From PAC-Chernoff Bounds to Practice: Smooth Interpolators and Near-Optimal Generalization in Deep Learning

Andrés R. Masegosa

Aalborg University (Copenhagen - Denmark)

RSME, January 2026

Masegosa, A. R., Ortega, L. A. (2025). PAC-Chernoff Bounds: Understanding Generalization in the Interpolation Regime. JAIR, ECAI (Spot Light).

Hu J., Ortega L. A., Laleg T. M. , Masegosa , A. R. Towards Near-Optimal Regularization in Deep Learning via the Inverse-Rate Regularizer. To be submitted to ICML 2026.

Why Do We Need Theory in Deep Learning?

The stated goal of learning theory is to guide practice.

Yet modern deep learning success has been driven by:

- Empirical experimentation and trial-and-error
- Architectural heuristics (residual connections, normalization, attention)
- Scale — more data, more parameters, more compute

> *"Deep learning is more alchemy than science."*

The concern: Most theory is **postdictive** — explaining known successes after the fact.

Goal of This Talk

Develop **prescriptive theory** — theory that leads to *new* algorithms.

This Talk: From Explanation to Prescription

We explore a different path: Distribution-dependent generalization bounds

Part I — Postdictive Theory (Masegosa et al. JAIR/ECAI, 2025):

- PAC-Chernoff bounds are *perfectly tight* for interpolators
- They explain why interpolation can generalize well
- The **inverse rate function** emerges as the key complexity measure

Part II — Prescriptive Theory (Ju et al. Submitted to ICML 2026):

- The inverse rate function is an *optimal regularizer*
- We develop a **practical estimator** that can be used in training
- Empirical results confirm improved generalization

Goal: Move from postdictive explanations to prescriptive, actionable theory.

The Generalization Puzzle in Modern ML

Modern machine learning: Models are so large they interpolate training data (zero training error)

Classical Learning Theory: Interpolation = overfitting = poor generalization

Empirical reality: Many interpolators generalize remarkably well! (Zhang et al., 2017)

The Central Question

Why do some interpolators generalize well while others don't?

Inception	Crop	L2	Train Acc.	Test Acc.	Test NLL
Standard	no	no	99.99%	84.36%	0.65
Crop	yes	no	99.94%	86.89%	0.58
L2	no	yes	100.0%	86.60%	0.49
L2-Crop	yes	yes	99.98%	88.45%	0.42
Random	no	no	100.0%	10.13%	5.52
Initial	-	-	10.00%	10.00%	2.30

Inception models on Cifar10.

Traditional Generalization Bounds Fail

Standard PAC-style bounds:

$$L(\theta) \leq \hat{L}(\theta, D) + C(\theta, D, \delta)$$

where $L(\theta)$ = expected loss, $\hat{L}(\theta, D)$ = empirical loss, C = complexity term

Problem: Recent works show these bounds are provably vacuous for over-parameterized interpolators:

- Zhang et al. (2017): Empirical evidence with deep networks
- Nagarajan & Kolter (2019): Theoretical impossibility results
- Gastpar et al. (2024): Bounds without distributional assumptions cannot be tight
- Wang et al. (2024): Parameter-norm bounds necessarily loose

Emerging Conclusion

Bounds that **solely depend on training data** are provably vacuous for over-parameterized model classes.

Our Approach: Distribution-Dependent Bounds

Key insight: We need bounds that depend on the data-generating distribution ν

$$L(\theta) \leq \hat{L}(\theta, D) + C(\theta, n, \nu, \delta)$$

Why distribution-dependent?

- Training data alone cannot distinguish good from bad interpolators
- The data-generating distribution captures what matters for generalization
- Enables explaining techniques like data augmentation and invariant architectures

Roadmap

We now introduce the **rate function** from large deviation theory:

1. Quantifies how concentrated empirical loss is around expected loss
2. Defines a natural, distribution-dependent notion of **smoothness**
3. Yields **perfectly tight** bounds for interpolators

The Rate Function: Definition

Setup: Model θ, loss $\ell(y, x, \theta)$, data distribution $\nu(y, x)$

Definition (Rate Function)

$$\mathcal{I}_\theta(a) = \sup_{\lambda > 0} \{\lambda a - J_\theta(\lambda)\} \quad \forall a \in [0, L(\theta) - m_\theta)$$

where $J_\theta(\lambda) = \ln \mathbb{E}_\nu \left[e^{\lambda(L(\theta) - \ell(y,x,\theta))} \right]$ is the cumulant-generating function.

