PAC-Chernoff Bounds: Understanding Generalization in the Interpolation Regime

Andrés R. Masegosa and **Luis A. Ortega**

ECAI, November 2025

https://www.jair.org/index.php/jair/article/view/17036

Aalborg University and Autonomous University of Madrid

Motivations

Generalization bounds that solely depend on the training data are **provably vacuous** for overparameterized model classes; unable to explain generalization.

$$L(\boldsymbol{\theta}) \leq \hat{L}(D, \boldsymbol{\theta}) + \mathcal{O}\left(\sqrt{\frac{p}{n}}\right)$$

Why current machine learning techniques find overparameterized **interpolators with strong generalization** performance is an **open question**.

Contributions

A **perfectly tight distribution-dependent PAC-Chernoff bound** for **interpolators**, even in over-parameterized models.

$$L(\boldsymbol{\theta}) \leq \hat{L}(D, \boldsymbol{\theta}) + C_{\nu,\boldsymbol{\theta}}\left(\frac{p}{n}\right)$$

where ν is the data-generating distribution.

Contributions

A **perfectly tight distribution-dependent PAC-Chernoff bound** for **interpolators**, even in over-parameterized models.

$$L(\boldsymbol{\theta}) \leq \hat{L}(D, \boldsymbol{\theta}) + C_{\nu,\boldsymbol{\theta}}\left(\frac{p}{n}\right)$$

where ν is the data-generating distribution.

A **theoretical framework** that explains why some interpolators generalize well, while others do not, based on a novel **characterization of smoothness**.

Contributions

A **perfectly tight distribution-dependent PAC-Chernoff bound** for **interpolators**, even in over-parameterized models.

$$L(\boldsymbol{\theta}) \leq \hat{L}(D, \boldsymbol{\theta}) + C_{\nu,\boldsymbol{\theta}}\left(\frac{p}{n}\right)$$

where ν is the data-generating distribution.

A **theoretical framework** that explains why some interpolators generalize well, while others do not, based on a novel **characterization of smoothness**.

We explain why regularization, data augmentation, invariant architectures, and over-parameterization produce smoother interpolators with superior generalization.

The Rate Function

Chernoff's Theorem. For any fixed $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $a > 0$, it satisfies

$$\mathbb{P}_{D\sim\nu^n}\Big(L(\boldsymbol{\theta}) - \hat{L}(D,\boldsymbol{\theta}) \geq a\Big) \leq e^{-n\mathcal{I}_{\boldsymbol{\theta}}(a)}\,.$$

with

$$\mathcal{I}_{\boldsymbol{\theta}}(a) := \sup_{\lambda>0}\ \lambda a - J_{\boldsymbol{\theta}}(\lambda) \quad \text{and} \quad J_{\boldsymbol{\theta}}(\lambda) := \ln \mathbb{E}_{\nu}\Big[e^{\lambda(L(\boldsymbol{\theta})-\ell(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\theta}))}\Big]\,,$$

- **Cramér's Theorem**: For large n, the bound is tight
- **Proposition 3.4**: When a is large, the bound is tight.

Smoother Interpolators Generalize Better

Smoothness: A model $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is β-smoother than a model $\boldsymbol{\theta}' \in \boldsymbol{\Theta}'$

$$\forall a \in (0, \beta] \quad \mathcal{I}_{\boldsymbol{\theta}}(a) \geq \mathcal{I}_{\boldsymbol{\theta}'}(a) \, .$$

Theorem 4.6. For any $\epsilon \geq 0$ with h.p. $1 - \delta$, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\boldsymbol{\theta}' \in \boldsymbol{\Theta}'$, simultaneously,

$$\hat{L}(D, \boldsymbol{\theta}) \leq \epsilon \quad \text{and} \quad \boldsymbol{\theta} \text{ is } \beta\text{-smoother than } \boldsymbol{\theta}'$$

$$\Downarrow$$

$$L(\boldsymbol{\theta}) \leq L(\boldsymbol{\theta}') + \epsilon$$

PAC-Chernoff Bound

Theorem 4.1. With h.p. $1-\delta$, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, simultaneously,

