A new Mutual Information Bound for Statistical Inference

Pierre Alquier

OIST ML Workshop
March 4, 2025

Talk based on the preprint :

EL Mahdi Khribch, Pierre Alquier (2024).

Convergence of Statistical Estimators via Mutual Information Bounds.

Preprint arXiv :2412.18539.

EL Mahdi (Mehdi) Khribch

ESSEC Business School

(Paris campus)

Generalisation error in machine learning

- Risk :

$$R(\theta) := \mathbb{E}_{(X,Y)\sim P}\left[\ell\left(Y, f_\theta(X)\right)\right].$$

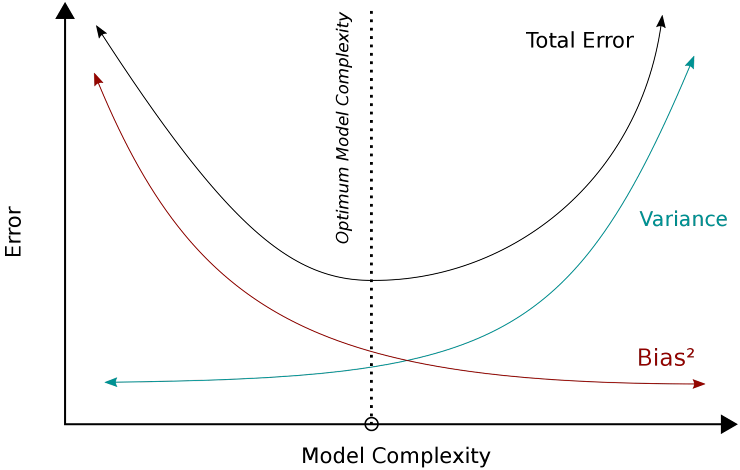
- Data $\mathcal{S} = ((X_1, Y_1), \dots, (X_n, Y_n))$ i.i.d. from P. Empirical risk :

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y_i, f_\theta(X_i)\right).$$

- Randomized estimator : $\hat{\theta}$, sampled from a data-dependent probability distribution $\hat{\rho} = \hat{\rho}(\mathcal{S})$.
- Generalization gap :

$$\text{gen}(\hat{\theta}, \mathcal{S}) = R(\hat{\theta}) - R_n(\hat{\theta}).$$

Classical visualization



Source : wikipedia ("bias-variance tradeoff" page).

Mutual information : definition

Küllback-Leibler divergence

$$\mathrm{KL}(\nu\|\mu) = \mathbb{E}_{\theta\sim\nu}\left[\log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(\theta)\right]$$

and $\mathrm{KL}(\nu\|\mu) = \infty$ is ν has no density $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ w.r.t. μ...

$$\mathrm{KL}(\nu\|\mu) \geq 0 \text{ and } \mathrm{KL}(\nu\|\mu) = 0 \Leftrightarrow \nu = \mu.$$

Let $(U, V) \sim Q$. Let Q_U and Q_V denote their marginals. If U and V were independent, $Q = Q_U \otimes Q_V$.

Mutual information between two random variables

$$\mathcal{I}(U, V) := \mathrm{KL}(Q\|Q_U \otimes Q_V).$$

Mutual information bound

Mutual information bound (Catoni, 2007 ; Russo & Zou, 2019)

Assumption : $0 \leq \ell(Y, f_\theta(X)) \leq 1$, then

$$\left| \mathbb{E}_{\mathcal{S}} \mathbb{E}_{\hat{\theta}} \operatorname{gen}(\hat{\theta}, \mathcal{S}) \right| \leq \sqrt{\frac{\mathcal{I}(\hat{\theta}, \mathcal{S})}{2n}}.$$

Warning

Notation

In this talk, MIB will not be used in its usual meaning. It will stand for "Mutual Information Bound".