Inverse Rate Function:

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

Key property: $\mathcal{I}_\theta(\cdot)$ is convex, increasing, and characterizes tail probabilities

The Rate Function: Why It Matters

Chernoff's Bound: For any model θ and $a > 0$:

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

Cramér's Theorem: This bound is exponentially tight as $n \to \infty$:

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

Key Insight

- **Higher rate function** $\Rightarrow$ Empirical loss more concentrated around expected loss
- Chernoff bound is tighter for **larger** n and for **interpolators** ($a \approx L(\theta)$)
- These are exactly the settings of modern ML!

Theorem 1: PAC-Chernoff Bound

Theorem 1 (PAC-Chernoff Bound)

With high probability $1 - \delta$ *over* $D \sim \nu^n$*, for all* $\theta \in \Theta$ *simultaneously:*

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

Key features:

- k^p = size of model class (p parameters, k precision levels)
- Complexity term depends on rate function (distribution-dependent!)
- Monotonically increases with model class size, decreases with n

Perfect Tightness for Interpolators

Proposition 2 (Tightness)

With h.p. $1-\delta$ over $D \sim \nu^n$, for all $\theta \in \Theta$ simultaneously:

$$\text{If } \hat{L}(D,\theta) \leq \epsilon \text{ then} \quad 0 \leq L(\theta) - \mathcal{I}_\theta^{-1}\left(\frac{1}{n} \ln \frac{k^p}{\delta}\right) \leq \epsilon$$

What this means:

- For interpolators ($\hat{L} \approx 0$), the bound is perfectly tight!
- The inverse rate function $\mathcal{I}_\theta^{-1}$ **equals** the expected loss (up to ϵ)
- Works even for over-parameterized model classes

Answer to Open Question 1

Yes! There exist tight distribution-dependent bounds for over-parameterized interpolators.

Definition of Smoothness

Definition 3 (β-Smoother)

Model θ is β-**smoother** than model θ' if:

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

Intuition: Smoother models have:

- Higher rate function $\Rightarrow$ empirical loss more concentrated
- Less likely to see large deviations between train and test

Theorem: Smoother Interpolators Generalize Better

Theorem 4 (Main Result)

For any $\epsilon \geq 0$, with h.p. $1 - \delta$ over $D \sim \nu^n$, for all $\theta \in \Theta$, $\theta' \in \Theta'$ simultaneously:

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

$$\Downarrow$$

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

Key Message

Smoother interpolators generalize better (with high probability)!

- Works for any loss function (including log-loss)
- Works for over-parameterized model classes
- Provides a clear criterion for comparing interpolators

Smoother Interpolators Generalize Better

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

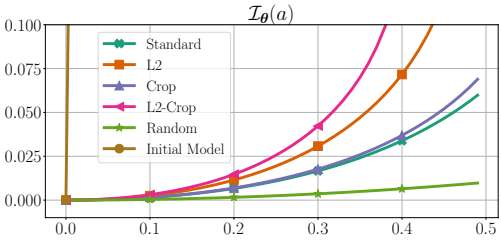


Figure: Metrics of Inception models on Cifar10 using ℓ_2 regularization and/or random cropping (Crop), and randomly sampled class labels (Random). The corresponding *rate functions* are shown on the right.

Regularization

The Inverse-Rate can be interpreted as a Regularizer (complexity control):

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

Regularization

The Inverse-Rate can be interpreted as a Regularizer (complexity control):

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

The best possible interpolator is $\boldsymbol{\theta_\epsilon^\star}$:

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

Regularization

The Inverse-Rate can be interpreted as a Regularizer (complexity control):

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

The best possible interpolator is $\boldsymbol{\theta_\epsilon^\star}$:

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

Theorem 5

With h.p. $1-\delta$*:* $|L(\theta_\epsilon^\star) - L(\theta_\epsilon^\times)| \leq \epsilon$

Regularization

The Inverse-Rate can be interpreted as a Regularizer (complexity control):

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

The best possible interpolator is $\boldsymbol{\theta_\epsilon^\star}$:

