PAC-Bayesian Bounds based on the Rényi Divergence

Luc Bégin[1], Pascal Germain[2], François Laviolette[3], Jean-Francis Roy[3]

[1] Campus d'Edmundston, Université de Moncton, Nouveau-Brunswick, Canada
[2] INRIA - Sierra Project-Team, École Normale Supérieure, Paris, France
[3] Département d'informatique et de génie logiciel, Université Laval, Québec, Canada

Classification Setting and PAC-Bayesian Basics

Training set: We draw m examples *i.i.d.* from a distribution D on $\mathcal{X}\times\mathcal{Y}$:

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \sim D^m \,.$$

Learner: Learn a *classifier* $h : \mathcal{X} \to \mathcal{Y}$ that has a low generalization risk on D:

$$R_D(h) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(x,y)\sim D} I\big[h(x) \neq y\big] \,, \text{ where } I(a) = 1 \text{ if } a \text{ is true and 0 otherwise.}$$

PAC-Bayesian learning: Given a set $\mathcal{H}$ of *voters* $\mathcal{X} \to \{-1, 1\}$ and a *training set* S, we consider

1. The *prior* distribution P on $\mathcal{H}$ encodes prior knowledge.
2. The *posterior* distribution Q on $\mathcal{H}$ is obtained by learning from S.

PAC-Bayesian theory traditionally bounds the *Gibbs risk* $R_D(G_Q) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim Q} R_D(h)$ using its **empirical value** $R_S(G_Q) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim Q} R_S(h)$ and a the **Kullback-Leibler** divergence between Q and P. General PAC-Bayesian theorems are tools to derive various PAC-Bayesian bounds using **any convex function** $\Delta : [0,1] \times [0,1] \to \mathbb{R}$.

Classical PAC-Bayesian Theory

The *Kullback-Leibler divergence* between distributions Q and P is given by

$$\mathrm{KL}(Q\|P) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim Q} \ln \frac{Q(h)}{P(h)} \,.$$

Lemma 3 *(Kullback-Leibler change of measure)* For any set $\mathcal{H}$, for any distributions P and Q on $\mathcal{H}$, and for any measurable function $\phi : \mathcal{H} \to \mathbb{R}$, we have

$$\mathop{\mathbf{E}}_{h\sim Q} \phi(h) \;\leq\; \mathrm{KL}(Q\|P) + \ln\left(\mathop{\mathbf{E}}_{h\sim P} e^{\phi(h)}\right).$$

Theorem 4 For any distribution D, for any set $\mathcal{H}$ of voters $\mathcal{X} \to \{-1, 1\}$, for any prior P on $\mathcal{H}$, for any $\delta \in (0, 1]$, for any $m' > 0$, and for any convex function Δ, with probability at least $1-\delta$ over the choice of $S \sim D^m$, we have

$$\forall Q \text{ on } \mathcal{H}\colon \quad \Delta\Big(R_S(G_Q), R_D(G_Q)\Big) \;\leq\; \frac{1}{m'}\left[\mathrm{KL}(Q\|P) + \ln \frac{\mathcal{I}^K_\Delta(m, m')}{\delta}\right],$$

where
$\mathcal{I}^K_\Delta(m, m') \stackrel{\text{def}}{=} \sup_{r\in[0,1]} \left[\sum_{k=0}^{m} \mathrm{Bin}^m_k(r)\, e^{m'\Delta\left(\frac{k}{m}, r\right)}\right], \quad \mathrm{Bin}^m_k(r) \stackrel{\text{def}}{=} \binom{m}{k} r^k (1-r)^{m-k}.$

Two commons Δ are $\Delta_{\mathrm{KL}}(q,p) \stackrel{\text{def}}{=} q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p} \geq \Delta_{V^2}(q,p) \stackrel{\text{def}}{=} 2(q-p)^2$:

