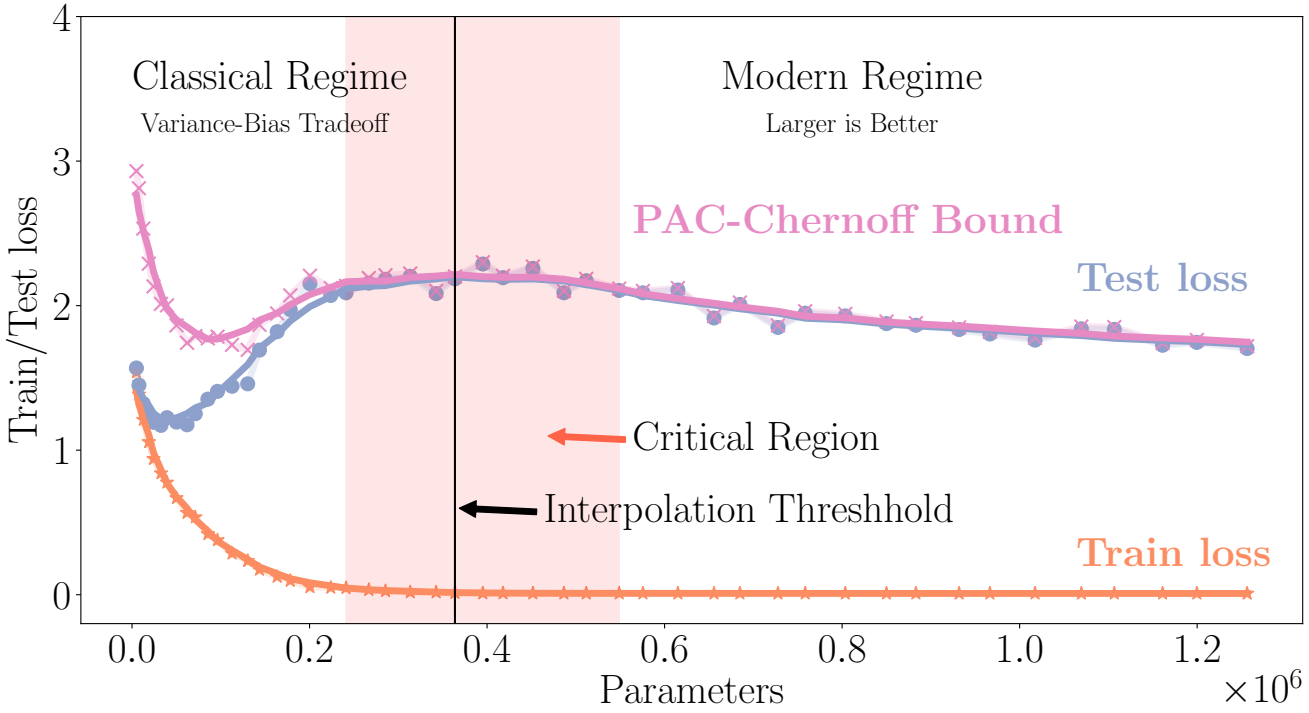
PAC-Chernoff Bounds: Understanding Generalization in the Interpolation Regime

Andrés R. Masegosa, Luis A. Ortega

ECAI, November 2025
https://arxiv.org/abs/2306.10947

Motivations



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

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

- 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.

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

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

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

with

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

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

Inception	Crop	L2	Train Acc.	Test Acc.	Test NLL	ℓ_2-norm
Standard	no	no	99.99%	84.36%	0.65	304
Crop	yes	no	99.94%	86.89%	0.58	309
L2	no	yes	100.0%	86.60%	0.49	200
L2-Crop	yes	yes	99.98%	88.45%	0.42	130
Random	no	no	100.0%	10.13%	5.52	311
Initial	-	-	10.00%	10.00%	2.30	593

