# A New Mutual Information Bound for Statistical Inference

EL Mahdi Khribch[1] Pierre Alquier[1]

[1]ESSEC Business School, Paris

## Problem: From ML to Statistics

Machine Learning Setting:

- Population risk: $R(\theta) := \mathbb{E}_{(X,Y)\sim P}[\ell(Y, f_\theta(X))]$
- Data $\mathcal{S} = ((X_1, Y_1), \dots, (X_n, Y_n))$ i.i.d. from $P$
- Empirical risk: $R_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i, f_\theta(X_i))$
- Randomized estimator $\hat{\theta}$ from $\hat{\rho} = \hat{\rho}(\mathcal{S})$
- Generalization gap: $\text{gen}(\hat{\theta}, \mathcal{S}) = R(\hat{\theta}) - R_n(\hat{\theta})$

**Mutual Information Definition:** For random variables $(U, V) \sim Q$ with marginals $Q_U$ and $Q_V$:

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

Measures the statistical dependence between $U$ and $V$.

Classic MIB (Catoni 2007; Russo & Zou 2019):
For **bounded** losses $0 \le \ell \le 1$:

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

## Wait... MIB? You mean Men In Black?

*No, but this bound also protects us from statistical aliens!*
MIB = Mutual Information Bound

**Statistical Challenge:** Log-likelihood ratios are UNBOUNDED!

$$\ell(\theta, X_i) = \log\frac{p_{\theta_0}(X_i)}{p_\theta(X_i)} \in (-\infty, +\infty).$$

Classic MIB breaks down for statistical inference!

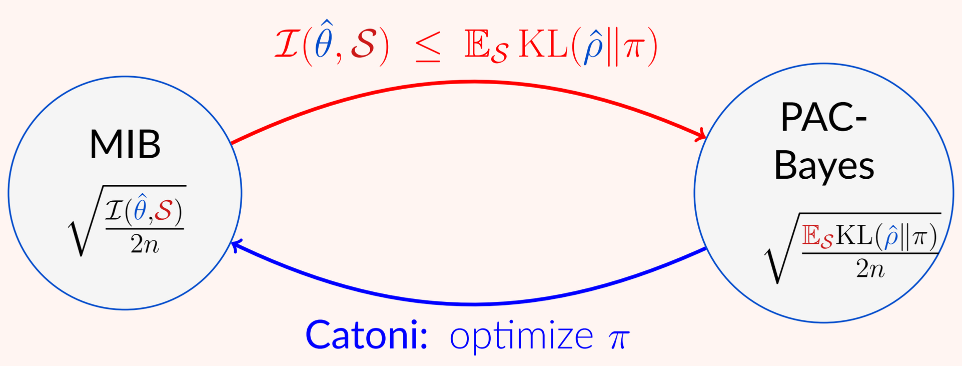
## The MIB PAC-Bayes Bridge

Fundamental Connection:

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

## Corollary - PAC Bayes bound (in expectation)

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




The connection between mutual information bounds and PAC-Bayesian theory provides a framework for developing new bounds for Statistical Inference.

## Statistical Divergences

**Key Innovation:** Replace KL with $\alpha$-Rényi divergence

**$\alpha$-Rényi divergence** for $\alpha \in (0, 1)$:

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

Hellinger distance:

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

## Statistical Inference Framework

**Setting:** We observe a sample $\mathcal{S} = (X_1, \dots, X_n)$ of $n$ variables i.i.d from $P = P_{\theta_0}$ in a model $(P_\theta, \theta \in \Theta)$ for some unknown $\theta_0 \in \Theta$.

**Objective:** Estimate $\theta_0$ from $\mathcal{S}$.

Assuming that the $P_\theta$'s have densities $p_\theta$, a classical estimation methods is the maximum likelihood estimator (MLE):

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

**Key Notation:** "Log-likelihood ratio"

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

## Our Main Result

### *Theorem (MIB for Statistical Inference)*

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

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

Special case ($\alpha = 1/2$):

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

Key features:

- Fast rate: $O(1/n)$ instead of $O(1/\sqrt{n})$
- Trade-off: Weaker metric ($\mathcal{H}^2 \le \mathrm{KL}$)

## PAC-Bayes for Statistics

### *Corollary (PAC-Bayes bound for statistics)*

*For $\alpha \in (0, 1)$ and prior $\pi$:*

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

Tempered posterior (minimizes bound):

$$\pi_{n,\alpha}(\mathrm{d}\theta) \propto \left(\prod_{i=1}^n p_\theta(X_i)\right)^\alpha \pi(\mathrm{d}\theta).$$

