A short introduction to information theory and Bayesian Model Averaging

Olivier Catoni

May 3, 2012

1. Information theory and lossless codes

1.1. Binary codes. Let us consider a finite alphabet A and a random sequence $(X_n)_{n=1}^{\infty}$ taking its values in A. (The alphabet can be any finite set here.)

Let $\{0,1\}^* = \bigcup_{n=1}^{\infty} \{0,1\}^n$ be the set of finite binary sequences.

Definition 1.1 *Given some block length n, a binary code c is an injective map from A^n to $\{0,1\}^n$. The mean length of c is*

$$\mathbb{E}\big\{\ell\big[c(X^n)\big]\big\},$$

where the expectation is taken with respect to the distribution of the sequence $X^n \stackrel{\text{def}}{=} (X_1, \dots, X_n)$ and where the length function ℓ is defined as $\ell(w) = k$ for any $w \in \{0,1\}^k$.

Minimizing the mean code length under various assumptions on the source and the set of authorized codes is the main subject of lossless coding theory.

1.2. Optimal code length for a known source. In the case when $\mathbb{P}_{X^n}$ (the distribution of $X^n = (X_1, \dots, X_n)$) is known and c is arbitrary, minimizing the code length is achieved by sorting A^n in order of decreasing probabilities. More precisely, let us write $A^n = \{b_i, i = 1, \dots, d^n\}$, where $d = |A|$ is the size of the alphabet and where the blocks of n letters are indexed by order of decreasing probabilities:

$$\mathbb{P}_{X^n}(b_i) \leq \mathbb{P}_{X^n}(b_{i-1}), i = 2, \dots, d^n.$$

Let us sort $\{0,1\}^*$ by order of increasing lengths, writing

$$\{0,1\}^* = \big\{w(j), j \in \mathbb{N} \setminus \{0\}\big\},$$

CNRS – UMR 8553, Département de Mathématiques et Applications, Ecole Normale Supérieure, 45, rue d'Ulm, F75230 Paris cedex 05, and INRIA Paris-Rocquencourt – CLASSIC team.

where $w(1) = 0$, $w(2) = 1$, $w(3) = 00$, $w(4) = 01$, ... and more generally

$$w\left(2^\ell + \sum_{k=1}^{\ell} d_{\ell-k} 2^k - 1\right) = \big(d_k\big)_{k=1}^{\ell},$$

where $d_k \in \{0, 1\}$ (this means that $w(i)$ is the binary representation of $i + 1$ read from left to right, after removing the leftmost bit, which is always equal to one). Let us remark that $\ell\big[w(i)\big] = \lfloor \log_2(i + 1) \rfloor$.

Proposition 1.1 *The code $c(b_i) = w(i)$ minimizes the mean code length among all binary codes.*

Proof. Let c' be some other code. Its mean length can be written as

$$\mathbb{E}\Big\{\ell\big[c'(X^n)\big]\Big\} = \sum_{k=1}^{\infty} \mathbb{P}\Big\{\ell\big[c'(X^n)\big] \geq k\Big\}. \tag{1}$$

On the other hand, since $i \mapsto \ell\big[w(i)\big]$ is non-decreasing, for any code length k,

$$\big\{i \in \mathbb{N} \setminus \{0\} : \ell\big[w(i)\big] < k\big\} = [\![1, i_k]\!],$$

where $i_k = 2^k - 2$ is the number of binary codes of length less than k, and $[\![1, i_k]\!] \stackrel{\text{def}}{=} \{1, \ldots, i_k\}$. Thus

$$\mathbb{P}\Big\{\ell\big[c'(X^n)\big] < k\Big\} = \mathbb{P}_{X^n}\Big\{c'^{-1}\big[w\big([\![1, i_k]\!]\big)\big]\Big\}.$$

As c' is an injective map, $\big|c'^{-1}\big[w\big([\![1, i_k]\!]\big)\big]\big| \leq i_k$, and can thus be written as $\big\{b_{j_1}, \ldots, b_{j_m}\big\}$, where $j_1 < j_2 < \cdots < j_m$ and $m \leq i_k$. Thus

$$\begin{aligned}
\mathbb{P}\Big\{\ell\big[c'(X^n)\big] < k\Big\} &= \mathbb{P}_{X^n}(b_{j_1}) + \cdots + \mathbb{P}_{X^n}(b_{j_m}) \\
&\leq \mathbb{P}_{X^n}(b_1) + \cdots + \mathbb{P}_{X^n}(b_m) \\
&\leq \mathbb{P}_{X^n}(b_1) + \cdots + \mathbb{P}_{X^n}(b_{i_k}) = \mathbb{P}\Big\{\ell\big[c(X^n)\big] < k\Big\},
\end{aligned}$$

since $s \leq j_s$, $i \mapsto \mathbb{P}_{X^n}(b_i)$ is non-increasing and $m \leq i_k$. This proves that

$$\mathbb{P}_{X^n}\Big\{\ell\big[c'(X^n)\big] \geq k\Big\} \geq \mathbb{P}_{X^n}\Big\{\ell\big[c(X^n)\big] \geq k\Big\},$$

and therefore according to equation (1) that the mean code length of c is optimal. □

The mean length of optimal codes is related to the Shannon entropy $H(\mathbb{P}_{X^n})$ of the distribution of X^n.

Definition 1.2 *The Shannon entropy $H(p)$ of a probability measure $p \in \mathcal{M}_+^1(\mathcal{X})$ on a finite set $\mathcal{X}$ is defined as*

$$H(p) = -\sum_{x\in\mathcal{X}} p(x)\log_2\big[p(x)\big],$$

where $\log_2(z) = \log(z)/\log(2)$. The entropy is measured in bits, for reasons that will become clear later.

Another quantity of interest is the Kullback Leibler divergence, or relative entropy, between two probability measures P and Q.

Definition 1.3 *Let P and $Q \in \mathcal{M}_+^1(\mathcal{X})$ be two probability measures defined on some measurable space $\mathcal{X}$ (that needs not be finite). The Kullback Leibler divergence of P with respect to Q is defined as*

$$\mathcal{K}(P,Q) = \begin{cases} \displaystyle\int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}P & \text{when } P \ll Q, \\ +\infty & \text{otherwise.} \end{cases}$$

Let us remark that when $P \ll Q$, $\log\left(\dfrac{\mathrm{d}P}{\mathrm{d}Q}\right)$ has an integrable negative part, because

$$\int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)_- \mathrm{d}P = \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)_- \frac{\mathrm{d}P}{\mathrm{d}Q}\,\mathrm{d}Q < \infty,$$

where $z_- = \max\{-z, 0\}$, due to the fact that $z \mapsto z\log(z) : \mathbb{R}_+ \to \mathbb{R}$ is bounded from below (by $-1/e$). Thus the integral $\displaystyle\int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}P$ is defined as a generalized integral, taking values in $\mathbb{R}\cup\{+\infty\}$.

Exercise 1 *Give an example, where $P \ll Q$ and $\mathcal{K}(P,Q) = +\infty$.*

Proposition 1.2 *The Kullback Leibler divergence is a non-negative function.*

Proof. When $P \ll Q$, we can write the divergence as

$$\mathcal{K}(P,Q) = \int 1 - \frac{\mathrm{d}P}{\mathrm{d}Q} + \frac{\mathrm{d}P}{\mathrm{d}Q}\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}Q.$$

Since the function $z \mapsto 1 - z + z\log(z) : \mathbb{R}_+ \to \mathbb{R}$ has a positive range, it shows that $\mathcal{K}(P,Q) \in \mathbb{R}_+ \cup \{+\infty\}$. □

Exercise 2 *Show that for any $P_1, P_2, Q \in \mathcal{M}_+^1(\mathcal{X})$, any $\lambda \in [0,1]$,*

$$\mathcal{K}\big(\lambda P_1 + (1-\lambda)P_2, Q\big) \leq \lambda \mathcal{K}(P_1, Q) + (1-\lambda)\mathcal{K}(P_2, Q),$$
$$\mathcal{K}(Q, \lambda P_1 + (1-\lambda)P_2) \leq \lambda \mathcal{K}(Q, P_1) + (1-\lambda)\mathcal{K}(Q, P_2).$$

Proposition 1.3 *The map $p \mapsto H(p) : \mathcal{M}_+^1(\mathcal{X}) \to \mathbb{R}_+$ is concave.*

Proof. The function $z \mapsto z \log_2(z)$ is convex. □

Proposition 1.4 *The Shannon entropy is sub-additive: For any random sequence X^{n+m},*

$$H(\mathbb{P}_{X^{n+m}}) \leq H(\mathbb{P}_{X^n}) + H(\mathbb{P}_{X_{n+1}^{n+m}}),$$

where $X_{n+1}^{n+m} \stackrel{\text{def}}{=} (X_{n+1}, \ldots, X_{n+m})$.

Proof. It is enough to prove that for any couple (X, Y) of random variables,

$$H(\mathbb{P}_{X,Y}) \leq H(\mathbb{P}_X) + H(\mathbb{P}_Y).$$

We can then remark that

$$\begin{aligned} H(\mathbb{P}_{X,Y}) &= -\int \log_2\big[\mathbb{P}_{X,Y}(x,y)\big]\, \mathrm{d}\mathbb{P}_{X,Y}(x,y) \\ &= -\int \log_2\big[\mathbb{P}_X(x)\big]\, \mathrm{d}\mathbb{P}_X(x) - \int \log_2\big[\mathbb{P}_{Y|X=x}(y)\big]\, \mathrm{d}\mathbb{P}_{Y|X=x}(y)\, \mathrm{d}\mathbb{P}_X(x) \\ &= H(\mathbb{P}_X) + \int H\big(\mathbb{P}_{Y|X=x}\big)\, \mathrm{d}\mathbb{P}_X(x) \\ &\leq H(\mathbb{P}_X) + H\left(\int \mathbb{P}_{Y|X=x}\, \mathrm{d}\mathbb{P}_X(x)\right) = H(\mathbb{P}_X) + H(\mathbb{P}_Y), \end{aligned}$$

where we have used the fact that H is concave. □

Definition 1.4 *A random sequence $(X_n)_{n=1}^{+\infty}$ is said to be stationary when, for any $n, m \in \mathbb{N} \setminus \{0\}$, $\mathbb{P}_{X_{n+1}^{n+m}} = P_{X^m}$.*

Proposition 1.5 *If $(X_n)_{n=1}^{+\infty} \stackrel{\text{def}}{=} X^{+\infty}$ is a stationary source, the following limit exists*

$$\lim_{n \to +\infty} \frac{1}{n} H(\mathbb{P}_{X^n}) = \inf_{n=1,\ldots,+\infty} \frac{1}{n} H(\mathbb{P}_{X^n}) \stackrel{\text{def}}{=} \overline{H}(\mathbb{P}_{X^{+\infty}}).$$

It is called the Shannon entropy of $X^{+\infty}$.

