A mutual information bound for statistical inference

Pierre Alquier
(joint work with El Mahdi Khribch)

The mutual information (MI) bound is a well known result in classification theory. Given a sample $\mathcal{S} = ((X_1, Y_1), \ldots, (X_n, Y_n))$ of i.i.d. pairs object-label, a set of classifiers $(f_w, w \in \Theta)$ and a data-dependent stochastic algorithm returning a (random) $\theta \in \Theta$, the MI bound states that

$$\left|\mathbb{E}\Big(R(\theta) - R_n(\theta)\Big)\right| \leq \sqrt{\frac{\mathcal{I}(\mathcal{S}, \theta)}{2n}}$$

where $R(\theta)$ is the out-of-sample missclassification probability of θ, $R_n(\theta)$ its in-sample missclassification frequency and $\mathcal{I}(\mathcal{S}, \theta)$ the mutual information between $\mathcal{S}$ and θ. While the result was first stated in this form by [6], even though similar results can be found as early as [8, 3]. This result reveals profound connections between information theory, PAC-Bayesian bounds, and Bayesian machine learning.

In this talk, I will introduce a variant of this result taylored for statistical inference rather than for classification. In the preprint [5], we provided a general inequality; I will only state a special case here.

Theorem. Assume we observe a sample $\mathcal{S} = (X_1, \ldots, X_n)$ of i.i.d. random variables from P^0 and consider a statistical model $(P_w, w \in \Theta)$. Let p^0 and p_w denote the respective densities of P_0 and P_w with respect to some common measure μ. Consider a stochastic algorithm producing a random parameter θ, distributed according to some $\mathcal{S}$-dependent probability distribution $\rho_{\mathcal{S}}$ on Θ. Then

$$\mathbb{E}\left(h^2(P_\theta, P^0) - \frac{1}{n}\sum_{i=1}^{n} \log\frac{p^0(X_i)}{p_\theta(X_i)}\right) \leq \frac{2\mathcal{I}(\mathcal{S}, \theta)}{n}$$

where h^2 is the squared-Hellinger metric and $\mathcal{I}(\mathcal{S}, \theta)$ the mutual information between $\mathcal{S}$ and θ.

As an example of application, I will study the fractional Bayesian posteriors of [2]:

$$\rho_{\mathcal{S}}(\mathrm{d}\theta) \propto \left(\prod_{i=1}^{n} p_\theta(X_i)\right)^{\frac{1}{2}} \pi(\mathrm{d}\theta)$$

for some prior π. Under suitable conditions (including that a certain measure d of the complexity of the model Θ is finite),

$$\mathbb{E}\Big(h^2(P_\theta, P^0)\Big) \leq \inf_{w\in\Theta} KL(P^0\|P_w) + \frac{d}{n}.$$

This improves on the $d\log(n)/n$ rates in similar contexts [1, 2].

The strength of the above theorem is its generality: it can be applied in principle to any stochastic algorithm $\rho_{\mathcal{S}}$. For example, I will discuss an application to variational approximations of the posterior, in the same context as [1, 7, 9]. The paper discusses further examples of applications: nonparametric models (under

prior mass conditions similar to the ones in [4]), and even convergence of the MLE.

References

[1] P. Alquier and J. Ridgway. *Concentration of tempered posteriors and of their variational approximations*. The Annals of Statistics **48**(3):1475–1497 (2020).

[2] A. Bhattacharya, D. Pati and Y. Yang. *Bayesian fractional posteriors*. The Annals of Statistics **47**(1):39–66 (2019).

[3] O. Catoni. PAC-Bayesian supervised classification: The thermodynamics of statistical learning, Institute of Mathematical Statistics Lecture Notes – Monograph Series 56 (2007).

[4] S. Ghosal, J. K. Ghosh and A. W. Van Der Vaart. *Convergence rates of posterior distributions*. The Annals of Statistics **28**(2):500–531 (2000).

[5] E. M. Kribch and P. Alquier. Convergence of Statistical Estimators via Mutual Information Bounds. Preprint arXiv:2412.18539 (2024).

[6] D. Russo and J. Zou. *Controlling bias in adaptive data analysis using information theory*, Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, PMLR **51**:1232–1240 (2016).

[7] Y. Yang, D. Pati. and A. Bhattacharya. α*-variational inference with statistical guarantees*. The Annals of Statistics **48**(2):886–905 (2020).

[8] T. Zhang, *Information-Theoretic Upper and Lower Bounds for Statistical Estimation*, IEEE Transactions on Information Theory **52**(4):1307–1321 (2006).

[9] Fengshuo Zhang and Chao Gao. Convergence rates of variational posterior distributions. *Ann. Statist.*, 48(4):2180–2207, 2020.