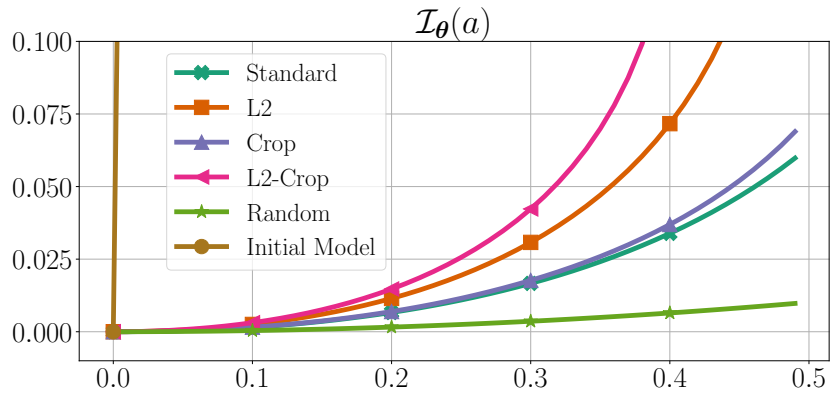


Figure 1: Metrics of Inception models on Cifar10 using ℓ_2 regularization and/or random cropping (Crop), and randomly sampled class labels (Random). The corresponding *rate functions* are shown on the right.

PAC-Chernoff Bound

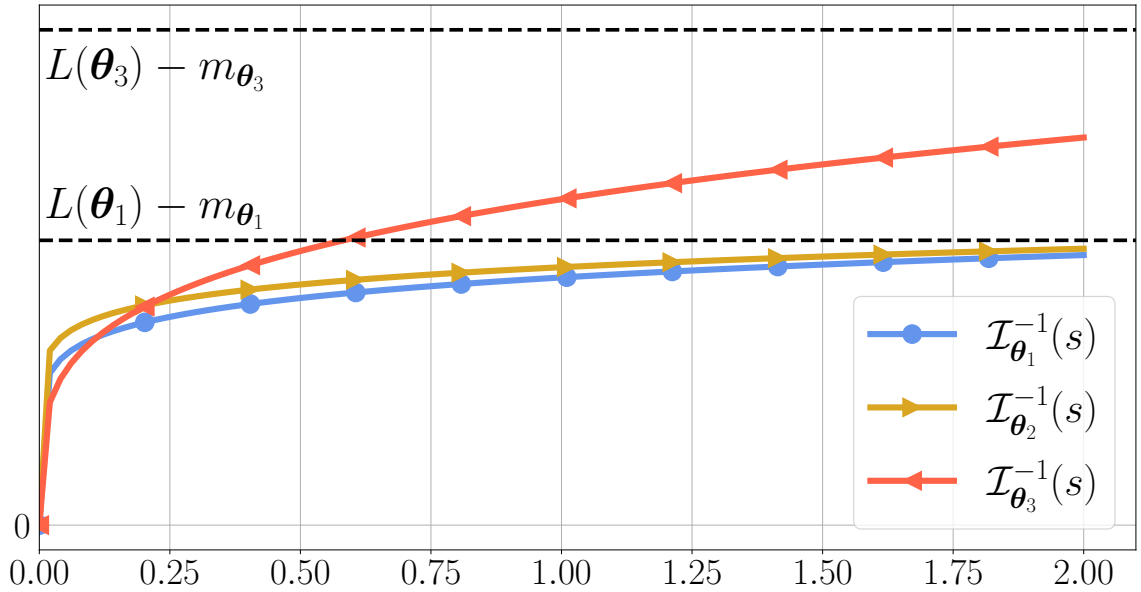
Theorem 4.1. With h.p., for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, simultaneously,

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

with

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

Where p is the number of **parameters** of the model class.



Proposition 4.3. The bound is **perfectly tight** for interpolators.

Smoother Interpolators Generalize Better

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

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

Theorem 4.5. For any $\epsilon \geq 0$, with h.p., for all $\boldsymbol{\theta}, \boldsymbol{\theta}'$, simultaneously,

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

Inception	Crop	L2	Train Acc.	Test Acc.	Test NLL	ℓ_2-norm
Standard	no	no	99.99%	84.36%	0.65	304
Crop	yes	no	99.94%	86.89%	0.58	309
L2	no	yes	100.0%	86.60%	0.49	200
L2-Crop	yes	yes	99.98%	88.45%	0.42	130
Random	no	no	100.0%	10.13%	5.52	311
Initial	-	-	10.00%	10.00%	2.30	593