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

Theorem 5

With h.p. $1-\delta$*:* $|L(\theta_\epsilon^\star) - L(\theta_\epsilon^\times)| \leq \epsilon$

Optimal Regularzation

The inverse-rate $\mathcal{I}_\theta^{-1}(s)$ *characterizes an optimal regularizer for interpolators*

Connections to existing regularizers:

Technique	Connection to inverse rate
ℓ_2-norm	$\mathcal{I}_\theta^{-1}(s) \leq \sqrt{2M} \cdot \|\theta - \theta_0\|_2$ (if Lipschitz)
Distance from init	Same bound with $\theta_0 =$ initial weights
Input-gradient	$\mathcal{I}_\theta^{-1}(s) \leq \sqrt{s \cdot M \cdot \mathbb{E}[\|\nabla_x \ell\|^2]}$
Lipschitz constant	Upper bounds inverse rate

Unified explanation for many existing regularzers

All these regularizers are proxies for the inverse rate function!

Part II: From Explanation to Prescription

So far: PAC-Chernoff bounds *explain* why smooth interpolators generalize.

Key theoretical insight from Part I:

- The inverse rate function $\mathcal{I}_\theta^{-1}(s)$ is an optimal regularizer for interpolators.
- Existing regularizers (ℓ_2-norm, Lipschitz, etc.) are *proxies* for it

Natural question: Can we use the inverse rate function *directly* as a regularizer?

Part II (Ju et al., submitted ICML 2026):

- Develop a **practical regularizer** from the inverse rate function

The Inverse-Rate Regularizer: Practical Estimation

Challenge: The inverse rate depends on the *true* distribution ν — unknown!

Our Solution: Estimate from training data using **overlapping batch splitting**:

$$\widehat{J}_\theta(\lambda) = \log\left(\frac{1}{|S_1|}\sum_{(x,y)\in S_1} e^{-\lambda\ell(y,x,\theta)}\right) + \frac{\lambda}{|S_2|}\sum_{(x,y)\in S_2}\ell(y,x,\theta)$$

where S_1, S_2 are overlapping subsets controlled by parameter $\rho \in [0,1]$.

The Inverse-Rate Regularizer

$$\widehat{\mathcal{I}}_\theta^{-1}(s) = \inf_{\lambda>0}\frac{s+\widehat{J}_\theta(\lambda)}{\lambda}$$

Implicit Re-weighting Mechanism

Theorem: For log-loss $\ell(y, x, \theta) = -\log p(y|x, \theta)$:

$$\boxed{\widehat{\mathcal{I}}_\theta^{-1}(s) = \widehat{\mathbb{E}}\left[w_{y,x} \cdot \ell(y, x, \theta)\right]}$$

where the weights are:

$$w_{y,x} = 1 - \frac{p(y|x, \theta)^{\lambda^\star}}{\widehat{\mathbb{E}}\left[p(y|x, \theta)^{\lambda^\star}\right]}$$

Interpretation:

- $w_{y,x} < 0$ (over-confident): Increase loss — penalize overconfidence
- $w_{y,x} > 0$ (under-confident): Decrease loss — focus on hard examples

$\Rightarrow$ **Connects to focal loss and sample-adaptive methods!**

Connection to Distributionally Robust Optimization

KL-DRO Objective: Minimize worst-case loss over distributions near empirical:

$$\hat{L}^{s}_{\mathrm{DRO}}(\theta) = \sup_{q \in \Delta_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} q_i \, \ell(y_i, x_i, \theta) \;\middle|\; D_{\mathrm{KL}}(q \| \widehat{P}_n) \leq s \right\}$$

Theorem: Under symmetry of the empirical loss distribution:

$$\boxed{\hat{L}^{s}_{\mathrm{DRO}}(\theta) = \hat{L}(D, \theta) + \widehat{\mathcal{I}}_{\theta}^{-1}(s)}$$

$\Rightarrow$ **Inverse-rate regularization $\equiv$ KL-constrained DRO!**

This provides a **robustness interpretation**: we're optimizing for worst-case perturbations of the training distribution.

Experimental Setup

Datasets: CIFAR-10 and CIFAR-100

Architectures: 19 CNN models spanning diverse design paradigms

- ResNet (20, 32, 44, 56)
- VGG (11, 13, 16, 19 + BN)
- MobileNetV2 (0.5x–1.4x)
- ShuffleNetV2 (0.5x–2.0x)
- RepVGG (A0, A1, A2)

Why this selection?