Proof. Let $h(n) = n^{-1}H(\mathbb{P}_{X^n})$. From the previous proposition $h(n+m) \leq \frac{nh(n)+mh(m)}{n+m}$. Let $n = pq + r$ where $0 \leq r < p$ be the result of the Euclidean division of n by p. We obtain

$$h(n) \leq \frac{pqh(pq) + rh(r)}{pq+r} \leq \frac{pqh(p) + rh(r)}{pq+r} \leq h(p) + \frac{p-1}{n}\max\bigl\{h(1), \ldots, h(p-1)\bigr\}.$$

Thus $\limsup_n h(n) \leq h(p)$, for any $p \in \mathbb{N}\backslash\{0\}$. This implies that $\limsup_n h(n) \leq \inf_n h(n) \leq \liminf_n h(n)$, proving that $\lim_n h(n)$ exists and is equal to $\inf_n h(n)$. □

Proposition 1.6 *For any source, any optimal code c,*

$$(1-n^{-1})\bigl[H(\mathbb{P}_{X^n}) - \log_2(n-1)\bigr] - 1 \leq \sup_{\alpha>1} \alpha^{-1}\bigl[H(\mathbb{P}_{X^n}) + \log_2(\alpha-1)\bigr] - 1 \leq \mathbb{E}_{X^n}\bigl[\ell\bigl(c(X^n)\bigr)\bigr] \leq H(\mathbb{P}_{X^n}) + 1.$$

Consequently, if the source is stationary,

$$\lim_{n\to+\infty} \frac{1}{n}\mathbb{E}\bigl[\ell\bigl(c(X^n)\bigr)\bigr] = \overline{H}(\mathbb{P}_{X^{+\infty}}).$$

Proof. Let $\mathbb{P}_{X^n}(b_i) \stackrel{\text{def}}{=} p(b_i)$. As

$$ip(b_i) \leq \sum_{j=1}^{i} p(b_j) \leq 1,$$

$p(b_i) \leq i^{-1}$. Since all optimal codes have the same mean length, we can choose the optimal code of Proposition 1.1 (page 2). We get

$$\mathbb{E}\bigl[\ell\bigl(c(X^n)\bigr)\bigr] = \sum_{i=1}^{d^n} p(b_i)\lfloor\log_2(i+1)\rfloor \leq \sum_i p(b_i)\log_2\bigl(p(b_i)^{-1}+1\bigr) \leq 1 - \sum_i p(b_i)\log_2\bigl(p(b_i)\bigr) = H(p) + 1.$$

On the other hand,

$$\mathbb{E}\big[\ell\big(c(X^n)\big)\big] \geq \sum_i p(b_i)\big[\log_2(i+1)-1\big]$$
$$\geq -\alpha^{-1}\sum_i p(b_i)\log_2\left(\frac{\alpha-1}{(i+1)^\alpha}\right) - 1 + \frac{\log_2(\alpha-1)}{\alpha}.$$

Since $\sum_i \frac{\alpha-1}{(i+1)^\alpha} \leq (\alpha-1)\int_1^{+\infty}\frac{dz}{z^\alpha} = 1$, there is a probability measure $q \in \mathcal{M}_+^1(D^n)$ such that $q(b_i) \geq \frac{\alpha-1}{(i+1)^\alpha}$. We can then remark that

$$-\sum_i p(b_i)\log_2\left(\frac{\alpha-1}{(i+1)^\alpha}\right) - H(p) \geq \log(2)^{-1}\mathcal{K}(p,q) \geq 0.$$

This proves the second inequality of the proposition. The first one is obtained by choosing $\alpha = \big(1-n^{-1}\big)^{-1}$. □

1.3. Instantaneous codes. When transmitting $c(X^n)$ through a channel, it is useful for the receiver to know without delay when the message ends. More precisely, if $c(A^n) = W$, the receiver would like to dispose of a function $f : \{0,1\}^* \to \{0,1\}$ such that $\mathbb{1}\big(i = \ell(w)\big) = f\big(w(1),\dots,w(i)\big)$. If such a function exists, it is necessarily equal to $\mathbb{1}\big(w(1,\dots,i) \in W\big)$, so that in this case

$$\mathbb{1}\big(i = \ell(w)\big) = \mathbb{1}\big(w(1,\dots,i) \in W\big), \quad w \in W, 1 \leq i \leq \ell(w),$$

which can also be written as

$$w(1,\dots,i) \notin W, \qquad w \in W, 1 \leq i < \ell(w).$$

A code having this property is called an instantaneous code, and the previous reasoning proves

Proposition 1.7 *The binary code $c : A^n \to \{0,1\}^*$ is an instantaneous code if and only if it is a prefix code, which means that no codeword is the prefix of another codeword. In this case, the range $c(A^n) = W \subset \{0,1\}^*$ of c is called a prefix set.*

Proposition 1.8 (Kraft inequality) *For any prefix set $W \subset \{0,1\}^*$,*

$$\sum_{w\in W} 2^{-\ell(w)} \leq 1.$$

Proof. Consider $a(w) = \sum_{i=1}^{\ell(w)} w(i)\, 2^{-i}$ and $b(w) = a(w) + 2^{-\ell(w)}$. Let us write $w \prec w'$ when $w(1,\dots,\ell(w)) = w'(1,\dots,\ell(w))$, that is when w is a prefix of w'. To each finite binary word $w \in \{0,1\}^*$ corresponds an interval $I(w) = [a(w), b(w)[\subset [0,1[$. It is easy to see that $I(w') \subset I(w)$ when $w \prec w'$ and that $I(w') \cap I(w) = \varnothing$ otherwise. Thus W is a prefix set if and only if the intervals $I(w)$ are disjoint. In this case, considering the Lebesgue measure λ, we see that $1 = \lambda([0,1[) \geq \sum_{w\in W} \lambda\big(I(w)\big) = \sum_{w\in W} 2^{-\ell(w)}$. □

Proposition 1.9 (Inverse Kraft inequality) *If*

$$\sum_{i=1}^{T} 2^{-r_i} \leq 1, \quad (r_i, i = 1,\dots,T) \in \mathbb{N}^T,$$

then there is a prefix set $\{w_1,\dots,w_T\}$ *such that* $\ell(w_i) = r_i$.

Proof. Let us sort r_i so that $r_{i-1} \leq r_i$, $i = 2,\dots,T$. Let

$$\alpha_i = \sum_{j=1}^{i-1} 2^{-r_j} = \sum_{k=1}^{r_i} w_i(k)\, 2^{-k}, \text{ where } w_i \in \{0,1\}^{r_i}$$

is the binary representation of α_i up to precision 2^{-r_i}. As $a(w_i) = \alpha_i$ and $b(w_i) = \alpha_{i+1}$, it is clear that $I(w_i) \cap I(w_j) = \varnothing$, for any $i \neq j$, proving that $\{w_i, i = 1,\dots,T\}$ is indeed a prefix set. □

Definition 1.5 (Complete prefix code and prefix set) *A prefix set W is said to be complete if no other prefix set W' is such that $W \subsetneq W'$. A prefix code is complete if it uses a complete prefix set of codewords.*

Proposition 1.10 (Kraft equality) *A prefix set $W \subset \{0,1\}^*$ is complete if and only if*

$$\sum_{w\in W} 2^{-\ell(w)} = 1.$$

Proof. We are going to show that W is *not* complete if and only if $\sum_{w\in W} 2^{-\ell(w)} < 1$. If W is not complete, $W \subsetneq W'$, where W' is another prefix set, so that $\sum_{w\in W} 2^{-\ell(w)} < \sum_{w\in W'} 2^{-\ell(w)} \leq 1$, and $\sum_{w\in W} 2^{-\ell(w)} < 1$. On the other hand, if

$$\sum_{w\in W} 2^{-\ell(w)} < 1, \text{ then } \bigcup_{w\in W} I(w) \subsetneq [0,1[.$$

Let $L = \max\{\ell(w), w \in W\}$ and consider the partition of $[0,1[$ made by the intervals $\big\{J(k) = \big[k2^{-L}, (k+1)2^{-L}\big[,\ k = 0, \dots, 2^L - 1\big\}$. Since for any $w \in W$, $I(w) = \bigcup_{k \in K(w)} J(k)$, we see that $\bigcup_{w \in W} K(w) \subsetneq \{0, \dots, 2^L - 1\}$. Therefore, there is $k \in \{0, \dots, 2^L - 1\}$ such that $J(k) \notin \bigcup_{w \in W} I(w)$. Define $w' \in \{0,1\}^L$ by the formula $k2^{-L} = \sum_{i=1}^{L} w'(i)2^{-i}$. Since $I(w') = J(k)$, $W \cup \{w'\}$ forms a prefix set, hence W is not complete. □

Proposition 1.11 *For any prefix code c, and any random source $X^n \in A^n$,*

$$\mathbb{E}\big[\ell\big(c(X^n)\big)\big] \geq H(\mathbb{P}_{X^n}).$$

Moreover, there exists a prefix code c such that

$$\mathbb{E}\big[\ell\big(c(X^n)\big)\big] < H(\mathbb{P}_{X^n}) + 1.$$

Proof. Due to the Kraft inequality, we can find a probability measure $Q \in \mathcal{M}_+^1(A^n)$ such that $Q(x^n) \geq 2^{-\ell(c(x^n))}$, for any $x^n \in A^n$. We can then remark that

$$\mathbb{E}\big[\ell\big(c(X^n)\big)\big] - H(\mathbb{P}_{X^n}) \geq \mathcal{K}(P, Q)/\log(2) \geq 0,$$

as proved in Proposition 1.2 (page 3). As for the second part of the proposition, let us put

$$r(x^n) = \Big\lceil -\log_2\big[\mathbb{P}_{X^n}(x^n)\big] \Big\rceil.$$

Since $\sum_{x^n \in A^n} 2^{-r(x^n)} \leq 1$, there is, according to Proposition 1.9 (page 7), a prefix code c such that $\ell\big(c(x^n)\big) = r(x^n) < -\log_2\big(\mathbb{P}_{X^n}(x^n)\big) + 1$, for any $x^n \in A^n$, and consequently such that $\mathbb{E}\big[\ell\big(c(X^n)\big)\big] < H(\mathbb{P}_{X^n}) + 1$. □

Proposition 1.12 (Huffman code) *Let $p \in \mathcal{M}_+^1\big(\{1, \dots, d\}\big)$ be a probability measure on the d first integers. Let i and $j \leq d$ be such that $p(i)$ and $p(j)$ are the smallest, meaning that*

$$\max\{p(i), p(j)\} \leq \min\{p(k), k \neq i, k \neq j\}.$$

Let c be some optimal prefix code for the probability vector

$$\big(p(k), k \notin \{i, j\}, p(i) + p(j)\big)$$

of length $d-1$, that is some prefix code with minimum mean code length. Let $c' : \{1, \dots, d\} \to \{0,1\}^*$ be the prefix binary code defined as

$$c'(k) = \begin{cases} c(k), & k \notin \{i, j\}, \\ \big(c(i,j), 0\big), & k = i, \\ \big(c(i,j), 1\big), & k = j. \end{cases}$$

Then, c' is an optimal code for the source with probability p.

Proof. Let $c'' : \{1,\dots,d\} \to \{0,1\}^*$ be some optimal prefix code and let $W = c''(\{1,\dots,d\})$. Let us choose $w \in W$ such that $\ell(w) = \max\big\{\ell(v), v \in W\big\}$. Let $w' = \big(w(1,\dots,\ell(w)-1), 1 - w(\ell(w))\big)$ be its brother. If $w' \notin W$, then $w(1,\dots,\ell(w)-1)$ would be the prefix of w only, so that it would be possible to replace w with it and to decrease the mean code length of c'', leading to a contradiction. Thus necessarily $w' \in W$ also. We can assume that $c''(i) = w$ and $c''(j) = w'$, indeed, if this is not the case, we can exchange w and $c''(i)$ and w' and $c''(j)$ without increasing the mean length of c''. We can then consider the prefix code c''' on $\{k \notin \{i,j\}, 1 \le k \le d\} \cup \{(i,j)\}$ defined as

$$c'''(k) = \begin{cases} c''(k), & k \notin \{i,j\}, \\ w\big(1,\dots,\ell(w)-1\big), & k = (i,j). \end{cases}$$

It is easy to see that

$$\begin{aligned} \mathbb{E}\big[\ell(c'')\big] &= \mathbb{E}\big[\ell(c''')\big] + p(i) + p(j), \\ \mathbb{E}\big[\ell(c')\big] &= \mathbb{E}\big[\ell(c)\big] + p(i) + p(j). \end{aligned}$$

Since c is optimal, $\mathbb{E}\big[\ell(c)\big] \le \mathbb{E}\big[\ell(c''')\big]$, showing that $\mathbb{E}\big[\ell(c')\big] \le \mathbb{E}\big[\ell(c'')\big]$, so that c' is also optimal. □

Exercise 3 *Find an optimal code for the probability vector* $(1/3, 1/3, 1/6, 1/6)$.

