Bound to Disagree: Generalization Bounds via Certifiable Surrogates

Mathieu Bazinet[1] Valentina Zantedeschi[1,2] Pascal Germain[1]

[1]Université Laval [2]ServiceNow Research

UAI 2026

Setup

Basic notation:

- A dataset $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is i.i.d.-sampled from a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$
- A predictor/model $h : \mathcal{X} \to \mathbb{R}^C$ is learned from S. (today, $\mathcal{Y} = \{1, \ldots, C\}$)

We are interested in certifying the predictor h without relying on a test set

Generalization Bounds

Given a loss function $\ell : \mathbb{R}^C \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ and a confidence $\delta \in [0, 1]$:

$$\mathop{\mathbb{P}}_{S \sim \mathcal{D}^n} \Big(\underbrace{\mathcal{L}_{\mathcal{D}}(h)}_{\text{true loss}} \leq \underbrace{\widehat{\mathcal{L}}_S(h)}_{\text{empirical loss}} + \underbrace{\epsilon(n, \delta, \ldots)}_{\text{complexity term}} \Big) \geq 1 - \delta.$$

As we want to use the bound as a numerical certificate the complexity term $\epsilon(n, \delta, \ldots)$ must be small and computable.

Actionable Generalization Bounds

Three classical frameworks for tight bounds:

- **Model compression** (e.g., Zhou et al., 2019; Lotfi et al., 2024)
  — complexity $\epsilon(n, \delta, l_{\mathcal{C}}) \approx$ description length $l_{\mathcal{C}}$ of a compressed model.
- **Sample compression** (e.g., Littlestone and Warmuth, 1986; Paccagnan et al., 2024)
  — complexity $\epsilon(n, \delta, S_{\mathbf{i}}) \approx$ size of a compression subset $S_{\mathbf{i}} \subset S$.
- **PAC-Bayes** (e.g., McAllester, 1998; Dziugaite and Roy, 2018; Pérez-Ortiz et al., 2021)
  — complexity $\epsilon(n, \delta, Q, P) \approx$ KL divergence between a prior P and posterior Q over models.

Common limitations:

- They rarely give tight bounds for deep/large models learned with everyday scheme.
- They require special training recipes to control the model's complexity $\epsilon(n, \delta, \ldots)$

But what if one wants to certify an already trained large model?

Main Result

Key ideas:

- Train a **certifiable surrogate** h close to the **target model** f.
- **Bound their disagreement** on a unlabeled set $U \sim \mathcal{D}^m$

Bound sketch

*Given **two predictors** f & h, a loss ℓ, a confidence δ, and a proper **disagreement bound** D_U :*

$$\mathop{\mathbb{P}}_{U\sim\mathcal{D}^m}\Big(\underbrace{\mathcal{L}_{\mathcal{D}}(f)}_{\text{Target's loss}} \leq \underbrace{\mathcal{L}_{\mathcal{D}}(h)}_{\text{Surrogate's loss}} + \underbrace{D_U(f,h;\ell,\delta)}_{\text{disagreement term}} \Big) \geq 1-\delta.$$

Proposed Disagreement Bounds

Zero-one loss – classification $\hat{y} = \operatorname{argmax}_j f(\boldsymbol{x})_j$ (inspired by Langford, 2005)

$$D_U(f, h; \ell^{0\text{-}1}, \delta) = \overline{\mathrm{Bin}}\left(\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left[\operatorname*{argmax}_j f(\boldsymbol{x}_i)_j \neq \operatorname*{argmax}_k h(\boldsymbol{x}_i)_k\right], m, \delta\right)$$

where $\overline{\mathrm{Bin}}$ is the inverse of the binomial CDF.

L-Lipschitz loss – soft-max output $\hat{\boldsymbol{y}} = \boldsymbol{\sigma}(f(\boldsymbol{x}))$ (inspired by Foong et al., 2022)

$$D_U(f, h; \ell^{(L)}, \delta) = 2L\,\mathrm{kl}^{-1}\left(\frac{1}{2m}\sum_{i=1}^{m}\left\|\boldsymbol{\sigma}(f(\boldsymbol{x}_i)) - \boldsymbol{\sigma}(h(\boldsymbol{x}_i))\right\|_1,\ \tfrac{1}{m}\log\tfrac{1}{\delta}\right)$$

where $\mathrm{kl}^{-1}(q, \epsilon)$ is the inverse of the binary KL divergence.

Also in the paper:

- Extension to a disagreement between a target f and a distribution Q over surrogates.
- Weaker result for non-Lipschitz losses (required labeled U)

The Bounds in Action

Three use cases:

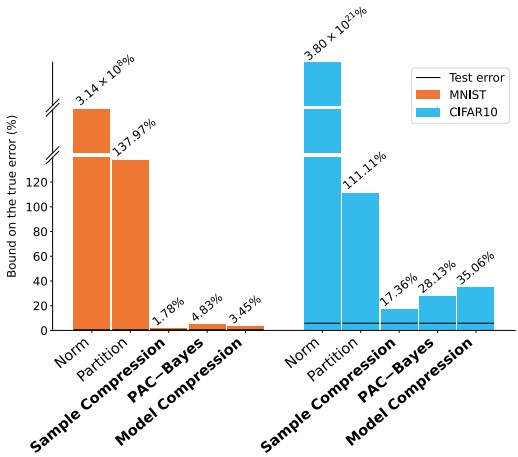
- **Certify** a target model using an independently-learned surrogate.
- **Minimize** the surrogate disagreement via distillation.
- **Replace** a legacy model by a preferred one while bounding the performance gap.