Toy example

Finite set of predictors $\{\theta_1, \ldots, \theta_M\}$, then $\mathcal{I}(\hat{\theta}, \mathcal{S}) \leq \log(M)$.

The MIB gives :

$$\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\hat{\theta}} R(\hat{\theta}) \leq \mathbb{E}_{\mathcal{S}} \mathbb{E}_{\hat{\theta}} R_n(\hat{\theta}) + \sqrt{\frac{\log(M)}{2n}}.$$

If we take $\hat{\rho}$ as a point mass on the empirical risk minimizer (ERM) : $\hat{\theta} = \hat{\theta}_{\mathrm{ERM}}$. Then

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}} R(\hat{\theta}_{\mathrm{ERM}}) &\leq \mathbb{E}_{\mathcal{S}} \min_{1 \leq j \leq M} R_n(\theta_j) + \sqrt{\frac{\log(M)}{2n}} \\
&\leq \min_{1 \leq j \leq M} \mathbb{E}_{\mathcal{S}} R_n(\theta_j) + \sqrt{\frac{\log(M)}{2n}} \\
&= \min_{1 \leq j \leq M} R(\theta_j) + \sqrt{\frac{\log(M)}{2n}}.
\end{aligned}$$

Corollary : PAC-Bayes bounds

Define a new probability measure $\mathbb{E}_{\mathcal{S}}[\hat{\rho}]$ by

$$\forall \text{ event } E,\ \mathbb{E}_{\mathcal{S}}[\hat{\rho}](E) = \mathbb{E}_{\mathcal{S}}[\hat{\rho}(E)].$$

Classical property of the mutual information :

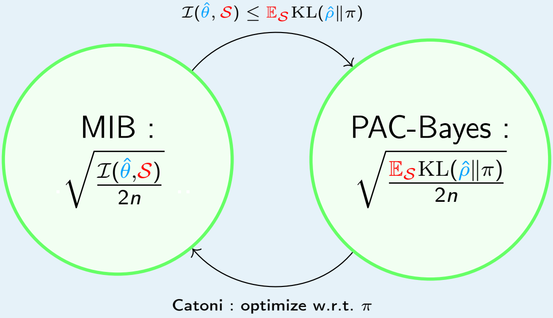
$$\mathcal{I}(\hat{\theta}, \mathcal{S}) = \mathbb{E}_{\mathcal{S}} \mathrm{KL}(\hat{\rho} \| \mathbb{E}_{\mathcal{S}}[\hat{\rho}]) = \inf_{\pi} \mathbb{E}_{\mathcal{S}} \mathrm{KL}(\hat{\rho} \| \pi).$$

Fix a "prior distribution" π, then the MIB implies the following

Corollary - PAC-Bayes bound (in expectation)

$$\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\hat{\theta}} R(\hat{\theta}) \leq \mathbb{E}_{\mathcal{S}} \mathbb{E}_{\hat{\theta} \sim \hat{\rho}} R_n(\hat{\theta}) + \sqrt{\frac{\mathbb{E}_{\mathcal{S}} \mathrm{KL}(\hat{\rho} \| \pi)}{2n}}.$$

MIB and PAC-Bayes bounds



Catoni, O. (2007). *PAC-Bayesian supervised classification : the thermodynamics of statistical learning*. IMS Monograph series.

Russo, D. and Zou, J. (2019). How much does your data exploration overfit ? controlling bias via information usage. *IEEE Transactions on Information Theory*.

Alquier, P. (2024). *User-friendly introduction to PAC-Bayes bounds*. Foundations and Trends® in Machine Learning.

Hellström, F., Durisi, G., Guedj, B. and Raginsky, M. (2025). *Generalization bounds : Perspectives from information theory and PAC-Bayes*. Foundations and Trends® in Machine Learning.

Statistical inference framework

We now observe a sample $\mathcal{S} = (X_1, \dots, X_n)$ of n variables i.i.d from P.