1.4. Arithmetic codes. The previous section shows that for prefix codes as well, finding an optimal code requires to sort A^n according to decreasing probabilities. If n is large, this may induce intensive computations. Arithmetic codes are almost optimal and do not share this weakness.

Definition 1.6 *A non ordered vector of probabilities* $p(1,\dots,d)$ *being given, let us define*

$$\xi_i = \sum_{j=1}^{i-1} p(j), \quad i = 1,\dots,d+1.$$

Let us then consider the intervals $J(i) = \big[\xi_i, \xi_{i+1}\big[$ *and define*

$$w_i \in \arg\max_w \big\{\lambda\big(I(w)\big); w \in \{0,1\}^*, I(w) \subset J(i)\big\}.$$

The code $c(i) = w_i$ *is called an arithmetic, or Shannon-Fano-Elias code, and satisfies* $\mathbb{E}[\ell(c)] < H(p) + 2$.

Proof. The code c is prefix because the intervals $J(i)$ being non overlapping, this is also the case for $I(w)$, $w \in c\big(\{1,\dots,d\}\big)$. Let us define $L = \big\lceil 1 - \log_2\big[p(i)\big]\big\rceil$ and let

$$w = \arg\min_w \big\{a(w), w \in \{0,1\}^*, \ell(w) = L, a(w) \geq \xi_i\big\}.$$

Since

$$b(w) = a(w) + 2^{-L} < \xi_i + 2^{-(L-1)} \leq \xi_i + p(i) = \xi_{i+1},$$

$I(w) \subset J(i)$ and therefore $\ell\big[c(i)\big] \leq \ell(w) < 2 - \log_2\big(p(i)\big)$, implying that $\mathbb{E}[\ell(c)] < H(p) + 2$. □ Arithmetic coding of $x^n \in A^n$ is fast when A^n is sorted in lexicographic order. Indeed,

$$\begin{aligned}\xi(x^n) &= \sum_{k=1}^{n} \sum_{y<x_k} p(x^{k-1}, y) \\ &= \sum_{k=1}^{n} \sum_{y<x_k} \prod_{j=1}^{k-1} p(x_j \,|\, x^{j-1}) p(y \,|\, x^{k-1}),\end{aligned}$$

which can be computed from at most dn conditional probabilities performing at most dn additions and dn multiplications.

Exercise 4 *Write a program that draws a probability vector $p \in \mathcal{M}_+^1(\{1,\dots,d\})$ at random according to the uniform probability measure on the simplex. (Hint: draw $(X_1,\dots,X_{d-1})$ i.i.d. according to the uniform measure on the interval $[0,1]$, set $X_0 = 0$, $X_d = 1$, consider the order statistics $X_{(i)}$ such that $\{X_i\} = \{X_{(i)}\}$ and $X_{(0)} \leq \cdots \leq X_{(d)}$ and put $p_i = X_{(i)} - X_{(i-1)}$.) Compute an optimal prefix code for p (following Huffman's algorithm). Compute also an arithmetic prefix code. Print the mean code lengths of the two codes, as well as the Shannon entropy of p, for various draws of p and various values of d.*

2. Universal codes and Bayesian model averaging

We have seen that prefix codes were closely related to probability distributions, due to the Kraft inequality. From now on, we will identify the two, and consider any probability measure on A^n as an *ideal code*. In the previous section, we described some ways of coding efficiently a source X^n whose distribution was known. Here, we are going to study the case of an unknown distribution.

Definition 2.1 *The performance of an ideal code $Q \in \mathcal{M}_+^1(A^n)$ applied to a block X^n of length n of a source with probability law $\mathbb{P}_{X^n}$, is measured, in bits, by its redundancy*

$$\mathcal{R}\big(\mathbb{P}_{X^n}, Q\big) = \log(2)^{-1}\mathcal{K}\big(\mathbb{P}_{X^n}, Q\big) = \mathbb{E}_{X^n}\Big\{-\log_2\big[Q(X^n)\big]\Big\} - H\big(\mathbb{P}_{X^n}\big)$$

More precisely, the redundancy measures the difference between the mean length of a prefix code built according to Q and the optimal code built using $\mathbb{P}_{X^n}$, up to the discretization errors of 1 (Huffman) or 2 (Shannon-Fano-Elias) bits introduced by the actual coding algorithms.

2.1. Bayes redundancy.

One important approach to choose Q when $\mathbb{P}_{X^n}$ is unknown is provided by the Bayesian approach.

Definition 2.2 *Let π be some probability measure on some measurable set of parameters Θ. Let $\big\{P_\theta \in \mathcal{M}_+^1(A^n); \theta \in \Theta\big\}$ be a family of measures on A^n, where A is some finite set (extensions to more general measurable sets being possible). Let us assume that $\theta \mapsto P_\theta(x^n)$ is measurable for any $x^n \in A^n$. The mean Bayesian redundancy with respect to π is defined as*

$$\mathcal{R}_\pi(Q) = \int \mathcal{R}\big(P_\theta, Q\big)\,\mathrm{d}\pi(\theta).$$

Thus, the Bayesian redundancy measures the average loss of efficiency of the coding distribution Q, when the source distribution is drawn from $\{P_\theta; \theta \in \Theta\}$ according to the "prior" probability measure π.

Proposition 2.1 *The infimum $\inf_{Q\in\mathcal{M}_+^1(A^n)} \mathcal{R}_\pi(Q)$ is reached for only one coding distribution Q, called the bayesian mixture ideal code, and defined as*

$$Q = \int P_\theta\,\mathrm{d}\pi(\theta) \stackrel{\text{def}}{=} P_\pi.$$

Proof. The Bayesian redundancy is composed of

$$\begin{aligned}\mathcal{R}_\pi(Q) &= \int \mathcal{R}\big(P_\theta, P_\pi\big)\,\mathrm{d}\pi(\theta) + \mathcal{R}\big(P_\pi, Q\big)\\ &= \mathcal{R}_\pi\big(P_\pi\big) + \mathcal{R}\big(P_\pi, Q\big).\end{aligned}$$

The conclusion comes from the fact that $\mathcal{R}(P,Q) > 0$ for $P \neq Q$ and $\mathcal{R}(P,P) = 0$, because it is proportional to the Kullback Leibler divergence. □

Definition 2.3 *The mutual information of the joint distribution of two random variables X and Y defined on the same probability space is*

$$\begin{aligned}\mathcal{I}\big(\mathbb{P}_{X,Y}\big) &= \log(2)^{-1}\mathcal{K}\big(\mathbb{P}_{X,Y}, \mathbb{P}_X \otimes \mathbb{P}_Y\big)\\ &= \log(2)^{-1}\mathbb{E}_X\big[\mathcal{K}(\mathbb{P}_{Y|X}, \mathbb{P}_Y)\big].\end{aligned}$$

Proposition 2.2 *The optimal Bayesian redundancy, achieved by the Bayesian ideal code, is equal to the mutual information between the parameter θ and the source $(X_1, \ldots, X_n)$, when their joint distribution is defined as*

$$\mathbb{P}_{(\theta, X^n)}(B) = \int \mathbb{1}\big[(\theta, x^n) \in B\big]\mathrm{d}\pi(\theta)\mathrm{d}P_\theta(x^n),$$

for any measurable set B. In other words,

$$\mathcal{R}_\pi(P_\pi) = \mathcal{I}\big(\mathbb{P}_{\theta, X^n}\big).$$

Proof. This is a straightforward consequence of the definitions. The mutual information between X and Y tells how many bits can be saved on the transmission of Y when X is known both to the sender and the receiver of the message, as is clear from the second expression of the mutual information. In the case of the Bayesian redundancy, we see that the average increase of code length due to the fact that the parameter of the source has to be learnt, depends on how much information about the unknown parameter θ is contained in the source X^n. □

2.2. Minimax redundancy. In this approach we want to make sure to do our best in the worst case. In the setting of Definition 2.2 (page 11), let us define the worst case redundancy as

$$\mathcal{R}(Q) = \sup_{\theta \in \Theta} \mathcal{R}\big(P_\theta, Q\big).$$

Definition 2.4 *The ideal code $\widehat{Q} \in \mathcal{M}_+^1(A^n)$ is said to be a minimax coding distribution when*

$$\mathcal{R}(\widehat{Q}) = \inf_{Q \in \mathcal{M}_+^1(A^n)} \mathcal{R}(Q)$$

Proposition 2.3 (Least favorable prior) *Under the same hypotheses as in Definition 2.2 (page 11),*

$$\operatorname*{ess\,inf}_{\mathrm{d}\pi(\theta)} \mathcal{R}\big(P_\theta, P_\pi\big) \leq \int \mathcal{R}\big(P_\theta, P_\pi\big)\, \mathrm{d}\pi(\theta) \leq \inf_Q \sup_\theta \mathcal{R}\big(P_\theta, Q\big) \leq \sup_{\theta \in \Theta} \mathcal{R}\big(P_\theta, P_\pi\big).$$

If moreover, for some $\widehat{\pi} \in \mathcal{M}_+^1(\Theta)$, $\mathcal{R}\big(P_\theta, P_{\widehat{\pi}}\big) = \sup_\theta \mathcal{R}\big(P_\theta, P_{\widehat{\pi}}\big)$, $\widehat{\pi}$ almost surely, then $\widehat{\pi}$ is called a least favorable prior. In this case $P_{\widehat{\pi}}$ is the unique minimax coding distribution. Moreover $\widehat{\pi}$ is solution of

$$\mathcal{R}_{\widehat{\pi}}\big(P_{\widehat{\pi}}\big) = \sup_{\pi \in \mathcal{M}_+^1(\Theta)} \mathcal{R}_\pi\big(P_\pi\big). \tag{2}$$

On the other hand, if $\widehat{\pi}$ satisfies equation (2), then $P_{\widehat{\pi}}$ is the unique minimax coding distribution.