## Optimal Rates of Convergence

Assumptions:

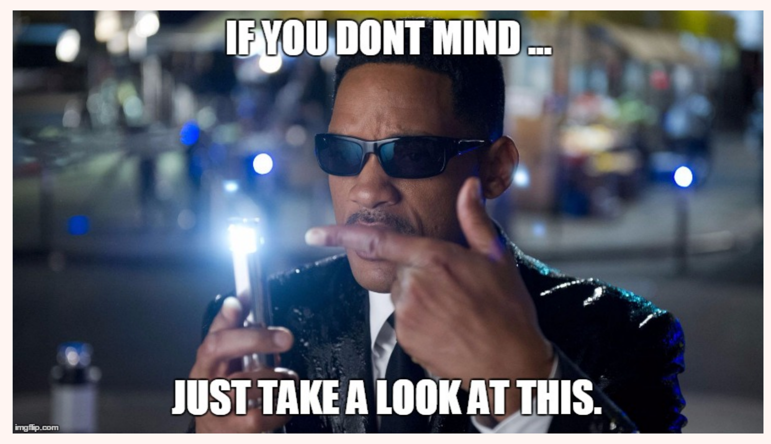
1. **Rényi-KL relation:** $\exists c_\alpha > 0 : \mathrm{KL}(P_{\theta_0}\|P_\theta) \le c_\alpha D_\alpha(P_\theta\|P_{\theta_0})$.
2. **Catoni's dimension:** $d := \sup_{\beta>0}\beta\,\mathbb{E}_{\theta\sim\pi_\beta}[\mathrm{KL}(P_{\theta_0}\|P_\theta)] < +\infty$.

Convergence Rate for Tempered Posteriors:

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

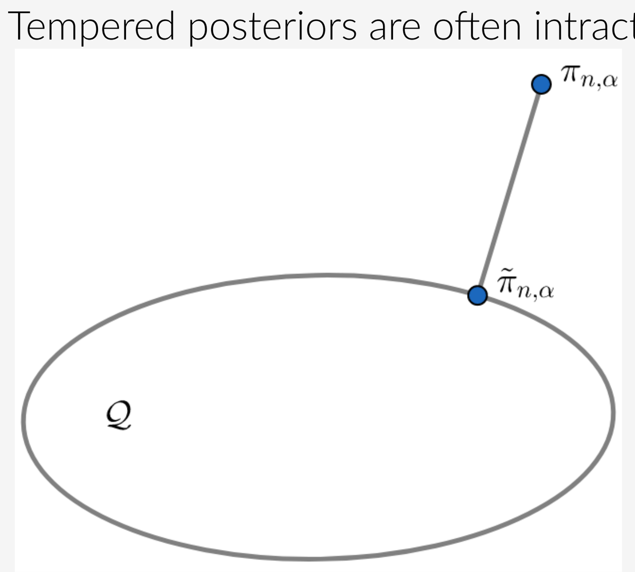
**Achievement:** We obtain the optimal $O(d/n)$ rate, eliminating the suboptimal $\log(n)$ factor present in traditional $O(d\log(n)/n)$ bounds!

## The $\log(n)$ factor was never here...




## Variational Approximation

**Challenge:** Tempered posteriors are often intractable




**Solution:** Variational approximation in family $\mathcal{Q}$:

$$\tilde{\pi}_{n,\alpha} = \underset{q\in\mathcal{Q}}{\mathrm{argmin}}\left\{\alpha\mathbb{E}_{\theta\sim q}LR_n(\theta_0, \theta) + \frac{\mathrm{KL}(q\|\pi)}{n}\right\}$$

Assumption for Variational Rates

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

Variational Rate Guarantee:

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

## Summary of Contributions

- ✓ **New MIB** for Statistical inference via Rényi divergences
- ✓ **Optimal rates:** $O(d/n)$ without $\log(n)$
- ✓ **Unified theory:** MIB ↔ PAC-Bayes
- ✓ **Practical:** Variational inference with guarantees

## References

1. Alquier, P., Ridgway, J. (2020). Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*.
2. Catoni, O. (2007). *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. Institute of Mathematical Statistics.
3. Russo, D., & Zou, J. (2019). How much does your data exploration overfit? Controlling bias via information usage. *IEEE Transactions on Information Theory*.
4. Bhattacharya, A., Pati, D. and Yang, Y. (2019). Bayesian fractional posteriors. *The Annals of Statistics*.