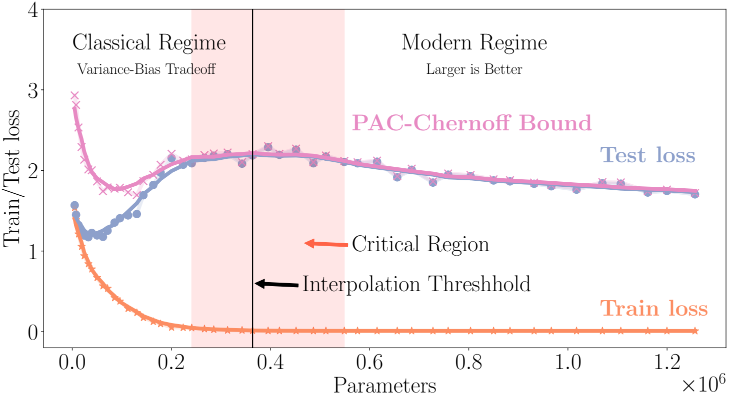
$$L(\boldsymbol{\theta}) \leq \hat{L}(D, \boldsymbol{\theta}) + \mathcal{I}_{\boldsymbol{\theta}}^{-1}(\tfrac{1}{n} \ln \tfrac{k^p}{\delta})\,.$$

with

$$\mathcal{I}_{\boldsymbol{\theta}}^{-1}(s) := \inf_{\lambda > 0} \frac{J_{\boldsymbol{\theta}}(\lambda) + s}{\lambda} \quad \forall s \geq 0\,.$$

Proposition 4.4. The bound is **perfectly tight** for interpolators.

$$L(\boldsymbol{\theta}) \leq \hat{L}(D,\boldsymbol{\theta}) + \mathcal{I}_{\boldsymbol{\theta}}^{-1}\left(\frac{1}{n}\ln\frac{k^p}{\delta}\right) \leq L(\boldsymbol{\theta}) + \hat{L}(D,\boldsymbol{\theta})\,.$$



Optimal Regularization

- The inverse rate is an **regularizer** towards **smoother** models.

$$\boldsymbol{\theta}_\epsilon^\times = \underset{\boldsymbol{\theta}\,:\,\hat{L}(D,\boldsymbol{\theta})\leq\epsilon}{\arg\min} \quad \hat{L}(D,\boldsymbol{\theta}) + \underbrace{\mathcal{I}_{\boldsymbol{\theta}}^{-1}\left(\tfrac{1}{n}\ln\tfrac{k^p}{\delta}\right)}_{\text{Regularizer}},$$

$$\boldsymbol{\theta}_\epsilon^\star = \underset{\boldsymbol{\theta}\,:\,\hat{L}(D,\boldsymbol{\theta})\leq\epsilon}{\arg\min} \quad L(\boldsymbol{\theta})\,.$$

- How close is $\boldsymbol{\theta}_\epsilon^\times$ from the best possible interpolator $\boldsymbol{\theta}_\epsilon^\star$.

Optimal Regularization

- The inverse rate is an **regularizer** towards **smoother** models.

$$\boldsymbol{\theta}_\epsilon^\times = \operatorname*{arg\,min}_{\boldsymbol{\theta}\,:\,\hat{L}(D,\boldsymbol{\theta})\leq\epsilon} \hat{L}(D,\boldsymbol{\theta}) + \underbrace{\mathcal{I}_{\boldsymbol{\theta}}^{-1}\left(\tfrac{1}{n}\ln\tfrac{k^p}{\delta}\right)}_{\text{Regularizer}},$$

$$\boldsymbol{\theta}_\epsilon^\star = \operatorname*{arg\,min}_{\boldsymbol{\theta}\,:\,\hat{L}(D,\boldsymbol{\theta})\leq\epsilon} L(\boldsymbol{\theta}).$$

- How close is $\boldsymbol{\theta}_\epsilon^\times$ from the best possible interpolator $\boldsymbol{\theta}_\epsilon^\star$.
- **Very close!!**

  Theorem 6.1 For any $\epsilon > 0$, with h.p. $1-\delta$ over $D \sim \nu^n$

$$|L(\boldsymbol{\theta}_\epsilon^\star) - L(\boldsymbol{\theta}_\epsilon^\times)| \leq \epsilon\,.$$

  The inverse rate is an **optimal regularizer**.