Remark 2.1 *Equation (2) does have a solution under mild assumptions, we refer to [1][theorem 1.2.1 page 17] for further details on this question.*

Proof. The chain of inequalities at the beginning of the proposition is a consequence of Proposition 3.1, saying that

$$\int \mathcal{R}(P_\theta, P_\pi) \mathrm{d}\pi(\theta) = \inf_{Q \in \mathcal{M}_+^!(\mathcal{X})} \int \mathcal{R}(P_\theta, Q) \mathrm{d}\pi(\theta).$$

In the case when $\mathcal{R}(P_\theta, P_{\widehat{\pi}}) = \sup_\theta \mathcal{R}(P_\theta, P_{\widehat{\pi}})$ almost surely, all the inequalities become equalities, showing that $P_{\widehat{\pi}}$ is a minimax coding distribution. To show that it is unique, let us consider another minimax distribution $\widehat{Q}$. It would satisfy necessarily

$$\sup_{\theta \in \Theta} \mathcal{R}\big(P_\theta, \widehat{Q}\big) = \int \mathcal{R}(P_\theta, P_{\widehat{\pi}}) \mathrm{d}\widehat{\pi}(\theta),$$

and therefore

$$\int \mathcal{R}(P_\theta, \widehat{Q})\, \mathrm{d}\widehat{\pi}(\theta) \leq \int \mathcal{R}(P_\theta, P_{\widehat{\pi}})\, \mathrm{d}\widehat{\pi}(\theta).$$

On the other hand

$$\int \mathcal{R}(P_{\widehat{\theta}}, \widehat{Q})\, \mathrm{d}\widehat{\pi}(\theta) = \int \mathcal{R}(P_\theta, P_{\widehat{\pi}})\, \mathrm{d}\widehat{\pi}(\theta) + \mathcal{R}(P_{\widehat{\pi}}, \widehat{Q}),$$

implying that $\mathcal{R}(P_{\widehat{\pi}}, \widehat{Q}) = 0$, and consequently that $\widehat{Q} = P_{\widehat{\pi}}$. Finally, equation (2) is a consequence of the equality

$$\int \mathcal{R}(P_\theta, P_{\widehat{\pi}}) \mathrm{d}\widehat{\pi}(\theta) = \inf_Q \sup_\theta \mathcal{R}(P_\theta, Q) \geq \int \mathcal{R}(P_\theta, P_\pi) \, \mathrm{d}\pi(\theta).$$

On the other hand, let us now assume that $\widehat{\pi}$ is solution of equation (2). In this case, we can write

$$\mathcal{R}_{\widehat{\pi}}(P_{\widehat{\pi}}) - \mathcal{R}_\pi(P_\pi) = \int \mathcal{R}(P_\theta, P_{\widehat{\pi}}) \, (\mathrm{d}\widehat{\pi} - \mathrm{d}\pi)(\theta) + \mathcal{R}(P_\pi, P_{\widehat{\pi}}).$$

Applying this inequality to $\pi = \lambda\nu + (1-\lambda)\widehat{\pi}$, for all $\lambda \in [0,1]$, and $\nu \in \mathcal{M}_+^1(\Theta)$, we see that the right-hand side should have a positive derivative at $\lambda = 0$, showing that

$$\int \mathcal{R}(P_\theta, P_{\widehat{\pi}}) (\mathrm{d}\widehat{\pi} - \mathrm{d}\nu)(\theta) \geq 0.$$

Considering for any $\theta \in \Theta$, $\nu = \delta_\theta$, the Dirac mass at θ, we deduce that

$$\int \mathcal{R}(P_\theta, P_{\widehat{\pi}}) \, \mathrm{d}\widehat{\pi}(\theta) = \sup_{\theta \in \Theta} \mathcal{R}(P_\theta, P_{\widehat{\pi}}),$$

showing from the first part of the proposition that $P_{\widehat{\pi}}$ is indeed the minimax coding distribution in this case. □

2.3. Oracle inequalities. Another point of view is to consider a family of coding distributions, $\{Q_\theta : \theta \in \Theta\}$, instead of considering a familiy of source distributions. Given some prior probability measure $\pi \in \mathcal{M}_+^1(\Theta)$, it is easy to compare the mean length of the code mixture

$$Q_\pi \stackrel{\text{def}}{=} \int Q_\theta \, \mathrm{d}\pi(\theta),$$

with the mean code length of any Q_θ, $\theta \in \Theta$. Indeed,

Proposition 2.4 *For any* $(x_1, \ldots, x_n) \in A^n$,

$$\begin{aligned} -\log_2\bigl[Q_\pi(x^n)\bigr] &\leq \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \int -\log_2\bigl[Q_\theta(x^n)\bigr] \mathrm{d}\rho(\theta) + \mathcal{R}(\rho, \pi) \\ &\leq \inf_{\theta \in \Theta} -\log_2\bigl[Q_\theta(x^n)\bigr] - \log_2[\pi(\{\theta\})]. \end{aligned}$$

Consequently, for any source distribution $P \in \mathcal{M}_+^1(A^n)$,

$$\mathcal{R}\big(P, Q_\pi\big) \leq \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \int \mathcal{R}(P, Q_\theta) \mathrm{d}\rho(\theta) + \mathcal{R}(\rho, \pi)$$
$$\leq \inf_{\theta \in \Theta} \mathcal{R}\big(P, Q_\theta\big) - \log_2\big[\pi(\{\theta\})\big].$$

Proof. If $\mathcal{R}(\rho, \pi) = +\infty$, there is nothing to prove. Let us assume consequently that $\mathcal{R}(\rho, \pi) < \infty$. Let $\pi = \pi_s + \pi_a$ where $\pi_a \ll \rho$ and π_s and ρ are singular. It is easy to see that $\dfrac{\mathrm{d}\pi_a}{\mathrm{d}\rho}(\theta) = \dfrac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta)^{-1}$, ρ almost surely, therefore

$$\begin{aligned}
\int Q_\theta(x^n) \mathrm{d}\pi(\theta)\, \mathrm{d}\pi(\theta) &\geq \int Q_\theta(x^n)\, \mathrm{d}\pi_a(\theta) \\
&= \int \exp\left\{ \log\big[Q_\theta(x^n)\big] + \log\left[\frac{\mathrm{d}\pi_a}{\mathrm{d}\rho}(\theta)\right] \right\} \mathrm{d}\rho(\theta) \\
&= \int \exp\left\{ \log\big[Q_\theta(x^n)\big] - \log\left[\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta)\right] \right\} \mathrm{d}\rho(\theta) \\
&\geq \exp\left\{ \int \log\big[Q_\theta(x^n)\big]\, \mathrm{d}\rho(\theta) - \mathcal{K}(\rho, \pi) \right\},
\end{aligned}$$

by Jensen's inequality. □

Corollary 2.5 (Double mixture codes and Bayesian Model Averaging) *Consider a family of models of ideal codes $\big\{Q_\theta \in \mathcal{M}_+^1(A^n); \theta \in \Theta_k\big\}$, $k \in \mathbb{N}$, where Θ_k are measurable parameter spaces. Let $\mu \in \mathcal{M}_+^1(\mathbb{N})$ and for each $k \in \mathbb{N}$ let $\nu_k \in \mathcal{M}_+^1\big(\Theta_k\big)$ be some prior measure on Θ_k. Let us consider $\Theta = \bigcup_{k \in \mathbb{N}} \big(\{k\} \times \Theta_k\big)$. Let the prior distribution π on Θ be defined for any measurable set $B \subset \Theta$ by the formula*

$$\pi(B) = \int \mathbb{1}\big[(k, \theta) \in B\big]\, \mathrm{d}\mu(k)\, \mathrm{d}\nu_k(\theta).$$

The coding distribution Q_π is called a double mixture code. The sequential prediction method that estimates $\mathbb{P}_{X_i|X^{i-1}}$ by

$$Q_\pi\big(x_i|x^{i-1}\big) \stackrel{\text{def}}{=} Q_\pi(x^i)/Q_\pi(x^{i-1})$$

is called Bayesian Model Averaging. It is such that

$$\sum_{i=1}^n -\log_2\big[Q_\pi\big(x_i|x^{i-1}\big)\big] \leq \inf_{k \in \mathbb{N}} -\log_2 Q_{\nu_k}\big(x_i|x^{i-1}\big) - \log_2\big[\mu(k)\big].$$

Consequently for any source $X^n \in A^n$,

$$\sum_{i=1}^{n} \mathbb{E}\Big\{\mathcal{K}\big[\mathbb{P}_{X_i|X^{i-1}}, Q_\pi(\cdot|X^{i-1})\big]\Big\}$$
$$\leq \inf_{k\in\mathbb{N}}\Big\{\sum_{i=1}^{n} \mathbb{E}\Big\{\mathcal{K}\big[\mathbb{P}_{X_i|X^{i-1}}, Q_{\nu_k}(\cdot|X^{i-1})\big]\Big\} - \log\big[\mu(k)\big]\Big\}.$$

PROOF. It is a consequence of the previous proposition and of the identity

$$\sum_{i=1}^{n} -\log_2\big[Q_\pi\big(x_i \,|\, x^{i-1}\big)\big] = -\log_2\big[Q_\pi(x^n)\big].$$

Let us remark that we could have replaced in this proof the family $\{Q_{\nu_k}, k \in \mathbb{N}\}$ by any family of ideal codes $\{Q_k, k \in \mathbb{N}\}$. □

This is a result about sequential prediction. We consider estimators $Q_k(\cdot|X^{i-1})$ of the conditional distributions $\mathbb{P}_{X_i|X^{i-1}}$ and are provided with an aggregated estimator $Q_\pi(\cdot|X^{i-1})$ that performs almost as well as the best choice of k, when performance is measured by the sum of the expected Kullback divergences over all sample sizes in the range $1, \ldots, n$. The sequential nature of the result comes from the fact that the criterion is a sum over all sample sizes : this is an example of cumulated risk function. The following two subsections are devoted to an important example of double mixture codes : the context tree weighting algorithm, where we are going to compute everything explicitly.

2.4. MIXTURES OF I.I.D. DISTRIBUTIONS. Let $\Theta = \mathcal{M}_+^1(A)$ and consider the Lebesgue measure λ on Θ. Let the prior $\nu \ll \lambda$ be defined by its density

$$\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}(\theta) = \frac{\Gamma\big(\frac{d}{2}\big)}{\sqrt{d}\,\Gamma\big(\frac{1}{2}\big)^d} \prod_{x\in A} \theta(x)^{-1/2}, \tag{3}$$

where $\Gamma(z) = \displaystyle\int_{\mathbb{R}_+} t^{z-1}\exp(-t)\,\mathrm{d}t = (z-1)\Gamma(z-1)$ is the usual Γ function.

LEMMA 2.6 *The mixture code based on the prior ν and the family of product measures $Q_\theta(x^n) = \prod_{i=1}^n \theta(x_i)$ is such that*

$$Q_\nu(x^n) = \frac{\Gamma\big(\frac{d}{2}\big)\prod_{y\in A}\Gamma\big(n\overline{\mathbb{P}}_{x^n}(y)+\frac{1}{2}\big)}{\Gamma\big(\frac{1}{2}\big)^d\Gamma\big(n+\frac{d}{2}\big)} = \frac{\displaystyle\prod_{y\in A}\prod_{j=1}^{n\overline{\mathbb{P}}_{x^n}(y)}\big(j-\tfrac{1}{2}\big)}{\displaystyle\prod_{j=1}^{n}\big(j+\tfrac{d}{2}-1\big)},$$

where

$$\overline{\mathbb{P}}_{x^n} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$$

is the empirical probability measure of the sequence $(x_1, \ldots, x_n)$*, so that* $n\overline{\mathbb{P}}_{x^n}(y) = \sum_{i=1}^{n} \mathbb{1}(x_i = y)$.

