Sample Compression Unleashed : New Generalization Bounds for Real Valued Losses

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

[1]Université Laval [2]ServiceNow Research

Fonds de recherche Québec

Summary

We present new sample-compression guarantees for any bounded losses and validate empirically their tightness when applied to deep neural networks.

Background and notation

1. A dataset $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ sampled *i.i.d.* from an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$.
2. A family of predictor $\mathcal{H}$ of the form $h : \mathcal{X} \to \mathcal{Y}$.
3. A learning algorithm A that returns a predictor $A(S) \in \mathcal{H}$.
4. A loss function $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.
5. The generalization loss of a predictor $\mathcal{L}_{\mathcal{D}}(h) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\, \ell(h, \boldsymbol{x}, y)$.
6. The empirical loss of a predictor $\widehat{\mathcal{L}}_S(h) = \frac{1}{n}\sum_{i=1}^n \ell(h, \boldsymbol{x}_i, y_i)$.

Sample compression theory

A predictor is called a sample-compressed predictor if it can be expressed as a function of a subset of S. To do so, we need :

1. A compression set $S_{\mathbf{i}} \subseteq S$, defined by a vector of indices $\mathbf{i} = (i_1, \ldots, i_{|\mathbf{i}|})$ such that $1 \leq i_1 \leq \ldots i_{|\mathbf{i}|} \leq n$. Each $\mathbf{i}$ are contained in the powerset $\mathcal{P}(n)$ of the numbers 1 to n. We denote its complement $S_{\mathbf{i}^c} = S \setminus S_{\mathbf{i}}$.
2. A reconstruction function $\mathcal{R}$ that takes a compression set and a message to output a predictor.

We denote a sample-compressed predictor $\mathcal{R}(S_{\mathbf{i}})$.

Example of sample compression : the SVM

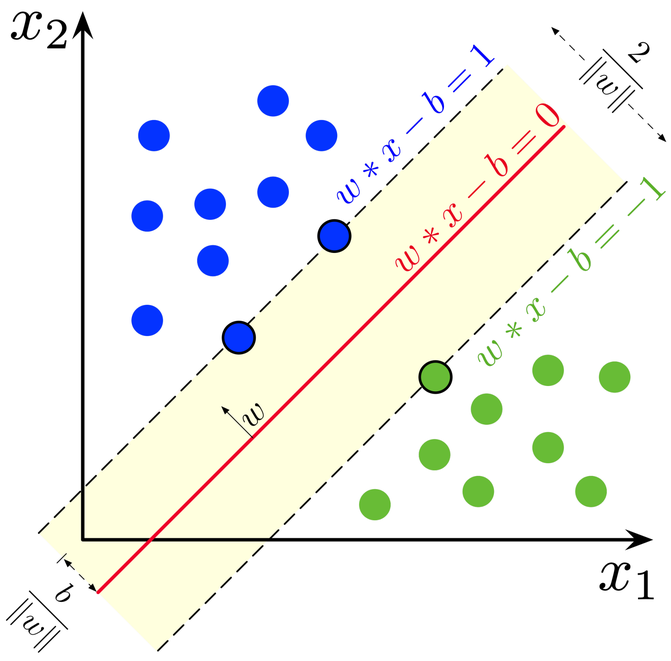


Figure 1. By Larhmam - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=73710028.

Main results

We present **new sample-compression generalization results** that

- hold for **any bounded losses** and some unbounded losses,
- **do not depend on the number of parameters** of the neural network,
- are **tight and easily computable**.

These new results can be applied

1. to **classification** problems, **regression** problems and more,
2. on a variety of models, such as **MLPs and transformers** (with the help of the meta-algorithm Pick-To-Learn [4]),

and still give **tight and non-vacuous guarantees** in practice.

Theorem 1 : New General Sample Compression bound

For any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, with $\mathcal{P}(n)$ the powersets of 1 to n, for any distribution $P_{\mathcal{P}(n)}$ over $\mathcal{P}(n)$, for any comparator function $\Delta : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and for any $\delta \in (0, 1]$, with probability at least $1 - \delta$ over the draw of $S \sim \mathcal{D}^n$, we have

$$\forall \mathbf{i} \in \mathcal{P}(n) : \Delta\Big(\widehat{\mathcal{L}}_{S_{\mathbf{i}^c}}(\mathcal{R}(\mathbf{i})), \mathcal{L}_{\mathcal{D}}(\mathcal{R}(\mathbf{i}))\Big) \leq \frac{1}{n - |\mathbf{i}|}\left[\log \boldsymbol{\mathcal{E}_\Delta(n, \mathbf{i})} + \log\left(\frac{1}{\boldsymbol{P_{\mathcal{P}(n)}(\mathbf{i})}\delta}\right)\right]$$

with

$$\mathcal{E}_\Delta(n, \mathbf{i}) = \mathop{\mathbb{E}}_{T_{\mathbf{i}}\sim\mathcal{D}^{|\mathbf{i}|}} \mathop{\mathbb{E}}_{T_{\mathbf{i}^c}\sim\mathcal{D}^{n-|\mathbf{i}|}} e^{(n-|\mathbf{i}|)\Delta\left(\widehat{\mathcal{L}}_{T_{\mathbf{i}^c}}(\mathcal{R}(T_{\mathbf{i}})), \mathcal{L}_{\mathcal{D}}(\mathcal{R}(T_{\mathbf{i}}))\right)}.$$

Corollary 1 : Optimal comparator function for bounded losses

In the setting of Theorem 1, for any loss function $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to [0, 1]$, with probability at least $1 - \delta$ over the draw of $S \sim \mathcal{D}^n$, we have

$$\forall \mathbf{i} \in \mathcal{P}(n) : \mathcal{L}_{\mathcal{D}}(\mathcal{R}(S_{\mathbf{i}})) \leq \sup_{0\leq p\leq 1}\left\{\mathrm{kl}\Big(\widehat{\mathcal{L}}_{S_{\mathbf{i}^c}}(\mathcal{R}(S_{\mathbf{i}})), p\Big) \leq \frac{1}{n - |\mathbf{i}|}\log\left(\frac{\boldsymbol{2\sqrt{n - |\mathbf{i}|}}}{\boldsymbol{P_{\mathcal{P}(n)}(\mathbf{i})}\delta}\right)\right\}.$$

The binary Kullback-Leibler divergence kl is **known to be optimal** for losses in the range $[0, 1]$ as per the results of [3].

