Sample compression unleashed : New generalization bounds for real valued losses

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

[1]Université Laval [2]ServiceNow Research

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 $1 \le i_1 \le \ldots i_{|\mathbf{i}|} \le n$. We denote its complement $S_{\mathbf{i}^c} = S \setminus S_{\mathbf{i}}$.
2. **A reconstruction function** $\mathscr{R}$ that takes a compression set and a message to output a predictor.

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

Example of sample compression : the SVM

The **support vectors** of the SVM form its **compression set**.

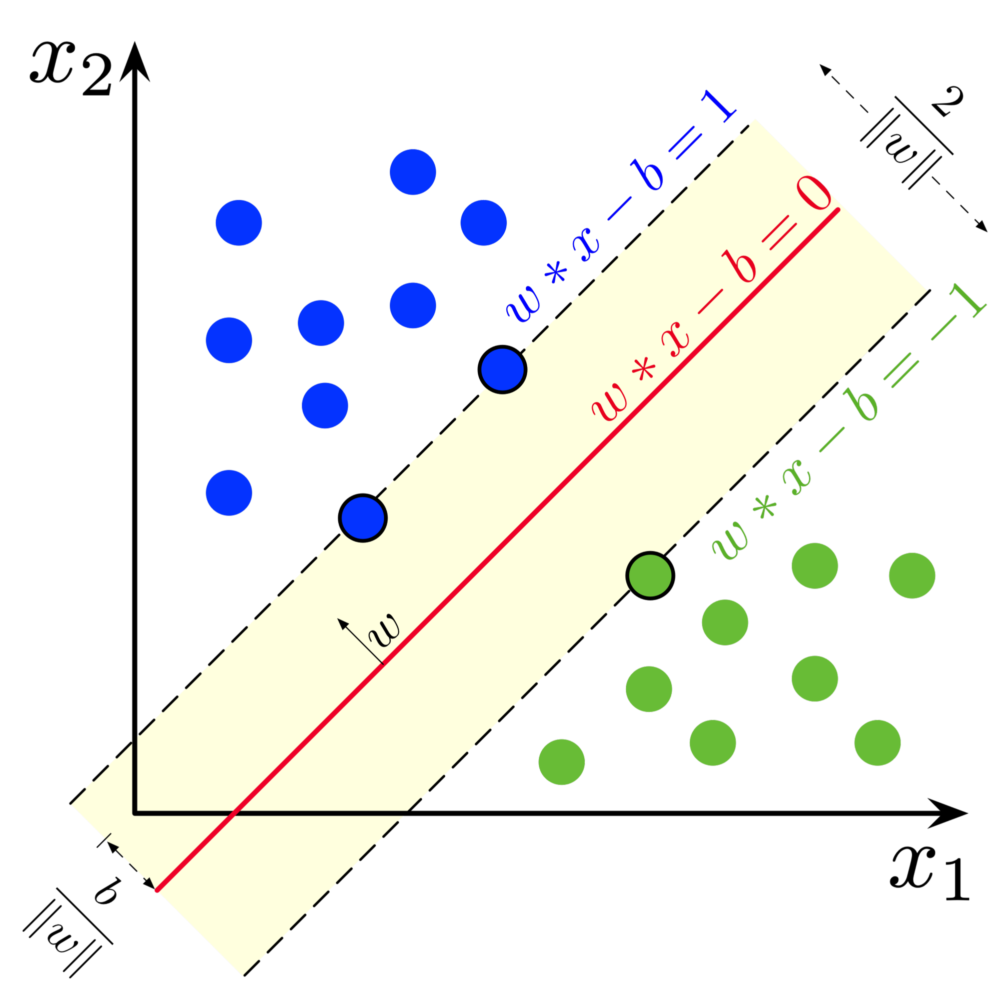

Figure 1. By Larhmam - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=73710028.

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 $\mathcal{X} \times \mathcal{Y}$, for any set of vectors of indices I, for any distribution P_I over I, 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

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

with

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

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

Corollary 1 : 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

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

Further extensions

In the future, we will extend our work to :

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

Experiments on Binary MNIST Problems

- Dataset : **Binary classification problems** created from MNIST (≈11000 datapoints)
- Model type : CNN with **1.1 million parameters**
- Training algorithm : **Pick-To-Learn**

Dataset	Validation error	Test error	kl bound	$\lvert\mathbf{i}\rvert$	Baseline test error
MNIST08	**0.33±0.17**	**0.25±0.10**	**5.05±0.16**	**92.0±3.6**	**0.22±0.05**
MNIST17	0.20±0.08	0.38±0.16	4.33±0.21	84.0±5.2	0.17±0.03
MNIST23	0.39±0.12	0.27±0.10	8.20±0.34	175.6±9.5	0.16±0.05
MNIST49	0.82±0.11	0.77±0.17	10.52±0.37	237.0±11.0	0.44±0.07
MNIST56	0.46±0.12	0.47±0.15	6.29±0.22	117.0±5.2	0.30±0.08

Table 1. All metrics presented are in percents (%), with the exception of $|\mathbf{i}|$.

Preliminary results on Amazon polarity

1. Dataset : **Binary classification problems** on Amazon Polarity dataset (we use 10%, 360000 datapoints)
2. Data type : **textual reviews**
3. Model type : DistilBERT [6] with **66 million parameters**
4. 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
P2I	3.04±0.77	3.66±0.84	5.18±0.14	10.23±1.91
Baseline	2.34±0.79	3.24±0.82	4.25±0.06	-

Table 2. All metrics present are in percent (%). The results for the baseline were computed on 2 seeds instead of 5.

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 1 and multi-class classification problems.

References

[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.

mathieu.bazinet.2@ulaval.ca