Corollary 6 *(Seeger, 2002; McAllester, 2003)* With probability at least $1-\delta$,

$$\forall Q \text{ on } \mathcal{H}\colon \quad \text{a) } \Delta_{\mathrm{KL}}\big(R_S(G_Q), R_D(G_Q)\big) \leq \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{m}}{\delta}\right],$$

$$\text{b) } R_D(G_Q) \;\leq\; R_S(G_Q) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(Q\|P) + \ln\frac{2\sqrt{m}}{\delta}\right]}.$$

Rényi PAC-Bayesian Theory

For any $\alpha > 1$, the *Rényi divergence* between distributions Q and P is given by

$$D_\alpha(Q\|P) \stackrel{\text{def}}{=} \frac{1}{\alpha-1}\ln\left[\mathop{\mathbf{E}}_{h\sim P}\left(\frac{Q(h)}{P(h)}\right)^\alpha\right]. \qquad D_\alpha(Q\|P) = \mathrm{KL}(Q\|P) \text{ when } \alpha \to 1.$$

Theorem 8 *(Rényi change of measure)* For any set $\mathcal{H}$, for any distributions P and Q on $\mathcal{H}$, for any $\alpha > 1$, and for any measurable function $\phi : \mathcal{H} \to \mathbb{R}$, we have

$$\frac{\alpha}{\alpha-1}\ln \mathop{\mathbf{E}}_{h\sim Q} \phi(h) \;\leq\; D_\alpha(Q\|P) + \ln\left(\mathop{\mathbf{E}}_{h\sim P} \phi(h)^{\frac{\alpha}{\alpha-1}}\right).$$

Theorem 9 For any distribution D, for any set $\mathcal{H}$ of voters $\mathcal{X} \to \{-1, 1\}$, for any prior P on $\mathcal{H}$, for any $\delta \in (0, 1]$, for any $\alpha > 1$, and for any convex function Δ, with probability at least $1-\delta$ over the choice of $S \sim D^m$, we have

$$\forall Q \text{ on } \mathcal{H}\colon \quad \ln\Delta\Big(R_S(G_Q), R_D(G_Q)\Big) \;\leq\; \frac{1}{\alpha'}\left[D_\alpha(Q\|P) + \ln\frac{\mathcal{I}^{\mathrm{R}}_\Delta(m, \alpha')}{\delta}\right],$$

where $\alpha' = \frac{\alpha}{\alpha-1}$, $\quad \mathcal{I}^{\mathrm{R}}_\Delta(m, \alpha') \stackrel{\text{def}}{=} \sup_{r\in[0,1]}\left[\sum_{k=0}^{m} \mathrm{Bin}^m_k(r)\, \Delta\big(\tfrac{k}{m}, r\big)^{\alpha'}\right].$

By choosing $\alpha = 2$, and $\Delta(q,p) = p - q$, we obtain as a special case:

Corollary 10 *($\approx$ Honorio and Jaakkola, 2014)* With probability at least $1-\delta$,

$$\forall Q \text{ on } \mathcal{H}\colon \quad R_D(G_Q) \leq R_S(G_Q) + \sqrt{\frac{\chi^2(Q\|P) + 1}{4m\,\delta}},$$