We are given a "model", that is a set $(P_\theta, \theta \in \Theta)$ of probability distributions, and the promise that $P = P_{\theta_0}$ for some $\theta_0 \in \Theta$.

Our objective is to estimate θ_0 from $\mathcal{S}$.

Assuming that the P_θ's have densities p_θ,a classical estimation methods is the maximum likelihood estimator (MLE) :

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} p_\theta(X_i).$$

Remarks on the MLE

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta\in\Theta} \prod_{i=1}^{n} p_\theta(X_i)$$
$$= \arg\max_{\theta\in\Theta} \frac{\prod_{i=1}^{n} p_\theta(X_i)}{\prod_{i=1}^{n} p_{\theta_0}(X_i)}$$
$$= \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)}.$$

The MLE can be seen a special case of ERM with the risk

$$R_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)} \xrightarrow[n\to\infty]{a.s.} KL(P_{\theta_0}\|P_\theta) =: R(\theta).$$

Better notation : "log-likelihood ratio"

$$LR_n(\theta_0,\theta) := \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)}.$$

What kind of bound can we hope for ?

By analogy with the MIB stated earlier, we could conjecture, for a parameter $\hat{\theta} \sim \hat{\rho}(\mathcal{S})$:

$$\left| \mathbb{E}_{\mathcal{S}} \mathbb{E}_{\hat{\theta}} \left(KL(P_{\theta_0} \| P_{\hat{\theta}}) - LR_n(\theta_0, \hat{\theta}) \right) \right| \leq \sqrt{\frac{\mathcal{I}(\hat{\theta}, \mathcal{S})}{2n}}.$$

However, the loss function $\log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)}$ is not bounded in general, and thus we cannot apply Russo & Zou's MIB here.

It appears that this conjecture is wrong, so you are going to forget I ever mentioned it !

Statistical divergences

The α-Rényi divergence for $\alpha \in (0,1)$

$$D_\alpha(Q\|R) = \frac{1}{\alpha - 1} \log \int [Q(\mathrm{d}x)]^\alpha [R(\mathrm{d}x)]^{1-\alpha}.$$

The Hellinger distance

$$\mathcal{H}(Q,R) = \sqrt{\frac{1}{2} \int \left(\sqrt{Q(\mathrm{d}x)} - \sqrt{R(\mathrm{d}x)} \right)^2}.$$

These are strongly related. For example, for $1/2 \leq \alpha$:

$$\mathcal{H}^2(Q,R) \leq D_\alpha(Q\|R) \xrightarrow[\alpha \nearrow 1]{} \mathrm{KL}(R\|Q).$$

T. Van Erven & P. Harremos (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*.

MIBfor statistical inference

Theorem – MIB for statistics

Fix $\alpha \in (0,1)$, then

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\hat{\theta}}\left(D_\alpha(P_{\hat{\theta}}\|P_{\theta_0}) - \frac{\alpha}{1-\alpha}LR_n(\theta_0,\hat{\theta})\right) \leq \frac{\mathcal{I}(\hat{\theta},\mathcal{S})}{n(1-\alpha)}.$$

In particular, for $\alpha = 1/2$, we obtain :

Corollary

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\hat{\theta}}\left(\mathcal{H}^2(P_{\hat{\theta}},P_{\theta_0}) - LR_n(\theta_0,\hat{\theta})\right) \leq \frac{2\,\mathcal{I}(\hat{\theta},\mathcal{S})}{n}.$$

Remarks on the MIB for statistics

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\hat{\theta}}\left(\mathcal{H}^2(P_{\hat{\theta}}, P_{\theta_0}) - LR_n(\theta_0, \hat{\theta})\right) \leq \frac{2\mathcal{I}(\hat{\theta}, \mathcal{S})}{n}.$$