PROOF. Let us remark that

$$Q_\theta(x^n) = \prod_{y\in A} \theta(y)^{n\overline{\mathbb{P}}_{x^n}(y)},$$

and thus

$$Q_\nu(x^n) = \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{d}\Gamma\left(\frac{1}{2}\right)^d} \int_\Theta \prod_{y\in A} \theta(y)^{n\overline{\mathbb{P}}_{x^n}(y)-\frac{1}{2}} \mathrm{d}\lambda(\theta). \tag{4}$$

On the other hand, for any $\alpha \in]-1, +\infty[^A$,

$$\int_\Theta \prod_{y\in A} \theta(y)^{\alpha_y}\, \mathrm{d}\lambda(\theta) = \frac{\sqrt{d}\prod_{y\in A}\Gamma(\alpha_y+1)}{\Gamma\left(\sum_{y\in A}\alpha_y + d\right)}. \tag{5}$$

Indeed, the change of variables $t_y = s\theta(y)$, where $t_y \in \mathbb{R}_+$, $\theta \in \mathcal{M}_+^1(A)$ and $s = \sum_{y\in A} t_y \in \mathbb{R}_+$ shows that

$$\int_{\mathbb{R}_+^d} \prod_{y\in A} t_y^{\alpha_i} \exp(-t_y)\, \mathrm{d}t_y = \int_{\mathbb{R}_+}\int_{\theta\in\Theta} \prod_{y\in A} \theta(y)^{\alpha_y} s^{\alpha_y} \exp(-s) s^{d-1} \mathrm{d}\lambda(\theta)\frac{\mathrm{d}s}{\sqrt{d}}.$$

Putting together equations (4) and (5) gives the desired result. □

PROPOSITION 2.7 (KRICHEVSKI, TROFIMOV) *For any* $x^n \in A^n$*,*

$$Q_\nu(x^n) \geq \frac{2^{-nH(\overline{\mathbb{P}}_{x^n})}}{dn^{(d-1)/2}} = d^{-1}n^{-(d-1)/2}\sup_{\theta\in\Theta} Q_\theta(x^n).$$

PROOF. Let us put for any $a = (a_i)_{i=1}^d \in \mathbb{N}^d$

$$\Delta(a) = \frac{\Gamma\left(\frac{d}{2}\right)\prod_{i=1}^d \Gamma\left(a_i+\frac{1}{2}\right)\left(\sum_{i=1}^d a_i\right)^{\sum_{i=1}^d a_i + (d-1)/2}}{\Gamma\left(\frac{1}{2}\right)^d \Gamma\left(\sum_{i=1}^d a_i + \frac{d}{2}\right)\prod_{i=1}^d a_i^{a_i}}$$

We have to show that $\Delta(a) \geq d^{-1}$. Let us notice first that $\Delta\big((1,0,\ldots,0)\big) = d^{-1}$, and that Δ is invariant under any permutation of the vector $(a_i)_{i=1}^d$. It

is therefore enough to check that $\Delta(a) \geq \Delta\big((a_1 - 1, a_2, \ldots, a_d)\big)$. Let us put $s = \sum_{i=1}^{d} a_i$ and $t = a_1$. With these notations,

$$\Delta(a) = \Delta(a_1 - 1, a_2^d) \frac{\big(t - \frac{1}{2}\big)\big(t-1\big)^{t-1} s^{s+\frac{d-1}{2}}}{\big(s + \frac{d}{2} - 1\big) t^t \big(s-1\big)^{s-1+\frac{d-1}{2}}}.$$

To end the proof, we have to check that

$$\frac{\big(t - \frac{1}{2}\big)\big(t-1\big)^{t-1} s^{s+\frac{d-1}{2}}}{\big(s + \frac{d}{2} - 1\big) t^t \big(s-1\big)^{s-1+\frac{d-1}{2}}} \geq 1, \qquad t \geq 1, s \geq 2.$$

This can also be written as $g(s) \geq f(t)$, where

$$g(s) = -\big(s + \tfrac{d-3}{2}\big) \log\left(\frac{s-1}{s}\right) - \log\left(\frac{s + \frac{d-3}{2}}{s}\right),$$
$$f(t) = t \log(t) - (t-1)\log(t-1) - \log\big(t - \tfrac{1}{2}\big).$$

Using the fact that $\log(1+z) \leq z$, we see that $g(s) \geq \big(s + \frac{d-3}{2}\big) s^{-1} - \frac{d-3}{2s} = 1$. On the other hand, making a Taylor expansion of $z \mapsto z\log(z)$, we see that

$$(z+u)\log(z+u) = z\log(z) + u\big[\log(z) + 1\big] + \int_0^u (u-v) \frac{\mathrm{d}v}{z+v}.$$

Applying this to $z = t - \frac{1}{2}$ and $u \in \big\{-\frac{1}{2}, \frac{1}{2}\big\}$, gives

$$t\log(t) - (t-1)\log(t-1) - \log\big(t - \tfrac{1}{2}\big) - 1$$
$$= \int_0^{\frac{1}{2}} \big(\tfrac{1}{2} - v\big) \left(\frac{1}{t - \frac{1}{2} + v} - \frac{1}{t - \frac{1}{2} - v} \right) \mathrm{d}v \leq 0,$$

proving that $f(t) \leq 1$. □

Corollary 2.8 *The redundancy of the KT ideal code is such that*

$$\mathcal{R}(P, Q_\nu) \leq \inf_{\theta \in \Theta} \mathcal{R}(P, Q_\theta) + \frac{d-1}{2} \log_2(n) + \log_2(d).$$

2.5. Universal codes and context tree weighting. Let $\mathcal{D}$ be the set of complete suffix sets of A^* and $\mathcal{D}_L$ the subset of complete suffix sets of length not greater than L. More explicitly, let us introduce the notations $A^*w = \{w'w, w' \in A^*\}$ and

$$\overline{D} = D \cup \big\{w \in A^*; A^*w \cap D \neq \varnothing\big\},$$

the set of suffixes of D. We can define $\mathcal{D}_L$ formely as

$$\mathcal{D}_L = \big\{ D \subset A^*; D \cap A^* D = \varnothing, A^* \overline{D} = A^*, \text{ and } \max_{w \in D} \ell(w) \leq L \big\}.$$

Let $D \in \mathcal{D}_L$ be a complete suffix set and let $f_D : A^{\mathbb{Z}_-} \to D$ be defined by

$$f_D\big(x_{-\infty}^0\big) = x_{-k}^0,$$

where k is the only index such that $x_{-k}^0 \in D$. Given a past context $x_{-\infty}^0$, we want to define a probability measure on $x_1^{+\infty}$. A stationary context tree distribution is a probability measure defined by

$$\mathbb{P}\big(X_n = x_n \,|\, X_{-\infty}^{n-1} = x_{-\infty}^{n-1}\big) = \mathbb{P}\big[X_1 = x_n \,|\, f_D(X_{-\infty}^0) = f_D(x_{-\infty}^{n-1})\big].$$

Its parameter set is $\big(D, \Theta_D\big)$ where $\Theta_D = \big\{\theta_w \in \mathcal{M}_+^1(A); w \in D\big\} = \mathcal{M}_+^1(A)^D$. Consider the parameter space

$$\Theta = \bigcup_{D \in \mathcal{D}_L} \{D\} \times \Theta_D.$$

Let us define on Θ the prior probability measure π by

$$\pi(B) = \int \mathbb{1}\big[(D, \theta) \in B\big] \, \mathrm{d}\mu(D) \prod_{w \in D} \mathrm{d}\nu(\theta_w),$$

where ν is the Krichevski Trofimov prior on $\mathcal{M}_+^1(A)$ and where for some real parameter $\alpha \in]0, 1[$

$$\mu(D) = \alpha^{(|D|-1)/(d-1)} (1-\alpha)^{|D \setminus A^L|}, \qquad D \in \mathcal{D}_L, \tag{6}$$

is the measure of a Galton Watson process on complete suffix sets with offspring probability α. More precisely, the distribution of $\overline{D}$ under μ is given by the formula

$$\begin{aligned}
\mu(\overline{D} = \overline{B}) &= \prod_{j=1}^{L} \mu\big(\overline{D} \cap A^j = \overline{B} \cap A^j \,|\, \overline{D} \cap \overline{A^{j-1}} = \overline{B} \cap \overline{A^{j-1}}\big) \\
&\stackrel{\text{def}}{=} \prod_{j=1}^{L} \prod_{w \in \overline{B} \cap A^{j-1}} \big[\alpha \mathbb{1}(A^* w \subset \overline{B}\big) + (1-\alpha) \mathbb{1}\big(A^* w \cap \overline{B} = \varnothing\big)\big] \\
&= \alpha^{|\overline{B} \setminus B| + 1} (1-\alpha)^{|B \setminus A^L|}, \qquad B \in \mathcal{D}_L.
\end{aligned}$$

Equation (6) is a consequence of the fact that

$$|\overline{B} \setminus B| = \big(|B| - 1\big)/(d-1) - 1,$$

for any $B \in \mathcal{D}_L$, a fact that can easily be checked by induction.

Proposition 2.9 *Let us be given some sequence $x_{1-L}^n \in A^{n+L}$. Let us consider the counters*

$$\begin{aligned} a_w^y(n) &= \sum_{k=1}^{n} \mathbb{1}\big(x_{k-\ell(w)}^{k-1} = w, x_k = y\big), \qquad w \in A^* \cup \{\varnothing\}, y \in A, \\ b_w(n) &= \sum_{y \in A} a_w^y(n), \\ K_w(n) &= \frac{\Gamma\big(\frac{d}{2}\big) \prod_{y \in A} \Gamma\big(a_w^y(n) + \frac{1}{2}\big)}{\Gamma\big(\frac{1}{2}\big)^d \Gamma\big(b_w(n) + \frac{d}{2}\big)}, \end{aligned}$$

The mixture code Q_π can be computed as

$$Q_\pi\big(x_1^n \,|\, x_{1-L}^0\big) = \sum_{D \in \mathcal{D}_L} \mu(D) \prod_{w \in D} K_w(n).$$

This computation can be made by induction according to the following scheme

$$\begin{aligned} p_w(n) &= K_w(n), && w \in A^L, \\ p_w(n) &= (1-\alpha) K_w(n) + \alpha \prod_{y \in A} p_{yw}(n), && w \in A^* \cup \{\varnothing\}, \ell(w) < L, \\ p_\varnothing(n) &= Q_\pi\big(x_1^n \,|\, x_{1-L}^0\big). \end{aligned}$$

Proof. For any $D \in \mathcal{D}_L$,

$$\begin{aligned} &\int_{\Theta_D} Q_{D,\theta}\big(x_1^n \,|\, x_{1-L}^0\big) \prod_{w \in D} \mathrm{d}\nu(\theta_w) \\ &\qquad = \int_{\Theta_D} \prod_{w \in D} \prod_{y \in A} \theta_w(y)^{a_w^y(n)} \prod_{w \in D} \mathrm{d}\nu(\theta_w) = \prod_{w \in D} K_w(n), \end{aligned}$$

so that

$$Q_\pi\big(x_1^n \,|\, x_{1-L}^0\big) = \int_{\mathcal{D}_L} \prod_{w \in D} K_w(n) \, \mathrm{d}\mu(D).$$

