theta \to \infty} s^* = 0.$$

  - Two computable upper bounds for K:
    - **Theorem (maximum norm regularization).** If $\|\theta_k\|_F \leq M$ for all layers, $K_\theta \leq \frac{M^{2L}}{16^L} = K$.
    - **Theorem (decimal-64 floating point system).** $K_\theta \leq \sqrt{\left(\frac{10^{36}\, d^2}{16}\right)^L} = K$.

Estimation of Bound Parameters (continued)

- **Estimating the differential entropy.**
  - In high dimensions we typically only have samples $\mathcal{I} = \{x^{(i)}\}_{i=1}^{N}$, not the density $p(\cdot)$.
  - Construct a multivariate kernel density estimator (KDE) following KNIFE.
    - Let $x = (x_1, \ldots, x_d)$ with marginal variances $\sigma_i^2 = \mathrm{Var}[X_i]$.
  - Silverman bandwidth and diagonal covariance (Silverman's rule of thumb):

$$h = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} N^{\frac{-1}{d+4}}, \qquad \Sigma_{ij} = \begin{cases} h^2\sigma_i^2, & i = j \\ 0, & i \neq j \end{cases}.$$

  - Gaussian kernel $\mathcal{K}(x) = \mathcal{N}(0_d, \Sigma)$ and KDE:

$$p(x) \approx p_{\mathrm{KDE}}(x) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{K}(x - x^{(i)}).$$

  - Plug-in sample mean estimator (in nats) of $h_{\mathcal{D}} = -\mathbb{E}_{x\sim\mathcal{D}}[\log p(x)]$:

$$\hat{h}_{\mathcal{D}} \;=\; -\frac{1}{|\mathcal{I}|}\sum_{x\in\mathcal{I}} \log p_{\mathrm{KDE}}(x).$$

Manifold Generalizations

- We generalize the lower bound to cases where earlier assumptions no longer hold.
- Notation: ${}_K\mathcal{A}_{d\mapsto l\mapsto d}$ denotes an autoencoder with input/output dimension d, latent dimension l, and decoder Lipschitz upper bound K.
- **Manifolds of dimension $m < d$.**
  - Many distributions lie on an m-dimensional manifold in a d-dimensional ambient space.
  - Here, $h_{\mathcal{D}} = -\infty$ for $X = (X_1, \ldots, X_d)$, but differential entropy may be finite on the manifold.
  - **Assumption 3**: there exist disjoint U, W with $|U| = m$ and $U \cup W = \{1, \ldots, d\}$ such that:

$$\forall w \in W,\ \exists \Lambda_w \text{ s.t. } x_w = \Lambda_w(x_U), \qquad \forall u \in U,\ h(X_u \mid (X_i)_{i \in U \setminus u}) > -\infty.$$

  - **Lemma 4**. Under Assumption 3, $h(X_U) > -\infty$.
  - **Theorem (valid manifold bound)**: suppose $\mathcal{D}$ violates Assumption 1 but obeys Assumption 3, let $\mathbb{U}$ be the collection of sets U satisfying Assumption 3, and let $X \sim \mathcal{D}$. Then:

$$\mathbb{E}\left[\left\| {}_K\mathcal{A}_{d\mapsto l\mapsto d}(X) - X \right\|^2\right] \geq \max_{U \in \mathbb{U}} \mathcal{F}^*_{(|U|, l, K, h(X_U))}.$$

Manifold Generalizations (continued)

- **Distributions with a degenerate marginal component.**
  - In practice, binning can produce effectively constant features (e.g., MNIST top-left pixel), yielding $h_{\mathcal{D}} = -\infty$ and $\mathcal{F}_\eta^* = 0$.
  - Separate indices into zero-variance components W and the remaining components U:

$$W = \{i : 1 \le i \le d,\ \sigma_i^2 = \mathrm{Var}[X_i] = 0\},$$
$$U = \{1, \dots, d\} \setminus W.$$

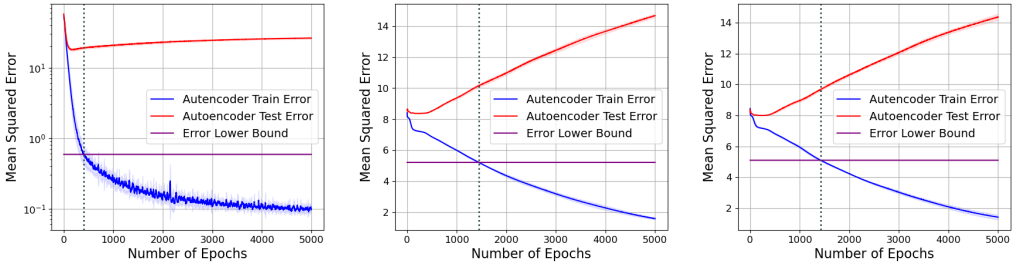
  - **Corollary (masking reconstructor).** Let $X = (X_1, \dots, X_d) \sim \mathcal{D}$. Then:

$$\mathbb{E}\left[\left\| {}_K\mathcal{A}_{d \mapsto l \mapsto d}(X) - X \right\|^2\right] \ge \mathcal{F}^*_{(|U|, l, K, h(X_U))}.$$

  - For numerical stability, allow W to include near-zero variance components below a threshold $\lambda \approx 0$; this works well when $(X_i)_{i \in W}$ is independent of X_U.