Corollary 2 : Unbounded losses with the linear distance

In the setting of Theorem 1, for any $\lambda > 0$, with $\Delta_\lambda(q, p) = \lambda(p - q)$, with a σ^2-sub-Gaussian loss function $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, with probability at least $1 - \delta$ over the draw of $S \sim \mathcal{D}^n$, we have

$$\forall \mathbf{i} \in \mathcal{P}(n) : \mathcal{L}_{\mathcal{D}}(\mathcal{R}(S_{\mathbf{i}})) \leq \widehat{\mathcal{L}}_{S_{\mathbf{i}^c}}(\mathcal{R}(S_{\mathbf{i}})) + \frac{\boldsymbol{\lambda\sigma^2}}{\boldsymbol{2}} + \frac{1}{\lambda(n - |\mathbf{i}|)}\log\left(\frac{1}{\boldsymbol{P_{\mathcal{P}(n)}(\mathbf{i})}\delta}\right).$$

Experiments

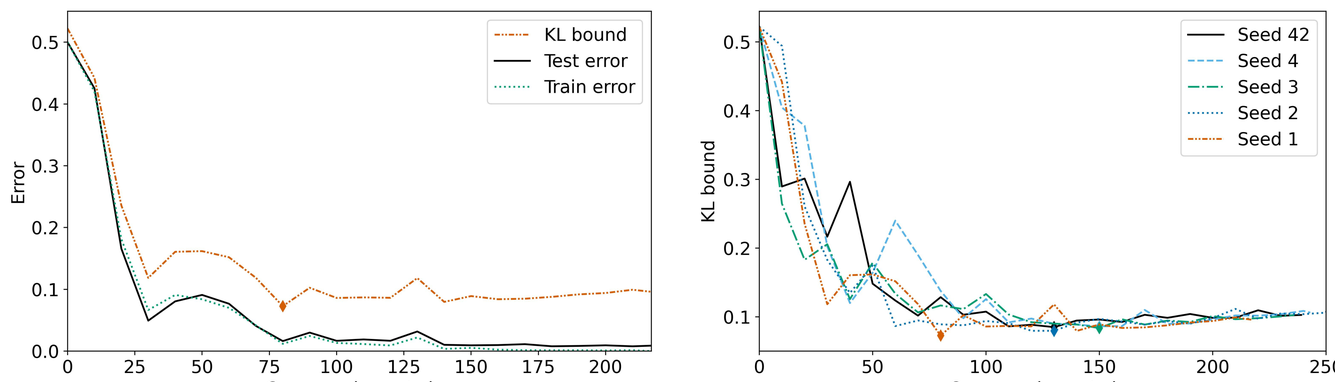
Amazon Polarity with DistilBERT

1. Dataset : **Binary classification problems** on Amazon Polarity dataset (we use 10%, i.e. 360000 datapoints).
2. Model type : DistilBERT [6] with **66 million parameters**.
3. Training algorithm : **Pre-training** on 50% of the dataset, then **Pick-To-Learn** on the other half.

Train method	Train error	Validation error	Test error	kl bound
Pick-To-Learn	4.73±1.09	5.41±1.05	5.60±1.19	13.91±2.73
Baseline	3.11±0.02	4.08±0.04	4.19±0.00	-

Table 1. All metrics present are in percent (%).

Binary MNIST with a CNN



(a) Behavior of the metrics for seed 1 (b) Behavior of the kl bound for all seeds

Figure 2. Illustration of Corollary 1's bound behavior throughout Pick-To-Learn iterations for the model that achieved the minimal Pick-To-Learn bound on the binarized MNIST dataset (4 vs 9). We mark the minimal kl bound for each seed with a diamond (⧫).

Further extensions and conclusion

In the future, we will extend our work to any:

1. **Unbounded losses** under the hypothesis-dependent range condition [2],
2. **Unbounded losses** under model-dependent assumptions [1],
3. Distributions with **more general tail behaviors** [5].

In conclusion, we presented a new general sample-compression theorem for real-valued losses and empirically verified its tightness for deep neural networks. In future experiments, we wish to tackle regression problems with the help of Corollary 2.

[1] Ioar Casado, Luis A. Ortega, Andrés R. Masegosa, and Aritz Pérez. Pac-bayes-chernoff bounds for unbounded losses, 2024.

[2] Maxime Haddouche, Benjamin Guedj, Omar Rivasplata, and John Shawe-Taylor. Pac-bayes unleashed: Generalisation bounds with unbounded losses. *Entropy*, 23(10):1330, 2021.

[3] Fredrik Hellström and Benjamin Guedj. Comparing comparators in generalization bounds. In *International Conference on Artificial Intelligence and Statistics*, pages 73–81. PMLR, 2024.

[4] 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.

[5] Borja Rodríguez-Gálvez, Ragnar Thobaben, and Mikael Skoglund. More pac-bayes bounds: From bounded losses, to losses with general tail behaviors, to anytime validity. *Journal of Machine Learning Research*, 25(110):1–43, 2024.

[6] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

Find the paper here!

Email me at : mathieu.bazinet.2@ulaval.ca