Let

$$p_w(n) = \int_{\mathcal{D}_L} \prod_{w' \in D \cap (A^* w \cup \{w\})} K_{w'}(n) \, \mathrm{d}\mu\big(D \,|\, w \in \overline{D} \cup \{\varnothing\}\big),$$

so that $p_\varnothing(n) = Q_\pi\big(x_1^n \,|\, x_{1-L}^0\big)$. We see from the definition of μ that

$$p_w(n) = \mu\big(w \in D \,|\, w \in \overline{D}\big) K_w(n)$$

$$+\mu\big(A^*w\subset\overline{D}\,|\,w\in\overline{D}\big)\prod_{y\in A}\int_{\mathcal{D}_L}\prod_{w'\in D\cap(A^*yw\cup\{yw\})}K_{w'}(n)\mathrm{d}\mu\big(D\,|\,yw\in\overline{D}\big)$$
$$=(1-\alpha)K_w(n)+\alpha\prod_{y\in A}p_{yw}(n),\qquad w\in A^*\cup\{\varnothing\},\ell(w)<L,$$

and that $p_w(n)=K_w(n)$ for any $w\in A^L$. □

Exercise 5 *Show that the computation of the context tree weighting ideal code can be updated online according to the following rules :*

$$\begin{aligned}
a_w^y(0)&=b_w(y)=0,\\
K_w(0)&=1,\\
p_w(0)&=1,\\
&\vdots\\
a_w^y(n)&=\begin{cases}a_w^y(n-1)+1, & \text{when } w=x_{n-\ell(w)}^{n-1} \text{ and } y=x_n,\\ a_w^y(n-1), & \text{otherwise,}\end{cases}\\
b_w(n)&=\begin{cases}b_w(n-1)+1, & \text{when } w=x_{n-\ell(m)}^{n-1},\\ b_w(n-1), & \text{otherwise,}\end{cases}\\
K_w(n)&=\begin{cases}K_w(n-1)\dfrac{a_w^{x_n}(n)-\frac12}{b_w(n)+\frac{d}{2}-1}, & \text{when } w=x_{n-\ell(m)}^{n-1},\\ K_w(n-1), & \text{otherwise,}\end{cases}\\
p_w(n)&=\begin{cases}p_w(n-1), & \text{when } w\neq x_{n-\ell(m)}^{n-1},\\ K_w(n), & \text{when } w=x_{n-L}^{n-1},\\ (1-\alpha)K_w(n)+\alpha\displaystyle\prod_{y\in A}p_{yw}(n), & \text{when } w=x_{n-\ell}^{n-1},\ \text{with } \ell<L.\end{cases}
\end{aligned}$$

Conclude that the number of operations needed to compute $Q_\pi(x_1^n\,|\,x_{1-L}^0)$ for a given value of $x_{1-L}^n\in A^{n+L}$ grows only linearly with n.

Proposition 2.10 *For any $x_{1-L}^n\in A^{n+L}$,*

$$\begin{aligned}
-\log_2\big[Q_\pi\big(x_1^n\,|\,x_{1-L}^0\big)\big]\leq\inf_{D\in\mathcal{D}_L}\inf_{\theta\in\Theta_D}\Big\{&-\log_2\big[Q_{D,\theta}\big(x_1^n\,|\,x_{1-L}^0\big)\big]\\
&+\tfrac{|D|(d-1)}{2}\log\big(\tfrac{n}{|D|}\big)+|D|\log_2(d)\\
&-\tfrac{|D|-1}{d-1}\log_2(\alpha)-|D|\log_2(1-\alpha)\Big\}.
\end{aligned}$$

As a consequence for any process X_{1-L}^n, and any $x_{1-L}^0\in A^L$,

$$\begin{aligned}\mathcal{R}\big[\mathbb{P}_{X_1^n|X_{1-L}^0=x_{1-L}^0}, Q_\pi(\cdot|x_{1-L}^0)\big] \\ \leq \inf_{D\in\mathcal{D}_L}\inf_{\theta\in\Theta_D}\Big\{\mathcal{R}\big[\mathbb{P}_{X_1^n|X_{1-L}^0=x_{1-L}^0}, Q_{D,\theta}(\cdot|x_{1-L}^0)\big] \\ + \tfrac{|D|(d-1)}{2}\log\big(\tfrac{n}{|D|}\big) + |D|\log_2(d) \\ - \tfrac{|D|-1}{d-1}\log_2(\alpha) - |D|\log_2(1-\alpha)\Big\}.\end{aligned}$$

Proof. Let $\gamma(z) = \frac{d-1}{2}\log_2(z) + \log_2(d)$. According to Proposition 2.7 (page 17),

$$\begin{aligned}-\log_2\bigg[\int_{\Theta_D} Q_{D,\theta}\big(x_1^n \,|\, x_{1-L}^0\big)\prod_{w\in D}\mathrm{d}\nu(\theta_w)\bigg] \\ \leq \inf_{\theta\in\Theta_D} -\log_2\big[Q_{D,\theta}\big(x_1^n \,|\, x_{1-L}^0\big)\big] + \sum_{w\in D}\gamma\big[b_w(n)\big] \\ \leq \inf_{\theta\in\Theta_D} -\log_2\big[Q_{D,\theta}\big(x_1^n \,|\, x_{1-L}^0\big)\big] + |D|\gamma\bigg(\frac{1}{|D|}\sum_{w\in D} b_w(n)\bigg),\end{aligned}$$

where we have used the fact that γ is concave. As $\sum_{w\in D} b_w(n) = n$, we get

$$\begin{aligned}-\log_2\big[Q_\pi\big(x_1^n \,|\, x_{1-L}^0\big)\big] \leq \inf_{D\in\mathcal{D}_L}\inf_{\theta\in\Theta_D} -\log_2\big[Q_{D,\theta}\big(x_1^n \,|\, x_{1-L}^0\big)\big] \\ + |D|\gamma\bigg(\frac{n}{|D|}\bigg) - \log_2\big[\mu(D)\big],\end{aligned}$$

as stated in the proposition. □

Exercise 6 (A simulation study of a switch code) *Let us consider the alphabet* $A = \{1,\dots,4\}^2$ *of size* 16. *Let us define* $q \in \mathcal{M}_+^1(A)$ *as*

$$q(i,j) = \frac{4(i-1)+j}{8\times 17}$$

and let $q_1(i) = \sum_{j=1}^4 q(i,j)$ *and* $q_2(j) = \sum_{i=1}^4 q(i,j)$ *be its two marginal distributions. Let*

$$p(i,j) = \frac{4}{5}q(i,j) + \frac{1}{5}q_1(i)q_2(j), \qquad i,j\in\{1,\dots,4\}.$$

For any $\theta_1,\theta_2 \in \Theta_1 \stackrel{\text{def}}{=} \mathcal{M}_+^1(\{1,\dots,4\})$, *let*

$$Q_{\theta_1,\theta_2}(x^n) = \prod_{i=1}^n \theta_1(x_{i,1})\theta_2(x_{i,2}), \qquad x^n\in A^n,$$

where $x_i = (x_{i,1}, x_{i,2}) \in \{1, \dots 4\}^2$. *In the same way, for any* $\theta \in \Theta_2 \stackrel{\text{def}}{=} \mathcal{M}_+^1(A)$, *let*

$$Q_\theta(x^n) = \prod_{i=1}^{n} \theta(x_i).$$

Let ν_1 *be the Krichevski Trofimov prior distribution on* Θ_1, *given by equation* (3, page 16), *and* ν_2 *the Krichevski Trofimov prior distribution on* Θ_2. *Let us consider the following ideal codes*

$$\begin{aligned}
Q_1(x^n) &= \int_{\Theta_1 \times \Theta_1} Q_{\theta_1,\theta_2}(x^n)\, \mathrm{d}\nu_1(\theta_1) \mathrm{d}\nu_1(\theta_2), \\
Q_2(x^n) &= \int_{\Theta_2} Q_\theta(x^n)\, \mathrm{d}\nu_2(\theta), \\
Q_m(x^n) &= \frac{1}{2} Q_1(x^n) + \frac{1}{2} Q_2(x^n), \\
Q_s(x^n) &= \begin{cases} Q_1(x^n), & \text{when } n < m, \\ Q_1(x^m) Q_2(x^n) / Q_2(x^m), & \text{when } n \geq m. \end{cases}
\end{aligned}$$

Take $m = 2000$, *and write a program to compute for* $n \in 200\mathbb{N}$, $n \leq 10000$ *the empirical mean ideal code length*

$$L(Q, n) = -\frac{1}{100} \sum_{j=1}^{100} \log_2 \big[Q\big(X_1^n(j)\big) \big],$$

where $X_1^n(j)$ *are independent trials of the i.i.d. random process distributed according to* $p^{\otimes n}$, *and where* $Q \in \{Q_1, Q_2, Q_m, Q_s\}$. *Plot the curves*

$$\begin{aligned}
n &\mapsto L(Q_2, n) - L(Q_1, n), \\
n &\mapsto L(Q_m, n) - L(Q_1, n), \\
n &\mapsto L(Q_s, n) - L(Q_1, n).
\end{aligned}$$

Make your comments (about the weak points and possible improvements of mixture codes).

Programming hint : all you need is to simulate the empirical measure of X_1^n, *something you can do in Octave, using the function*

`tabulate(rand(1,n), df)`,

where `df` *is the distribution function of your probability vector (to be computed with* `cumsum`*).*

3. PAC-Bayes bounds for supervised classification

3.1. Introduction. PAC-Bayes theory was first developed in the framework of supervised classification (see [6, 7, 8, 9, 5]) and subsequently extended to other settings. We will not deal with these extensions in the present notes, but instead focus on supervised classification, a setting that plays a central role in statistical learning theory and requires specific techniques of proofs.

In this section, we are given some i.i.d. sample $(W_i)_{i=1}^n \in \mathcal{W}^n$, where $\mathcal{W}$ is a measurable space, and some binary measurable loss function $L : \mathcal{W} \times \Theta \to \{0, 1\}$, where Θ is a measurable parameter space. Our aim is to minimize with respect to $\theta \in \Theta$ the expected loss

$$\int L(w, \theta) \, \mathrm{d}\mathbb{P}(w),$$

where $\mathbb{P}$ is the marginal distribution of the observed sample $(W_i)_{i=1}^n$. More precisely, assuming that $\mathbb{P}$ is unknown, we would like to find an estimator $\widehat{\theta}(W_1^n)$ depending on the observed sample W_1^n such that the excess risk

$$\int L(w, \widehat{\theta}) \, \mathrm{d}\mathbb{P}(w) - \inf_{\theta \in \Theta} \int L(w, \theta) \, \mathrm{d}\mathbb{P}(w)$$

is small. The previous quantity is random, since $\widehat{\theta}$ depends on the random sample W_1^n. Therefore, how small it is can be understood in different ways. Here we will focus on the *deviations* of the excess risk. Accordingly, we will look for estimators providing a small risk with a probability close to one.

A typical example of such a problem is provided by supervised classification. In this setting $\mathcal{W} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ is a finite set, $W_i = (X_i, Y_i)$, where (X_i, Y_i) are input-output pairs, a family of measurable classification rules $\big\{ f_\theta : \mathcal{X} \to \mathcal{Y}; \theta \in \Theta \big\}$ is considered and the loss function $L(w, \theta)$ is defined as the classification error

$$L\big[(x, y), \theta\big] = \mathbb{1}\big[f_\theta(x) \neq y\big].$$

Accordingly the aim is to minimize the expected classification error

$$\mathbb{P}_{X,Y}\big[f_\theta(X) \neq Y\big]$$

in view of a sample $(X_i, Y_i)_{i=1}^n$ of observations.

Let us remark that the point of view adopted in this section is different from the point of view of the previous section in some important ways. With the tools of the last section, we could have considered the loss function

$$L_c\big[(x, y), q_\theta\big] = -\log_2\big[q_\theta(y \,|\, x)\big],$$

where $q_\theta(y|x)$ is some family of conditional distributions indexed by θ.