Empirical Results: Overfitting Detection in Reconstruction

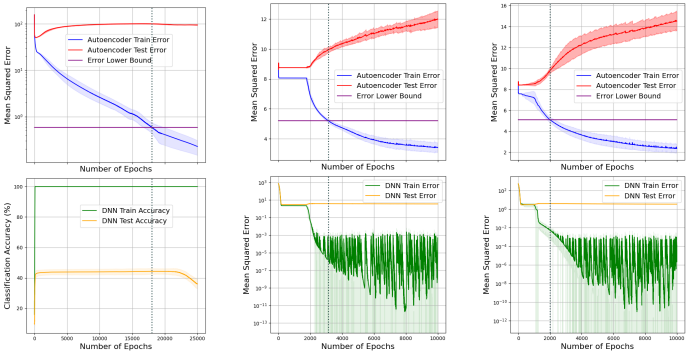
- **Setup**: autoencoders with $\mathcal{X} = [0,1]^d$. Overfitting detected when training MSE crosses $\hat{\mathcal{F}}^*_\eta$.
- MNIST: $d = 784,\ l = 60,\ |\mathcal{I}| = 500,\ |\mathcal{I}_{\text{val}}| = 10{,}000,\ \hat{h}_{\mathcal{D}} \approx -494.5,\ \lambda = 10^{-2}$.
- $\mathcal{TN}$: $d = 100,\ l = 2,\ h_{\mathcal{D}} \approx -1.036$.
- $\mathcal{U}([0,1]^d)$: $d = 100,\ l = 2,\ h_{\mathcal{D}} = 0,\ |\mathcal{I}| = 50,\ |\mathcal{I}_{\text{val}}| = 10{,}000$.
- Decoder constraint: maximum norm regularization with $M = 2.25,\ L = 2 \Rightarrow K_\theta \leq K \approx 0.1$.



(a) MNIST (b) $\mathcal{U}$ (c) $\mathcal{TN}$

Empirical Results: Overfitting Detection in Regression and Classification

- Network: $\hat{V} = t(f(X))$ with $f : \mathbb{R}^d \to \mathbb{R}^l$, $t : \mathbb{R}^l \to \mathbb{R}^k$.
- Append a decoder g to form $\mathcal{A}_\theta(\cdot) = g(f(\cdot))$; overfitting in $\mathcal{A}_\theta$ accompanies overfitting in $t(f(\cdot))$.
- MNIST: $|\mathcal{I}| = 20,\ k = 10$. $\mathcal{U}([0,1]^d)$ and $\mathcal{TN}$: $|\mathcal{I}| = 50,\ k = 1$; $K \approx 0.1$.



(a) MNIST (CLS) (b) $\mathcal{U}$ (REG) (c) $\mathcal{TN}$ (REG)

Empirical Results: Neural Architecture Search

- Search over latent dimensions $\mathbb{L} = \{l_1, \ldots, l_Q\}$ for a target validation threshold $\tau > 0$.
- For each $l_i \in \mathbb{L}$, compute $\hat{\mathcal{F}}^*_{\eta_i}$, $\eta_i = (d, l_i, K, h_{\mathcal{D}})$: if $\tau > \hat{\mathcal{F}}^*_{\eta_i}$, train; else **skip** l_i.
- $\mathcal{U}$: $d = 100,\ K \approx 0.1,\ L = 2,\ |\mathcal{I}| = 5000,\ |\mathcal{I}_{\text{val}}| = 10{,}000,\ h_{\mathcal{D}} = 0$.
- $\mathcal{TN}$: $d = 100,\ K \approx 0.1,\ L = 2,\ |\mathcal{I}| = 5000,\ |\mathcal{I}_{\text{val}}| = 10{,}000,\ h_{\mathcal{D}} \approx -1.036$.
- Example: $\mathbb{L} = \{1, 2, 4, 10, 20, 40, 46, 48\}$, $\tau = 4.5 \Rightarrow$ reduced search $\mathbb{L} \setminus \{1, 2, 4\}$.

$l_i \in \mathbb{L}$	$l_i = 1$		$l_i = 2$		$l_i = 4$		$l_i = 10$	
Dataset	**MSE**	$\hat{\mathcal{F}}^*_\eta$	**MSE**	$\hat{\mathcal{F}}^*_\eta$	**MSE**	$\hat{\mathcal{F}}^*_\eta$	**MSE**	$\hat{\mathcal{F}}^*_\eta$
$\mathcal{U}$	8.34±0.01	5.49	8.33±0.01	5.21	8.22±0.01	4.74	7.66±0.01	3.62
$\mathcal{TN}$	7.99±0.01	5.38	7.97±0.02	5.10	7.89±0.02	4.63	7.32±0.01	3.53
$l_i \in \mathbb{L}$	$l_i = 20$		$l_i = 40$		$l_i = 46$		$l_i = 48$	
Dataset	**MSE**	$\hat{\mathcal{F}}^*_\eta$	**MSE**	$\hat{\mathcal{F}}^*_\eta$	**MSE**	$\hat{\mathcal{F}}^*_\eta$	**MSE**	$\hat{\mathcal{F}}^*_\eta$
$\mathcal{U}$	6.72±0.01	2.20	4.98±4e-3	0.32	4.46±2e-3	0.03	4.29±4e-3	0.002
$\mathcal{TN}$	6.42±0.01	2.13	4.77±3e-3	0.29	4.27±3e-3	0.03	4.10±4e-3	1e-3

Conclusion

Takeaways

- **Theory:** We derive an information-theoretic lower bound on the generalization MSE of sigmoidal autoencoders via the EEDE inequality.
- **Regime:** In the practical regime, the most informative bound is strictly positive precisely when the latent dimension satisfies $l < d/2$.
- **Practice:** The bound enables overfitting detection without a validation set and can prune neural architecture search before expensive training.

Paper & Updates

Acknowledgements: Shyam Venkatasubramanian and Vahid Tarokh were supported in part by the U.S. Air Force Office of Scientific Research under award FA9550-21-1-0235.