Understanding Existing Regularizers

Many common regularization techniques are **approximations** to the **optimal regularizer**:

- **Distance from initialization and ℓ_2-norm:**

$$\mathcal{I}_{\boldsymbol{\theta}}^{-1}\left(\frac{1}{n} \ln \frac{k^p}{\delta}\right) \leq \sqrt{2Ma}\ \|\boldsymbol{\theta}\|_2 \,,$$

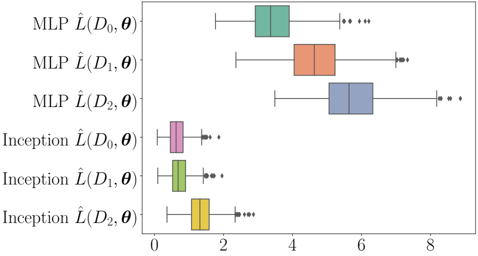
- **Input-gradient norm:**

$$\mathcal{I}_{\boldsymbol{\theta}}^{-1}\left(\frac{1}{n} \ln \frac{k^p}{\delta}\right) \leq \sqrt{\frac{1}{n} \ln \frac{k^p}{\delta}} \sqrt{M \mathbb{E}_\nu\left[\left\|\nabla_{\mathbf{x}} \ell(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta})\right\|_2^2\right]} \,.$$

Transformed Input Data

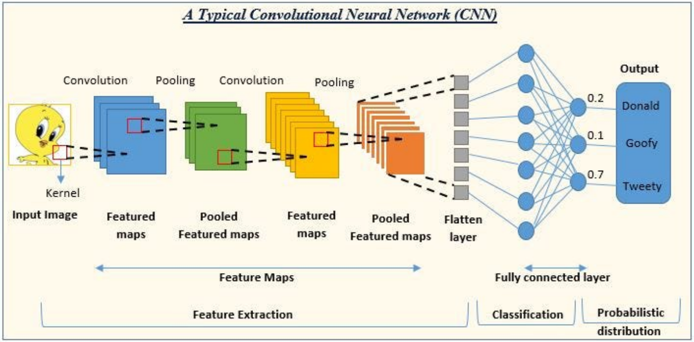
Input data in many machine learning problems undergo **transformations**, often due to the measuring process, such as sensor noise or image distortions.

Transformed input-data makes the expected loss $L(\theta)$ **higher** and the distribution of $\hat{L}(D, \theta)$ with $D \sim \nu^n$ **less concentrated**.



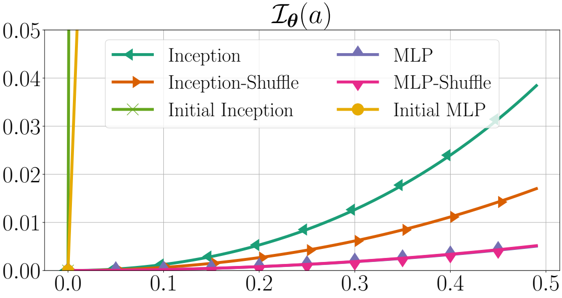
D_0 - CIFAR-10's test set.
D_1 - random translations.
D_2 - random rotations.

Invariant Architectures (I/II)



- PAC-Chernoff bounds explain why interpolating with **invariant architecture** leads to better generalization performance.
- The $\hat{L}(D, \theta)$ of invariant architectures is **more concentrated** under transformed inputs.

Invariant Architectures (II/II)



Model	Train Acc.	Test Acc.	Test NLL
Inception	100.0%	74.08%	1.00
Inception-Shuffle	100.0%	42.46%	2.45
MLP	99.99%	51.69%	3.29
MLP-Shuffle	99.99%	51.12%	3.29
Initial Inception	10.00%	10.00%	2.30
Initial MLP	10.00%	9.96%	2.30

Over-parameterization

Modern machine learning models are **highly overparametrized**.

Previous works have established links between overparametrization and generalization, but under very limited settings.

The distribution-dependent PAC-Chernoff Bound can be used to obtain **bounds over the number of parameters** of **interpolators**:

Theorem 8.1. For any $\epsilon \in (0, L^\star)$ and any $\delta \in (0, 1)$, with high probability $1 - \delta$ over $D \sim \nu^n$, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, simultaneously,

$$\text{if} \quad \hat{L}(D, \boldsymbol{\theta}) \leq \epsilon \quad \text{then} \quad p \geq \frac{n\mathcal{I}_{\boldsymbol{\theta}}(L^\star - \epsilon) + \ln \delta}{\ln k} \,.$$

where $L^\star = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$.

Conclusions and Limitations (I/II)

- Traditional bounds relying solely on training data are **unable** to **explain generalization of over-parameterized interpolators**.
- Distribution-dependent **PAC-Chernoff bounds** are a promising tool able to **explain a wide range of learning techniques**.
- **Smoother interpolators** generalize better.

Conclusions and Limitations (II/II)

- Connected to a **wide range of regularization methods**.
- Explain why **invariant architectures and data-augmentation** works under transformed input-data.
- Over-parameterization is a **neccessary condition** for smooth interpolation.
- **Limitation**: Assumption of a finite model class. It can be addressed by using PAC-Bayes Chernoff bounds **(Casado et al. 2024)**.