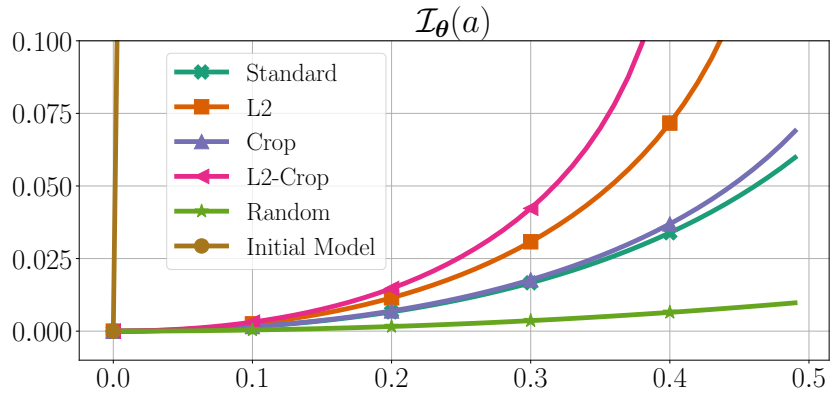


Figure 2: Metrics of Inception models on Cifar10 using ℓ_2 regularization and/or random cropping (Crop), and randomly sampled class labels (Random). The corresponding *rate functions* are shown on the right.

Optimal Regularization

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

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

Optimal Regularization

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

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

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

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

Optimal Regularization

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

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

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

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

- **Very close!!**

  Theorem 5.1 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:

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

Exponential Family and Large Data Sets:

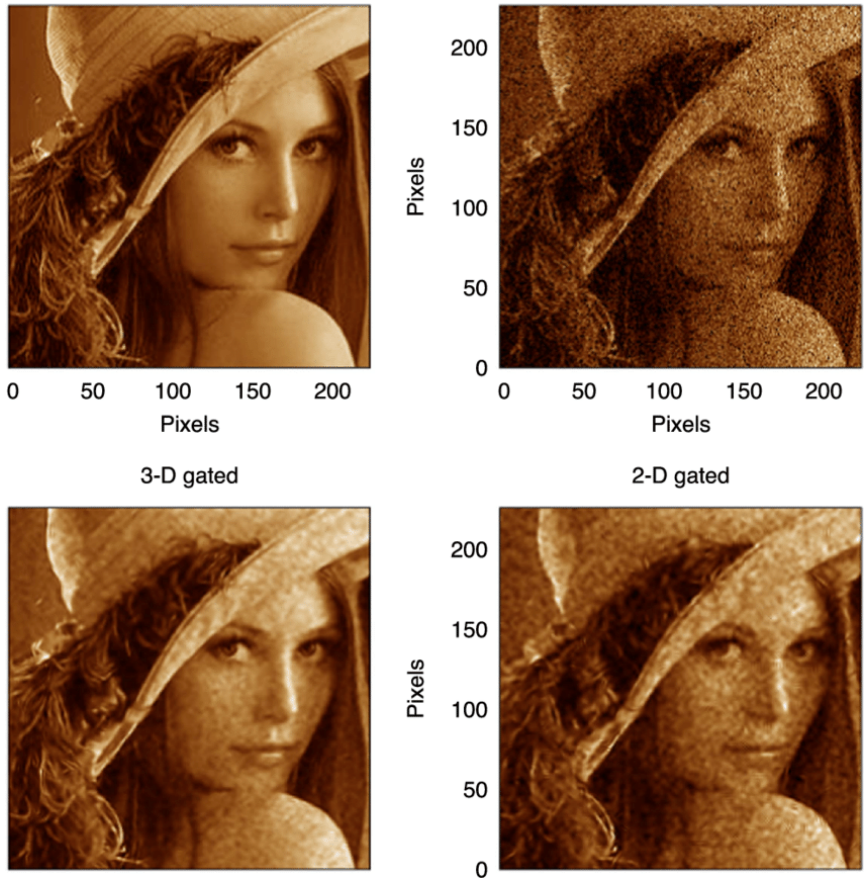
$$\left|\mathcal{I}_{\boldsymbol{\theta}}^{-1}\left(\tfrac{1}{n}\ln\tfrac{k^p}{\delta}\right) - \sqrt{2\tfrac{1}{n}\ln\tfrac{k^p}{\delta}}\sqrt{\boldsymbol{\theta}^T\mathsf{Cov}_\nu(s(\boldsymbol{y},\boldsymbol{x}))\boldsymbol{\theta}}\right| \leq \epsilon\,,$$

Input-gradient norm:

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

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

- **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 $L(\boldsymbol{\theta})$ **higher** and the distribution of $\hat{L}(D, \boldsymbol{\theta})$ with $D \sim \nu^n$ **less concentrated**.