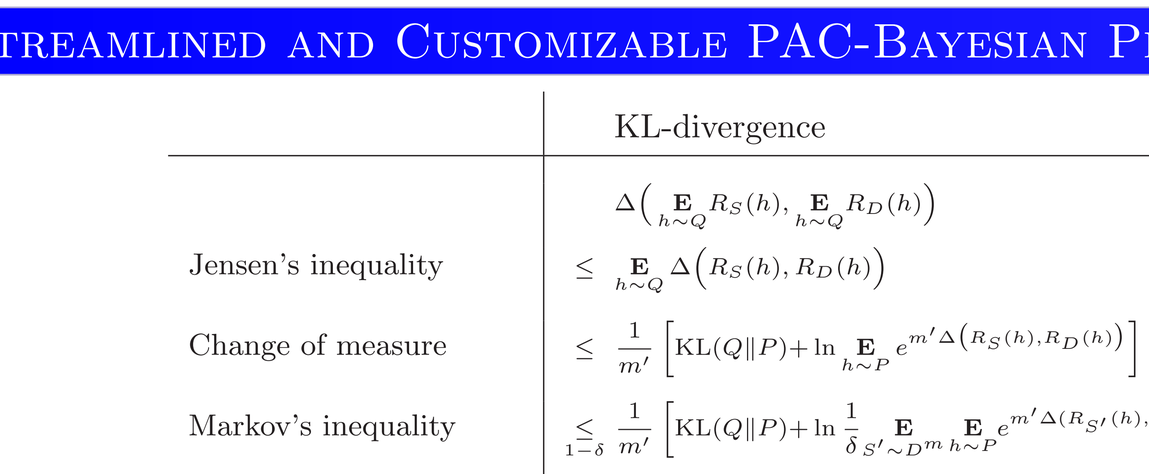
where $\chi^2(Q\|P) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim P}\left[\left(\frac{Q(h)}{P(h)}\right)^2 - 1\right]$ is the *chi-squared divergence*.