We could then have obtained results concerned with

$$\sum_{i=1}^{n} L_c\left[(X_i, Y_i), \int q_\theta \, \mathrm{d}\rho_i(\theta)\right],$$

where $\rho_i \in \mathcal{M}_+^1(\Theta)$ is a random probability measure defined with the help of some prior probability measure $\pi \in \mathcal{M}_+^1(\Theta)$ by its density

$$\frac{\mathrm{d}\rho_i}{\mathrm{d}\pi}(\theta) = \frac{\displaystyle\prod_{i=1}^{n} q_\theta(Y_i|X_i)}{\displaystyle\int \prod_{i=1}^{n} q_{\theta'}(Y_i|X_i) \, \mathrm{d}\pi(\theta')}.$$

Namely, the previous section provides the following almost sure upper bound:

$$\sum_{i=1}^{n} L_c\left[(X_i, Y_i), \int q_\theta \, \mathrm{d}\rho_i(\theta)\right] \leq \inf_{\theta\in\Theta} \sum_{i=1}^{n} L_c\big[(X_i, Y_i), q_\theta\big] - \log_2\big[\pi\big(\{\theta\}\big)\big]. \quad (7)$$

Importantly, the previous section is concerned with a *cumulated* notion of risk, whereas this one will deal with an *instantaneous* notion of risk. For the cumulated risk, we could obtain almost sure results, this will not be possible any more for the instantaneous risk we are considering here. We will have instead to make an hypothesis about the probabilistic nature of the observations, assuming that they are independent, and obtain results holding with a probability close to one but not equal to one.

Exercise 7 *To make the link with the previous section more specific, we can focus on*

$$q_\theta(y|x) = \frac{\exp\big\{-\beta \mathbb{1}\big[f_\theta(x) \neq y\big]\big\}}{\big[1 + \big(|\mathcal{Y}| - 1\big)\exp(-\beta)\big]}.$$

Show that

$$\big[1 - \exp(-\beta)\big] \sum_{i=1}^{n} \int \mathbb{1}\big[f_\theta(X_i) \neq Y_i\big] \, \mathrm{d}\rho_i(\theta)$$
$$\leq \inf_{\theta\in\Theta} \beta \sum_{i=1}^{n} \mathbb{1}\big[f_\theta(X_i) \neq Y_i\big] - \log_2\big[\pi\big(\{\theta\}\big)\big].$$

(Hint : consider a lower bound of the left-hand side of equation (7, page 25) *and an upper bound of its right-hand side, using the fact that* $\log(1+z) \leq z$*.) Conclude that, almost surely*

$$\sum_{i=1}^{n} \int L\big[(X_i,Y_i),\theta\big]\,\mathrm{d}\rho_i(\theta) \\ \leq \inf_{\theta\in\Theta} \frac{1}{1-\beta/2} \sum_{i=1}^{n} L\big[(X_i,Y_i),\theta\big] - \frac{\log_2\big[\pi\big(\{\theta\}\big)\big]}{\beta(1-\beta/2)}.$$

(Hint: use the inequality $\exp(-\beta) \leq 1-\beta+\beta^2/2$, $\beta \geq 0$*.) How would you choose the value of the parameter* β *?*

3.2. Deviation bounds for sums of Bernoulli random variables. Given some parameter $\lambda \in \mathbb{R}$, let us consider the (normalized) log-Laplace transform of the Bernoulli distribution :

$$\Phi_\lambda(p) \stackrel{\mathrm{def}}{=} -\frac{1}{\lambda}\log\big[1-p+p\exp(-\lambda)\big].$$

Let us also consider the Kullback-Leibler divergence of two Bernoulli distributions

$$K(q,p) \stackrel{\mathrm{def}}{=} q\log\left(\frac{q}{p}\right) + (1-q)\log\left(\frac{1-q}{1-p}\right).$$

In the sequel $\overline{\mathbb{P}}$ will be the empirical measure

$$\overline{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{W_i}$$

of an i.i.d. sample $(W_i)_{i=1}^n$ drawn from $\mathbb{P}^{\otimes n} \in \mathcal{M}_+^1(\mathcal{W}^n)$. We will use a short notation for integrals, putting for any $\rho,\pi \in \mathcal{M}_+^1(\Theta)$ and any integrable function $f \in \mathbb{L}_1\big(\mathcal{W}\times\Theta^2, \mathbb{P}\otimes\pi\otimes\rho\big)$

$$f(\mathbb{P},\rho,\pi) = \int f(w,\theta,\theta')\,\mathrm{d}\mathbb{P}(w)\mathrm{d}\rho(\theta)\mathrm{d}\pi(\theta'),$$

so that for instance $L(\mathbb{P},\rho) = \displaystyle\int L(w,\theta)\mathrm{d}\mathbb{P}(w)\mathrm{d}\rho(\theta)$.

Let us recall first Chernoff's bound.

Proposition 3.1 *For any fixed value of the parameter* $\theta\in\Theta$*, the identity*

$$\int \exp\big[-\lambda L(\overline{\mathbb{P}},\theta)\big]\,\mathrm{d}\mathbb{P}^{\otimes n} = \exp\Big\{-\lambda\Phi_\lambda\big[L(\mathbb{P},\theta)\big]\Big\}$$

shows that with probability at least $1-\epsilon$*,*

$$L(\mathbb{P},\theta) \leq B_+\big[L(\overline{\mathbb{P}},\theta), \log(\epsilon^{-1})/n\big],$$

$$\text{where} \quad B_+(q,\delta) = \inf_{\lambda\in\mathbb{R}_+} \Phi_\lambda^{-1}\left(q+\frac{\delta}{\lambda}\right)$$
$$= \sup\Big\{p\in[0,1] \,:\, K(q,p)\le\delta\Big\}, \qquad q\in[0,1],\ \delta\in\mathbb{R}_+.$$

Moreover

$$-\delta q \le B_+(q,\delta) - q - \sqrt{2\delta q(1-q)} \le 2\delta(1-q).$$

In the same way, the identity

$$\int \exp\big[\lambda L(\overline{\mathbb{P}},\theta)\big] d\mathbb{P}^{\otimes n} = \exp\Big\{\lambda\Phi_{-\lambda}\big[L(\mathbb{P},\theta)\big]\Big\}$$

shows that with probability at least $1-\epsilon$

$$L(\overline{\mathbb{P}},\theta) \le B_-\big[L(\mathbb{P},\theta), \log(\epsilon^{-1})/n\big],$$
$$\text{where} \quad B_-(q,\delta) = \inf_{\lambda\in\mathbb{R}_+} \Phi_{-\lambda}(q) + \frac{\delta}{\lambda}$$
$$= \sup\Big\{p\in[0,1] \,:\, K(p,q)\le\delta\Big\}, \qquad q\in[0,1],\ \delta\in\mathbb{R}_+,$$

and

$$-\delta q \le B_-(q,\delta) - q - \sqrt{2\delta q(1-q)} \le 2\delta(1-q).$$

Let us mention here some important identity.

Proposition 3.2 *For any probability measures* π *and* ρ *on some measurable space, such that* $\mathcal{K}(\rho,\pi)<\infty$*, and any bounded measurable function* h*, let us define the transformed probability measure* $\pi_{\exp(h)} \ll \pi$ *by its density*

$$\frac{\mathrm{d}\pi_{\exp(h)}}{\mathrm{d}\pi} = \frac{\exp(h)}{Z},$$

where $Z = \int \exp(h)\,\mathrm{d}\pi$*. Let us moreover define*

$$\mathbf{Var}\big(h\,\mathrm{d}\pi\big) = \int \big(h - \textstyle\int h\,\mathrm{d}\pi\big)^2\,\mathrm{d}\pi.$$

The expectations with respect to ρ *and* π *of* h *and the* log*-Laplace transform of* h *are linked by the identities*

$$\int h\,\mathrm{d}\rho - \mathcal{K}(\rho,\pi) + \mathcal{K}(\rho,\pi_{\exp(h)}) = \log\big[\textstyle\int \exp(h)\,\mathrm{d}\pi\big] \tag{8}$$
$$= \textstyle\int h\,\mathrm{d}\pi + \int_0^1 (1-\alpha)\,\mathbf{Var}\big[h\,\mathrm{d}\pi_{\exp(\alpha h)}\big]\,\mathrm{d}\alpha. \tag{9}$$

Proof. The first identity is a straightforward consequence of the definitions of $\pi_{\exp(h)}$ and of the Kullback-Leibler divergence function. The second one is the Taylor expansion of order one with integral remainder of the function

$$f(\alpha) = \log\bigl[\int \exp(\alpha h)\,\mathrm{d}\pi\bigr],$$

which says that $f(1) = f(0) + f'(0) + \int_0^1 (1-\alpha) f''(\alpha)\,\mathrm{d}\alpha$. □

Exercise 8 *Prove that $f \in \mathcal{C}^\infty$. Hint : write*

$$h^k \exp(\alpha h) = h^k + \int_0^{+\infty} \mathbb{1}(\gamma \leq \alpha) h^{k+1} \exp(\gamma h)\,\mathrm{d}\gamma$$

and use Fubini's theorem to show that $\alpha \mapsto \int h^k \exp(\alpha h)\,\mathrm{d}\pi$ belongs to $\mathcal{C}^1$ and compute its derivative.

Let us come now to the proof of Proposition 3.1 (page 26). Chernoff's inequality reads

$$\Phi_\lambda\bigl[L(\mathbb{P},\theta)\bigr] - \frac{\log(\epsilon^{-1})}{n\lambda} \leq L(\overline{\mathbb{P}},\theta),$$

where the inequality holds with probability at least $1-\epsilon$. Since the left-hand side is non-random, it can be optimized in λ, giving

$$L(\mathbb{P},\theta) \leq B_+\bigl[L(\overline{\mathbb{P}},\theta), \log(\epsilon^{-1})/n\bigr].$$

Exercise 9 *Prove this statement in more details. For any integer $k > 1$, consider the event*

$$A_k = \Bigl\{ \sup_{\lambda \in \mathbb{R}_+} F(\lambda) - k^{-1} > L(\overline{\mathbb{P}},\theta) \Bigr\},$$

where $F(\lambda) = \Phi_\lambda\bigl[L(\mathbb{P},\theta)\bigr] - \dfrac{\log(\epsilon^{-1})}{n\lambda}$. Show that $\mathbb{P}^{\otimes n}(A_k) \leq \epsilon$ by choosing some suitable value of λ. Remark that $A_k \subset A_{k+1}$ and conclude that $\mathbb{P}^{\otimes n}\bigl(\cup_k A_k\bigr) \leq \epsilon$.

Since

$$\lim_{\lambda\to+\infty} \Phi_\lambda^{-1}\left(q + \frac{\delta}{\lambda}\right) = \lim_{\lambda\to+\infty} \frac{1-\exp(-\lambda q - \delta)}{1-\exp(-\lambda)} \leq 1,$$

$B_+(q,\delta) \leq 1$.

Applying equation (8, page 27) to Bernoulli distributions gives

$$\lambda \Phi_\lambda(p) = \lambda q + K(q,p) - K(q,p_\lambda)$$

where

$$p_\lambda = \frac{p}{p + (1-p)\exp(\lambda)}.$$

This shows that

$$\begin{aligned} B_+(q,\delta) &= \sup\Big\{p \in [0,1] \,:\, \Phi_\lambda(p) \leq q + \frac{\delta}{\lambda},\ \lambda \in \mathbb{R}_+\Big\} \\ &= \sup\Big\{p \in [q,1[\,:\, K(q,p) \leq \delta + K(q,p_\lambda), \lambda \in \mathbb{R}_+\Big\} \\ &= \sup\Big\{p \in [q,1[\,:\, K(q,p) \leq \delta\Big\} \\ &= \sup\Big\{p \in [0,1] \,:\, K(q,p) \leq \delta\Big\}, \end{aligned}$$

because when $q \leq p < 1$ then $\lambda = \log\left(\frac{q^{-1}-1}{p^{-1}-1}\right) \in \mathbb{R}_+$, $q = p_\lambda$ and therefore $K(q,p_\lambda) = 0$.