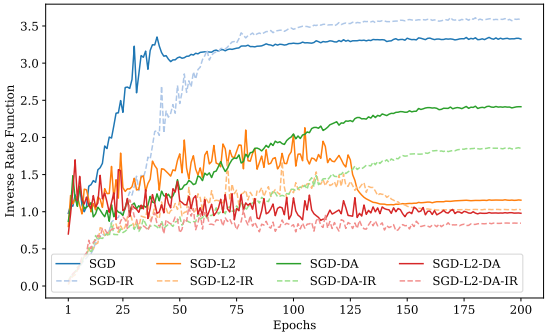
- **Varying depths and widths** — tests whether inverse-rate benefits scale
- **Different inductive biases** — residual, plain, efficient architectures
- **All reach interpolation** — the regime where our theory applies

Training: SGD with momentum 0.9, 200 epochs, standard augmentation

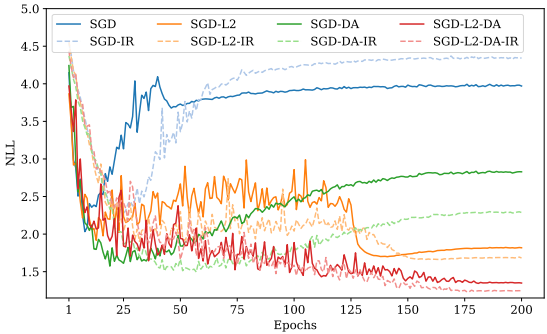
Hyperparameters: $\rho \in \{0.0, 0.5, 1.0\}$, $\lambda^\star \in \{0.001, 0.01, 0.1, 0.5, 1.0, 2.0\}$

Results: Ablation Study

Inverse Rate Function



Test NLL



Key observations:

- Standard regularizers (L2, data augmentation) lower the inverse rate curve
- Adding inverse-rate regularizer (IR) yields **lowest inverse-rate values**
- This translates to **better test NLL** (generalization)

Results: Performance Comparison

CIFAR-10	Top1-Acc	Top5-Acc	NLL	Variance	ECE
Wins (Ours/19)	7/19	**11/19**	**18/19**	**19/19**	17/19
CIFAR-100	Top1-Acc	Top5-Acc	NLL	Variance	ECE
Wins (Ours/19)	3/19	**14/19**	**19/19**	**16/19**	**19/19**

Key findings:

- **NLL:** 37/38 model-dataset pairs improved (better probabilistic quality)
- **Variance:** 35/38 improved (more uniform predictions)
- **Top-5 Accuracy:** 25/38 improved (better calibrated rankings)
- **ECE:** 35/38 improved (better calibration)

$\Rightarrow$ **Models are more trustworthy** — crucial for safety-critical applications!

Summary and Contributions PART I (Masegosa et al, JAIR/ECAI, 2025)

Main contributions:

1. **PAC-Chernoff Bound:** Perfectly tight for interpolators, even over-parameterized
2. **Smoothness via Rate Function:** Natural, distribution-dependent complexity measure
3. **Unified Framework:** Explains why modern techniques find good interpolators

Explained phenomena:

- Double-descent: Larger models become smoother
- Regularization: ℓ_2-norm, distance from init, Lipschitz are proxies for smoothness
- Invariances & DA: Increase concentration of empirical loss
- Over-parameterization: Necessary condition for smooth interpolation

Take-home Message:

Distribution-dependent bounds *(via rate function) are a powerful tool for understanding generalization in the interpolation regime.*

Summary and Contributions PART II (Ju et al, Submitted ICLM 2026)

Main Contributions:

1. **Practical estimator** of the theoretically optimal inverse rate function
2. **Seamless integration** into first-order optimization (SGD, Adam, etc.)
3. **Theoretical connections** to:
   - Sample-adaptive losses (focal loss)
   - Distributionally robust optimization (KL-DRO)
   - Existing regularizers (weight decay, Lipschitz)
4. **Consistent empirical improvements** in probabilistic quality

Take-home Message:

Improving estimators of the inverse rate function provides a well-defined path toward near-optimal regularization in deep learning.