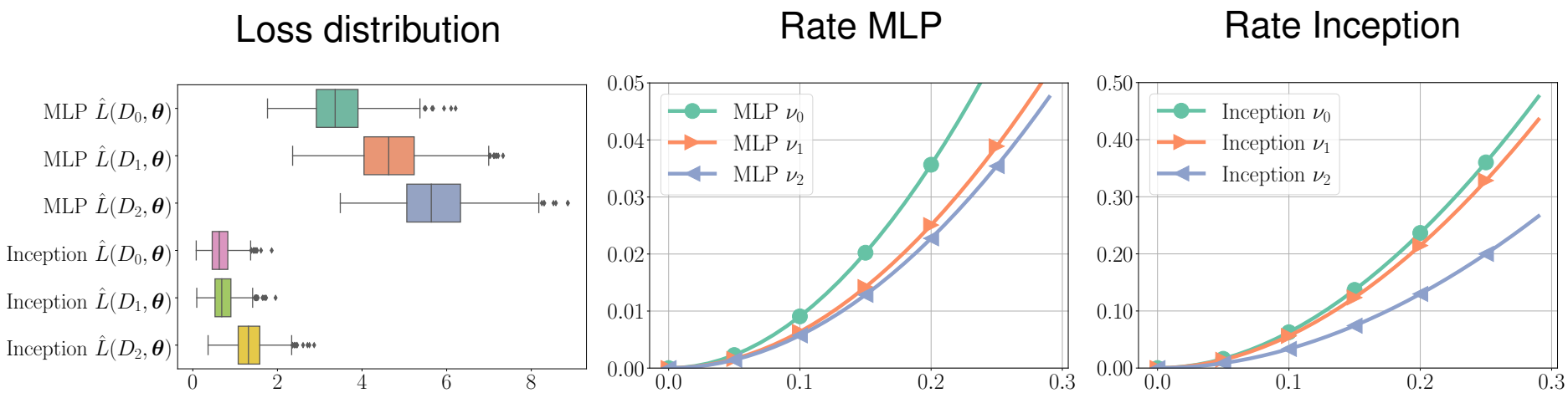
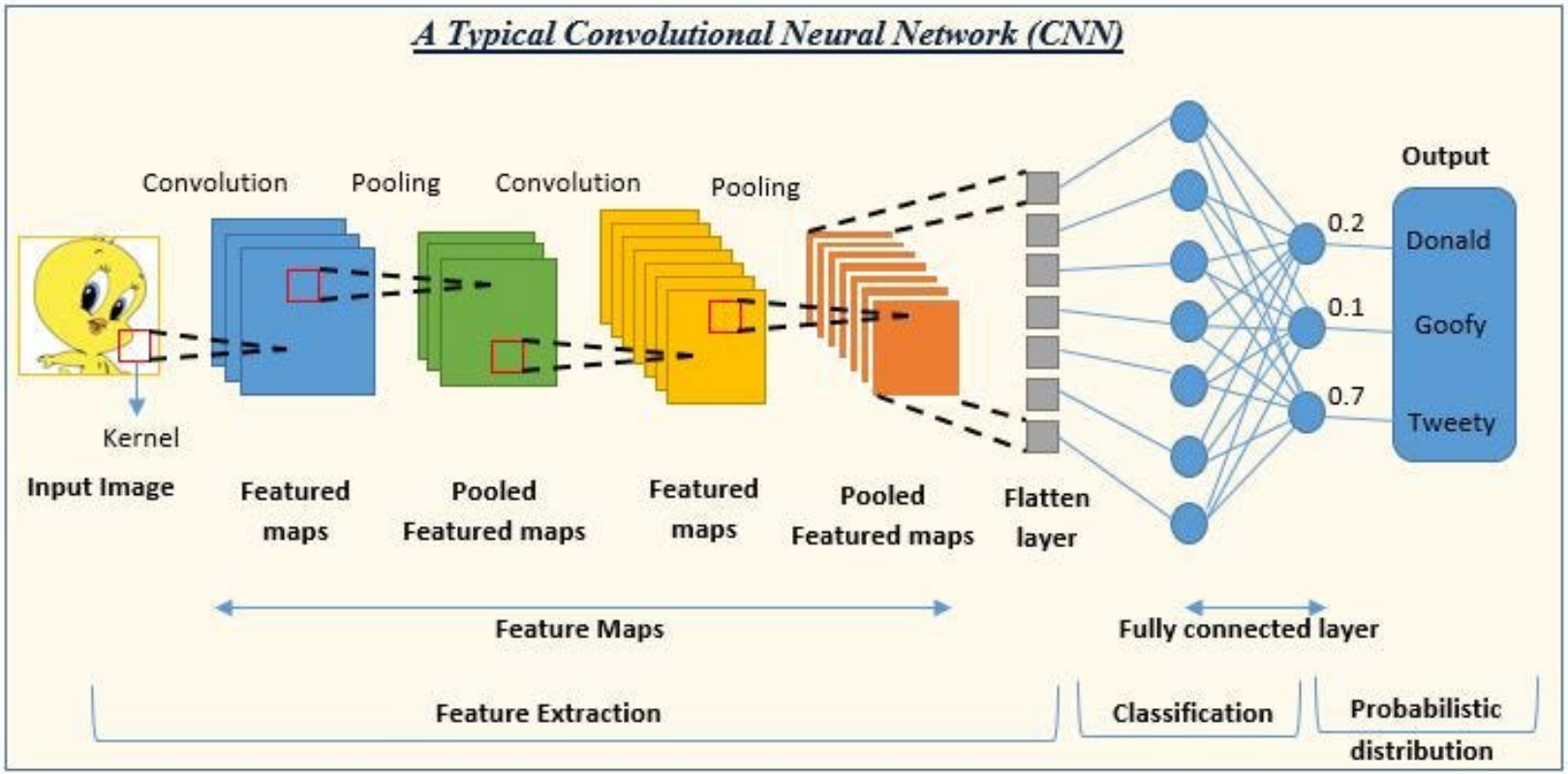