- Note the "fast rate" in $1/n$ instead of $1\sqrt{n}$.
- On the other hand, our risk $\mathcal{H}^2(P_\theta, P_{\theta_0}) \leq \mathrm{KL}(P_{\theta_0} \| P_\theta)$: this is weaker than what we were hoping for.
- Under suitable differentiability assumptions on $\log p_\theta(x)$,

$$\mathcal{H}^2(P_\theta, P_{\theta_0}) = \frac{1}{4}(\theta - \theta_0)^T J(\theta_0)(\theta - \theta_0)^T + o\left(\|\theta - \theta_0\|^2\right)$$
$$\mathrm{KL}(P_{\theta_0} \| P_\theta) = \frac{1}{2}(\theta - \theta_0)^T J(\theta_0)(\theta - \theta_0)^T + o\left(\|\theta - \theta_0\|^2\right)$$

where $J(\cdot)$ is the Fisher information,

$$J(\theta_0) = \mathbb{E}_{X \sim P_\theta}\left[\left(\frac{\partial}{\partial\theta} \log p_\theta(X)\right)^2\right].$$

Consequences of the MIB

Reminder – for MIB statistics

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\hat{\theta}}\left(D_{\alpha}(P_{\hat{\theta}}\|P_{\theta_0}) - \frac{\alpha}{1-\alpha}LR_n(\theta_0,\hat{\theta})\right) \leq \frac{\mathcal{I}(\hat{\theta},\mathcal{S})}{n(1-\alpha)}.$$

Until the end of the talk, let us investigate some consequences of this result :

1. PAC-Bayes bounds, which motivate "tempered posterior distributions",
2. rates of convergence of tempered posteriors,
3. rates of convergence of variational approximations,
4. rates for the MLE.

Corollary : PAC-Bayes bounds

Corollary – PAC-Bayes bound for statistics

Fix $\alpha \in (0,1)$ and a prior π,

$$\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\hat{\theta}} D_\alpha(P_{\hat{\theta}} \| P_{\theta_0}) \leq \frac{\mathbb{E}_{\mathcal{S}} \left[\mathbb{E}_{\hat{\theta}\sim\hat{\rho}} \left[\alpha \, LR_n(\theta_0, \hat{\theta}) \right] + \frac{\mathrm{KL}(\hat{\rho}\|\pi)}{n} \right]}{1-\alpha}.$$

This result was proven in :

Alquier, P. and Ridgway, J. (2020). Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*.

based on techniques from :

Bhattacharya, A., Pati, D. and Yang, Y. (2019). Bayesian fractional posteriors. *The Annals of Statistics*.

Key lemma for the minimization of the bound

Donsker and Varadhan variational inequality

Let π be a probability distribution. Let $h(\cdot)$ such that $\int \exp(-h(\vartheta))\pi(\mathrm{d}\vartheta) < \infty$. Define

$$\pi_h(\mathrm{d}\theta) = \frac{\exp(-h(\theta))}{\int \exp(-h(\vartheta))\pi(\mathrm{d}\vartheta)}\pi(\mathrm{d}\theta).$$

Then

$$\pi_h = \arg\min_p \left[\int h(\theta)p(\mathrm{d}\theta) + \mathrm{KL}(p\|\pi)\right].$$

Minimization of the PAC-Bayes bound

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\hat{\theta}} D_{\alpha}(P_{\hat{\theta}} \| P_{\theta_0}) \leq \frac{\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\hat{\theta}\sim\hat{\rho}}\left[\alpha\, LR_n(\theta_0, \hat{\theta})\right] + \frac{\mathrm{KL}(\hat{\rho}\|\pi)}{n}\right]}{1-\alpha}.$$

The right-hand side is minimized by

$$\begin{aligned}\hat{\rho}(\mathrm{d}\theta) &= \pi_{n\alpha LR_n}(\mathrm{d}\theta) \\ &\propto \exp\left(-\alpha n LR_n(\theta_0,\theta)\right)\pi(\mathrm{d}\theta) \\ &= \left(\prod_{i=1}^{n} p_{\theta}(X_i)\right)^{\alpha} \pi(\mathrm{d}\theta).\end{aligned}$$

Terminology from Bayesian statistics

The posterior distribution : $\left(\prod_{i=1}^{n} p_{\theta}(X_i)\right)\pi(\mathrm{d}\theta)$.