Use Case 1: Certifying the Target Model

Goal: Train a certifiable surrogate $h^\star$ and bound the target f:

$$\mathcal{L}_{\mathcal{D}}(f) \leq \underbrace{\widehat{\mathcal{L}}_S(h^\star) + \epsilon\left(n, \frac{\delta}{2}\right)}_{\text{surrogate's bound}} + \underbrace{D_U\left(f, h^\star; \ell, \frac{\delta}{2}\right)}_{\text{disagreement bound}}$$

- Bounds on the zero-one loss
- Targets: CNN on MNIST,
  ResNet18 on CIFAR-10
- Surrogates trained to optimize their bounds
- Compared against bounds directly on the target model:
  - **norm-based** (Galanti et al., 2023)
  - **partition-based** (Than and Phan, 2025)



Use Case 2: Minimizing the Disagreement (Distillation)

Goal: Given a distribution Q over the space of surrogates $\mathcal{H}$, learn a surrogate $h \in \mathcal{H}$ that minimizes disagreement with the target f:

$$\forall\, h \in \mathcal{H}: \quad \mathcal{L}_{\mathcal{D}}(f) \leq \underbrace{\widehat{\mathcal{L}}_S(h) + \epsilon\Big(n,\ \tfrac{Q(h)\delta}{2}\Big)}_{\text{surrogate's bound}} + \underbrace{D_U\Big(f, h; \ell,\ \tfrac{Q(h)\delta}{2}\Big)}_{\text{disagreement bound}}$$

Model	Surrogate h			Target f	
	Test error (%)	Size (KB)	Bound (%)	Test error (%)	Our bound (%)
DistilBERT	7.61±0.07	2.02±0.01	14.37±0.08	6.19±0.15	21.35±0.32
GPT2	16.37±0.94	2.04±0.43	25.05±0.91	5.51±0.11	43.34±1.52

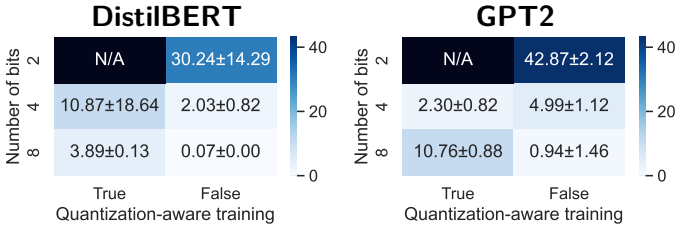
- Bounds on the Huber loss
- Targets: DistilBERT & GPT2 fine-tuned on Amazon Polarity dataset
- Surrogates: Compressed SubLoRA adapters (Lotfi et al., 2024) trained by distillation

Use Case 3: Replacing a Model

Goal: Bound the performance gap when replacing a legacy model f_1 by a preferred one f_2:

$$\underbrace{\left|\mathcal{L}_{\mathcal{D}}(f_1) - \mathcal{L}_{\mathcal{D}}(f_2)\right|}_{\text{performance gap}} \leq \underbrace{D_U(f_1, f_2; \ell, \delta)}_{\text{disagreement bound}}.$$

- Bounds on the Huber loss
- Legacy models:
  DistilBERT & GPT2 fine-tuned on Amazon Polarity dataset
- Candidate models:
  Quantized variants (Badri and Shaji, 2023; Or et al., 2025)



Conclusion

Summary & Contributions

- New **disagreement-based generalization bounds** certifying complex models via simpler certifiable surrogates.
- **Three use cases:**
  - **Certify** a fixed target model using a surrogate.
  - **Minimize** the disagreement via distillation.
  - **Replace** a model while bounding the performance gap.
- **Non-vacuous** bounds demonstrated on large NLP models (DistilBERT, GPT2).

Thank you!

Contact:
mathieu.bazinet.2@ulaval.ca

References I

Hicham Badri and Appu Shaji. Half-quadratic quantization of large machine learning models, November 2023. URL `https://dropbox.github.io/hqq_blog/`.

Gintare Karolina Dziugaite and Daniel M. Roy. Data-dependent PAC-Bayes priors via differential privacy. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, pages 8440–8450, 2018.

Andrew YK Foong, Wessel P Bruinsma, and David R Burt. A note on the Chernoff bound for random variables in the unit interval. *ArXiv preprint*, abs/2205.07880, 2022.

Tomer Galanti, Mengjia Xu, Liane Galanti, and Tomaso Poggio. Norm-based generalization bounds for sparse neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

John Langford. Tutorial on practical prediction theory for classification. *Journal of machine learning research*, 6 (3), 2005.

Nick Littlestone and Manfred Warmuth. Relating data compression and learnability. 1986.

Sanae Lotfi, Marc Anton Finzi, Yilun Kuang, Tim G. J. Rudner, Micah Goldblum, and Andrew Gordon Wilson. Non-vacuous generalization bounds for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32801–32818. PMLR, 2024.

References II

David A McAllester. Some PAC-Bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234, 1998.

Andrew Or, Apurva Jain, Daniel Vega-Myhre, Jesse Cai, Charles David Hernandez, Zhenrui Zhang, Driss Guessous, Vasiliy Kuznetsov, Christian Puhrsch, Mark Saroufim, et al. Torchao: Pytorch-native training-to-serving model optimization. In *Championing Open-source DEvelopment in ML Workshop@ ICML25*, 2025.

Dario Paccagnan, Marco Campi, and Simone Garatti. The pick-to-learn algorithm: Empowering compression for tight generalization bounds and improved post-training performance. *Advances in Neural Information Processing Systems*, 36, 2024.

María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22(227):1–40, 2021.

Khoat Than and Dat Phan. Non-vacuous generalization bounds for deep neural networks without any modification to the trained models. *arXiv preprint arXiv:2503.07325*, 2025.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous generalization bounds at the imagenet scale: a PAC-Bayesian compression approach. In *ICML*, 2019.