Figure 3: $D_0 \sim \nu_0^{50}$ estimated with CIFAR-10's test set; $D_1 \sim \nu_1^{50}$ adds random translations of 3 pixels; $D_2 \sim \nu_2^{50}$ also adds rotations up to $20°$.



Proposition 6.4 If a model $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is invariant to transformed-inputs ν_{t+1},

$$L^{\nu_{t+1}}(\boldsymbol{\theta}) = L^{\nu_t}(\boldsymbol{\theta}) \quad \text{and} \quad \mathcal{I}_{\boldsymbol{\theta}}^{\nu_{t+1}}(a) = \mathcal{I}_{\boldsymbol{\theta}}^{\nu_t}(a) \quad \forall a > 0\,.$$

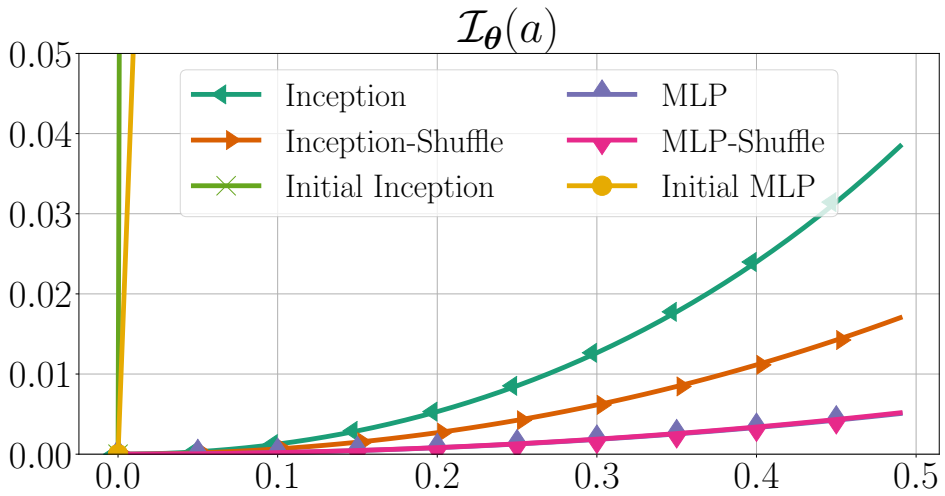
Invariant Architectures

Proposition 6.4 If a model $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is invariant to transformed-inputs ν_{t+1},

$$L^{\nu_{t+1}}(\boldsymbol{\theta}) = L^{\nu_t}(\boldsymbol{\theta}) \quad \text{and} \quad \mathcal{I}_{\boldsymbol{\theta}}^{\nu_{t+1}}(a) = \mathcal{I}_{\boldsymbol{\theta}}^{\nu_t}(a) \quad \forall a > 0\,.$$

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

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

Conclusions and Limitations

- 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.
- 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)**.