Tempered posterior : $\left(\prod_{i=1}^{n} p_{\theta}(X_i)\right)^{\alpha}\pi(\mathrm{d}\theta)$.

A complete example : Gaussian mean estimation

- $X_1, \dots, X_n$ i.i.d. $\mathcal{N}(\theta_0, I_d)$.
- $\pi = \mathcal{N}(0, \sigma^2 I_d)$.
- $\mathcal{D}_\alpha(P_\theta \| P_{\theta_0}) = \frac{\alpha}{2}\|\theta - \theta_0\|^2$ and $\mathrm{KL}(P_{\theta_0}\|P_\theta) = \frac{1}{2}\|\theta - \theta_0\|^2$.
- $\hat{\rho} = \pi_{n\alpha LR_n} = \mathcal{N}\left(\frac{\sum_{i=1}^n X_i}{n + \frac{1}{\alpha\sigma^2}}, \frac{\frac{1}{\alpha}}{n + \frac{1}{\alpha\sigma^2}} I_d \right)$.

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\hat{\theta}\sim\pi_{n\alpha LR_n}} D_\alpha(P_{\hat{\theta}}\|P_{\theta_0}) \leq \frac{\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\hat{\theta}}\left[\alpha\, LR_n(\theta_0, \hat{\theta})\right] + \frac{\mathrm{KL}(\hat{\rho}\|\pi)}{n}\right]}{1-\alpha}.$$

- $\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\hat{\theta}}\left[\alpha\, LR_n(\theta_0, \hat{\theta})\right] = \mathcal{O}\left(\frac{d}{n}\right)$.
- KL term :

$$\mathrm{KL}(\hat{\rho}\|\pi) = \frac{1}{2}\left[\frac{\frac{d}{\alpha\sigma^2}}{n + \frac{1}{\alpha\sigma^2}} - d + \frac{1}{2\sigma^2}\left\| \frac{\sum_{i=1}^n X_i}{n + \frac{1}{\alpha\sigma^2}} \right\|^2 + d \log \frac{n + \frac{1}{\alpha\sigma^2}}{\frac{1}{\alpha}} \right]$$

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\hat{\theta}\sim\pi_{n\alpha LR_n}} \|\hat{\theta} - \theta_0\|^2 \leq \mathcal{O}\left(\frac{d\log(n)}{n}\right).$$

A complete example : Gaussian mean estimation

More generally, when the parameter is d-dimensional, we obtain rates in $\mathcal{O}(\frac{d}{n}\log(n))$ as in :

Alquier, P. and Ridgway, J. (2020). Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*.

Bhattacharya, A., Pati, D. and Yang, Y. (2019). Bayesian fractional posteriors. *The Annals of Statistics*.

Solution : use the MIB bound !

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\hat{\theta}\sim\pi_{n\alpha LR_n}}\left(D_\alpha(P_{\hat{\theta}}\|P_{\theta_0}) - \frac{\alpha}{1-\alpha}LR_n(\theta_0,\hat{\theta})\right) \leq \frac{\mathcal{I}(\hat{\theta},\mathcal{S})}{n(1-\alpha)}$$

$$\mathcal{I}(\hat{\theta},\mathcal{S}) = \inf_{\pi}\mathbb{E}_{\mathcal{S}}\mathrm{KL}(\pi_{n\alpha LR_n}\|\pi) \leq \mathbb{E}_{\mathcal{S}}\mathrm{KL}(\pi_{n\alpha LR_n(\cdot)}\|\pi_{n\beta D_\alpha(P_\cdot\|P_{\theta_0})})$$

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\hat{\theta}\sim\pi_{n\alpha LR_n}}\|\hat{\theta}-\theta_0\|^2 \leq \frac{4d + \frac{\|\theta_0\|^2}{2\sigma^2}}{\alpha(1-\alpha)^2 n}.$$

Rates of convergence : general case

Fix $\alpha \in (0,1)$.