Streamlined and Customizable PAC-Bayesian Proofs

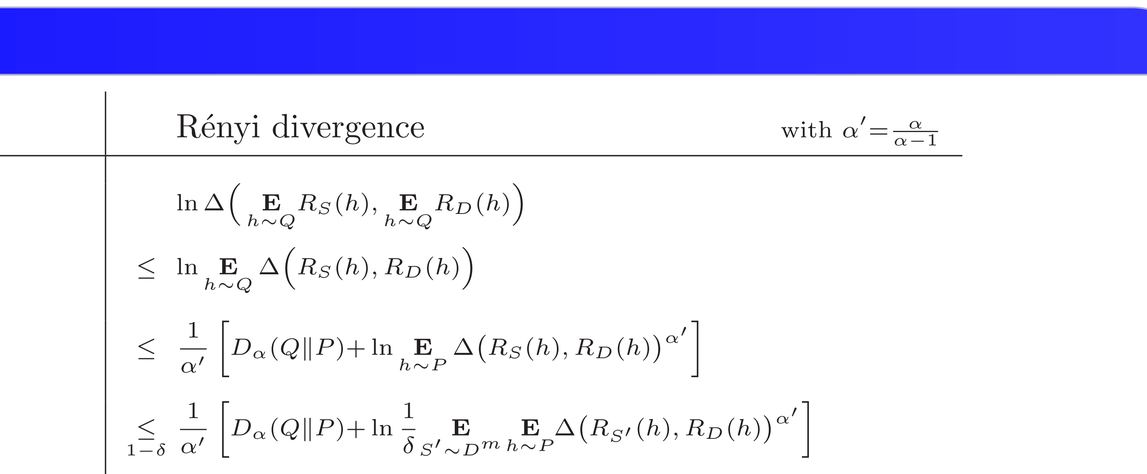
	KL-divergence	Rényi divergence (with $\alpha' = \frac{\alpha}{\alpha-1}$)
	$\Delta\Big(\mathop{\mathbf{E}}_{h\sim Q} R_S(h), \mathop{\mathbf{E}}_{h\sim Q} R_D(h)\Big)$	$\ln\Delta\Big(\mathop{\mathbf{E}}_{h\sim Q} R_S(h), \mathop{\mathbf{E}}_{h\sim Q} R_D(h)\Big)$
Jensen's inequality	$\leq \mathop{\mathbf{E}}_{h\sim Q} \Delta\Big(R_S(h), R_D(h)\Big)$	$\leq \ln \mathop{\mathbf{E}}_{h\sim Q} \Delta\Big(R_S(h), R_D(h)\Big)$
Change of measure	$\leq \frac{1}{m'}\left[\mathrm{KL}(Q\|P) + \ln \mathop{\mathbf{E}}_{h\sim P} e^{m'\Delta\left(R_S(h), R_D(h)\right)}\right]$	$\leq \frac{1}{\alpha'}\left[D_\alpha(Q\|P) + \ln \mathop{\mathbf{E}}_{h\sim P} \Delta\big(R_S(h), R_D(h)\big)^{\alpha'}\right]$
Markov's inequality	$\underset{1-\delta}{\leq} \frac{1}{m'}\left[\mathrm{KL}(Q\|P) + \ln \frac{1}{\delta} \mathop{\mathbf{E}}_{S'\sim D^m} \mathop{\mathbf{E}}_{h\sim P} e^{m'\Delta(R_{S'}(h), R_D(h))}\right]$	$\underset{1-\delta}{\leq} \frac{1}{\alpha'}\left[D_\alpha(Q\|P) + \ln \frac{1}{\delta} \mathop{\mathbf{E}}_{S'\sim D^m} \mathop{\mathbf{E}}_{h\sim P} \Delta\big(R_{S'}(h), R_D(h)\big)^{\alpha'}\right]$
Expectations swap	$= \frac{1}{m'}\left[\mathrm{KL}(Q\|P) + \ln \frac{1}{\delta} \mathop{\mathbf{E}}_{h\sim P} \mathop{\mathbf{E}}_{S'\sim D^m} e^{m'\Delta(R_{S'}(h), R_D(h))}\right]$	$= \frac{1}{\alpha'}\left[D_\alpha(Q\|P) + \ln \frac{1}{\delta} \mathop{\mathbf{E}}_{h\sim P} \mathop{\mathbf{E}}_{S'\sim D^m} \Delta\big(R_{S'}(h), R_D(h)\big)^{\alpha'}\right]$
Binomial law	$= \frac{1}{m'}\left[\mathrm{KL}(Q\|P) + \ln \frac{1}{\delta} \mathop{\mathbf{E}}_{h\sim P} \sum_{k=0}^{m} \mathrm{Bin}^m_k\big(R_D(h)\big) e^{m'\Delta\left(\frac{k}{m}, R_D(h)\right)}\right]$	$= \frac{1}{\alpha'}\left[D_\alpha(Q\|P) + \ln \frac{1}{\delta} \mathop{\mathbf{E}}_{h\sim P} \sum_{k=0}^{m} \mathrm{Bin}^m_k\big(R_D(h)\big)\, \Delta\big(\tfrac{k}{m}, R_D(h)\big)^{\alpha'}\right]$
Supremum over risk	$\leq \frac{1}{m'}\left[\mathrm{KL}(Q\|P) + \ln \frac{1}{\delta} \sup_{r\in[0,1]} \left\{\sum_{k=0}^{m} \mathrm{Bin}^m_k(r)\, e^{m'\Delta\left(\frac{k}{m}, r\right)}\right\}\right]$	$\leq \frac{1}{\alpha'}\left[D_\alpha(Q\|P) + \ln \frac{1}{\delta} \sup_{r\in[0,1]} \left\{\sum_{k=0}^{m} \mathrm{Bin}^m_k(r)\, \Delta\left(\frac{k}{m}, r\right)^{\alpha'}\right\}\right]$

Empirical Study (How the bound values are impacted by each inequality of our proof?)

Each example generated by D is a random draw among the 8124 examples of the *mushroom* dataset. That is, the training set $S \sim D^m$ contains m examples drawn *with replacement* and *uniform probability* from the full dataset. From training set S, we learn a majority vote using AdaBoost. We compare three different kinds of voters.



Values for each inequality computed with the three kinds of voters. The dashed lines correspond to the traditional bounds with the Kullback-Leibler divergence. The full lines correspond to the bounds considering the Rényi divergence.



Alternate representation of the quantities obtained using the weak decision trees. The blue curve corresponds to the function $\Delta(R_D(G_Q), r)$. Each dashed horizontal line corresponds to the value given by the right-hand side of the bound after each inequality.