Assumption 1

There is a constant c_α such that :

$$\forall \theta \in \Theta, \, \mathrm{KL}(P_{\theta_0} \| P_\theta) \leq c_\alpha D_\alpha(P_\theta \| P_{\theta_0}).$$

Assumption 2

$$\sup_{\beta > 0} \beta \, \mathbb{E}_{\theta \sim \pi_\beta} \left[\mathrm{KL}(P_{\theta_0} \| P_\theta)\right] =: d < +\infty.$$

Corollary of the MIB for tempered posteriors

Under Assumptions 1 and 2,

$$\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\hat{\theta} \sim \pi_{n\alpha L R_n}} \mathrm{KL}(P_{\theta_0} \| P_{\hat{\theta}}) \leq \alpha \left(\frac{2 c_\alpha}{1-\alpha} \right)^2 \frac{d}{n}.$$

Variational approximations

In general, the tempered posterior is intractable.

Define :

$$\hat{\rho}_{varia.} = \arg\min_{q \in \mathcal{F}} \left\{ \alpha \, \mathbb{E}_{\theta \sim q} LR_n(\theta_0, \theta) + \frac{\mathrm{KL}(q \| \pi)}{n} \right\}$$

where $\mathcal{F}$ is a specified set of "tractable" probability measures.

Assumption 2'

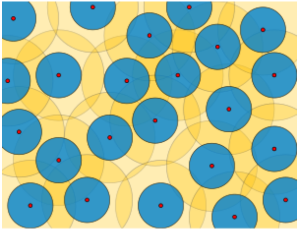
$$\sup_{\beta > 0} \inf_{\rho \in \mathcal{F}} \beta \left\{ \mathbb{E}_{\theta \sim \rho} [\mathrm{KL}(P_{\theta_0} \| P_\theta)] + \frac{\mathrm{KL}(\rho \| \pi_{n\beta D_\alpha})}{n} \right\} =: d' < +\infty.$$

Corollary of the MIB

Under Assumptions 1 and 2,

$$\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\hat{\theta} \sim \hat{\rho}_{varia.}} \mathrm{KL}(P_{\theta_0} \| P_{\hat{\theta}}) \leq \alpha \left(\frac{2 c_\alpha}{1 - \alpha} \right)^2 \frac{d'}{n}.$$

Study of the MLE



- Under a compacity assumption on Θ, there is a finite ε-cover of Θ with cardinality $\mathcal{N}(\Theta,\varepsilon)$.
- Let $\hat{\theta}^{\varepsilon}_{\mathrm{MLE}}$ be the MLE on this finite set.

$$\mathbb{E}_{\mathcal{S}}\left(D_{\alpha}(P_{\hat{\theta}^{\varepsilon}_{\mathrm{MLE}}}\|P_{\theta_0}) - \frac{\alpha}{1-\alpha} LR_n(\theta_0, \hat{\theta}^{\varepsilon}_{\mathrm{MLE}})\right) \leq \frac{\log \mathcal{N}(\Theta,\varepsilon)}{n(1-\alpha)}.$$

Under regularity assumptions (Lipschitz...) on D_{α} and on the log-likelihood,

$$\mathbb{E}_{\mathcal{S}} D_{\alpha}(P_{\hat{\theta}_{\mathrm{MLE}}}\|P_{\theta_0}) \leq C(\alpha)\varepsilon + \frac{\log \mathcal{N}(\Theta,\varepsilon)}{n(1-\alpha)}.$$

Thank you !

ありがとう ございました。