Let us remark now that $\frac{\partial^2}{\partial x^2} K(x,p) = x^{-1}(1-x)^{-1}$. Thus if $p \geq q \geq 1/2$, then

$$K(q,p) \geq \frac{(p-q)^2}{2q(1-q)},$$

so that if $K(q,p) \leq \delta$, then

$$p \leq q + \sqrt{2\delta q(1-q)}.$$

Now if $q \leq 1/2$ and $p \geq q$ then

$$K(q,p) \geq \left\{ \begin{matrix} \frac{(p-q)^2}{2p(1-p)}, & p \leq 1/2 \\ 2(p-q)^2, & p \geq 1/2 \end{matrix} \right\} \geq \frac{(p-q)^2}{2p(1-q)},$$

so that if $K(q,p) \leq \delta$, then

$$(p-q)^2 \leq 2\delta p(1-q),$$

implying that

$$p - q \leq \delta(1-q) + \sqrt{2\delta q(1-q) + \delta^2(1-q)^2} \leq \sqrt{2\delta q(1-q)} + 2\delta(1-q).$$

On the other hand,

$$K(q,p) \leq \frac{(p-q)^2}{2\min\{q(1-q), p(1-p)\}} \leq \frac{(p-q)^2}{2q(1-p)},$$

thus when $K(q,p)=\delta$ with $p>q$, then

$$(p-q)^2 \geq 2\delta q(1-p),$$

implying that

$$p-q \geq -\delta q + \sqrt{2\delta q(1-q)+\delta^2 q^2} \geq \sqrt{2\delta q(1-q)} - \delta q.$$

Exercise 10 *The second part of Proposition 3.1 (page 26) is proved in the same way and left as an exercise.*

3.3. PAC-Bayes bounds. We are now going to make Proposition 3.1 uniform with respect to θ. The PAC-Bayes approach to this is to randomize θ, so we will consider now joint distributions on $(W_1,\dots,W_n,\theta)$, where the distribution of $(W_1,\dots,W_n)$ is still $\mathbb{P}^{\otimes n}$ and the conditional distribution of θ given the sample is given by some transition probability kernel $\rho : \mathcal{W}^n \to \mathcal{M}_+^1(\Theta)$, called in this context a posterior distribution*. This posterior distribution ρ will be compared with a prior (meaning non-random) probability measure $\pi \in \mathcal{M}_+^1(\Theta)$.

Proposition 3.3 *Let us introduce the notation*

$$B_\Lambda(q,\delta) = \inf_{\lambda\in\Lambda} \Phi_\lambda^{-1}\left(q+\frac{\delta}{\lambda}\right).$$

For any prior probability measure $\pi \in \mathcal{M}_+^1(\Theta)$ and any $\lambda \in \mathbb{R}_+$,

$$\int \exp\left[\sup_{\rho\in\mathcal{M}_+^1(\Theta)} n\lambda\Big\{\Phi_\lambda\big[L(\mathbb{P},\rho)\big] - L(\overline{\mathbb{P}},\rho)\Big\} - \mathcal{K}(\rho,\pi)\right] \mathrm{d}\mathbb{P}^{\otimes n} \leq 1, \qquad (10)$$

and therefore for any finite set $\Lambda \subset \mathbb{R}_+$, with probability at least $1-\epsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$,

$$L(\mathbb{P},\rho) \leq B_\Lambda\left(L(\overline{\mathbb{P}},\rho), \frac{\mathcal{K}(\rho,\pi)+\log\big(|\Lambda|/\epsilon\big)}{n}\right),$$

Proof. The exponential moment inequality (10) is a consequence of equation (8, page 27), showing that

*We will assume that ρ is a regular conditional probability kernel, meaning that for any measurable set A the map $(w_1,\dots,w_n)\mapsto\rho(w_1,\dots,w_n)(A)$ is assumed to be measurable. We will also assume that the σ-algebra we consider on Θ is generated by a countable family of subsets. See [1][page 50] for more details

$$\exp\biggl\{\sup_{\rho\in\mathcal{M}_+^1(\Theta)} n\lambda\int\Bigl\{\Phi_\lambda\bigl[L(\mathbb{P},\theta)\bigr]-L(\overline{\mathbb{P}},\theta)\Bigr\}\,\mathrm{d}\rho(\theta)-\mathcal{K}(\rho,\pi)\biggr\}$$
$$\leq\int\exp\biggl[n\lambda\Bigl\{\Phi_\lambda\bigl[L(\mathbb{P},\theta)\bigr]-L(\overline{\mathbb{P}},\theta)\Bigr\}\biggr]\,\mathrm{d}\pi(\theta),$$

and of the fact that Φ_λ is convex, showing that $\Phi_\lambda\bigl[L(\mathbb{P},\rho)\bigr]\leq\int\Phi_\lambda\bigl[L(\mathbb{P},\theta)\bigr]\,\mathrm{d}\rho(\theta)$. The deviation inequality follows as usual. □

We cannot take the infimum on $\lambda\in\mathbb{R}_+$ as in Proposition 3.1 (page 26), because we can no more cast our deviation inequality in such a way that λ appears on some non-random side of the inequality. Nevertheless, we can get a more explicit bound from some specific choice of the set Λ.

Proposition 3.4 *Let us define the least increasing upper bound of the variance of a Bernoulli distribution of parameter $p\in[0,1]$ as*

$$\overline{v}(p)=\begin{cases}p(1-p), & p\leq 1/2,\\ 1/4, & \textit{otherwise.}\end{cases}$$

Let us choose some positive integer parameter m and let us put

$$t=\frac{1}{4}\log\biggl(\frac{n}{8\log\bigl[(m+1)/\epsilon\bigr]}\biggr).$$

With probability at least $1-\epsilon$, for any $\rho\in\mathcal{M}_+^1(\Theta)$,

$$L(\mathbb{P},\rho)\leq L(\overline{\mathbb{P}},\rho)+B_m\bigl[L(\overline{\mathbb{P}},\rho),\mathcal{K}(\rho,\pi),\epsilon\bigr],$$

where

$$B_m\bigl(q,e,\epsilon\bigr)=\max\biggl\{\sqrt{\frac{2\overline{v}(q)\bigl\{e+\log\bigl[(m+1)/\epsilon\bigr]\bigr\}}{n}}\cosh\bigl(t/m\bigr)$$
$$+\frac{2(1-q)\bigl\{e+\log\bigl[(m+1)/\epsilon\bigr]\bigr\}}{n}\cosh(t/m)^2,$$
$$\frac{2\bigl\{e+\log\bigl[(m+1)/\epsilon\bigr]\bigr\}}{n}\biggr\}$$
$$\leq\sqrt{\frac{2\overline{v}(q)\bigl\{e+\log\bigl[(m+1)/\epsilon\bigr]\bigr\}}{n}}\cosh\bigl(t/m\bigr)$$
$$+\frac{2\bigl\{e+\log\bigl[(m+1)/\epsilon\bigr]\bigr\}}{n}\cosh(t/m)^2.$$

Moreover, as soon as $n \geq 5$,

$$B_{\lfloor \log(n)^2 \rfloor - 1}(q, e, \epsilon) \leq B(q, e, \epsilon) \stackrel{\text{def}}{=} \sqrt{\frac{2\overline{v}(q)\big\{e + \log\big[\log(n)^2/\epsilon\big]\big\}}{n}} \cosh\big[\log(n)^{-1}\big] + \frac{2\big\{e + \log\big[\log(n)^2/\epsilon\big]\big\}}{n} \cosh\big[\log(n)^{-1}\big]^2, \quad (11)$$

so that with probability at least $1 - \epsilon$, *for any* $\rho \in \mathcal{M}_+^1(\Theta)$,

$$L(\mathbb{P}, \rho) \leq L(\overline{\mathbb{P}}, \rho) + \sqrt{\frac{2\overline{v}\big[L(\overline{\mathbb{P}}, \rho)\big]\Big\{\mathcal{K}(\rho, \pi) + \log\big[\log(n)^2/\epsilon\big]\Big\}}{n}} \cosh\big[\log(n)^{-1}\big] + \frac{2\Big\{\mathcal{K}(\rho, \pi) + \log\big[\log(n)^2/\epsilon\big]\Big\}}{n} \cosh\big[\log(n)^{-1}\big]^2.$$

Proof. Let us put

$$\begin{aligned} q &= L(\overline{\mathbb{P}}, \rho), \\ \delta &= \frac{\mathcal{K}(\rho, \pi) + \log\big[(m+1)/\epsilon\big]}{n}, \\ \lambda_{\min} &= \sqrt{\frac{8 \log\big[(m+1)/\epsilon\big]}{n}}, \\ \Lambda &= \Big\{\lambda_{\min}^{1-k/m}, k = 0, \ldots, m\Big\}, \\ p &= B_\Lambda(q, \delta) = \inf_{\lambda \in \Lambda} \Phi_\lambda^{-1}\left(q + \frac{\delta}{\lambda}\right), \\ \widehat{\lambda} &= \sqrt{\frac{2\delta}{\overline{v}(p)}}. \end{aligned}$$

According to equation (9, page 27) applied to Bernoulli distributions, for any $\lambda \in \Lambda$,

$$\Phi_\lambda(p) = p - \frac{1}{\lambda} \int_0^\lambda (\lambda - \alpha) p_\alpha (1 - p_\alpha) \, \mathrm{d}\alpha \leq q + \frac{\delta}{\lambda}.$$

As moreover $p_\alpha \leq p$,

$$p - q \leq \inf_{\lambda \in \Lambda} \frac{\lambda \overline{v}(p)}{2} + \frac{\delta}{\lambda} = \inf_{\lambda \in \Lambda} \sqrt{2\delta\overline{v}(p)} \cosh\left[\log\left(\frac{\widehat{\lambda}}{\lambda}\right)\right].$$

As $\overline{v}(p) \leq 1/4$ and $\delta \geq \dfrac{\log\bigl[(m+1)/\epsilon\bigr]}{n}$,

$$\sqrt{\frac{2\delta}{\overline{v}(p)}} = \widehat{\lambda} \geq \lambda_{\min} = \sqrt{\frac{8\log\bigl[(m+1)/\epsilon\bigr]}{n}}.$$

Therefore either $\lambda_{\min} \leq \widehat{\lambda} \leq 1$, or $\widehat{\lambda} > 1$. Let us consider these two cases separately.

If $\lambda_{\min} = \min\Lambda \leq \widehat{\lambda} \leq \max\Lambda = 1$, then $\log\bigl(\widehat{\lambda}\bigr)$ is at distance at most t/m from some $\log\bigl(\lambda\bigr)$ where $\lambda \in \Lambda$, because $\log(\Lambda)$ is a grid with constant steps of size $2t/m$. Thus

$$p - q \leq \sqrt{2\delta\overline{v}(p)}\cosh\bigl(t/m\bigr).$$

If moreover $q \leq 1/2$, then $\overline{v}(p) \leq p(1-q)$, so that we obtain a quadratic inequality in p, whose solution is less than

$$p \leq q + \sqrt{2\delta q(1-q)}\cosh\bigl(t/m\bigr) + 2\delta(1-q)\cosh\bigl(t/m\bigr)^2.$$

If on the contrary $q \geq 1/2$, then $\overline{v}(p) = \overline{v}(q) = 1/4$ and

$$p \leq q + \sqrt{2\delta\overline{v}(q)}\cosh\bigl(t/m\bigr),$$

so that in both cases

$$p - q \leq \sqrt{2\delta\overline{v}(q)}\cosh(t/m) + 2\delta(1-q)\cosh\bigl(t/m\bigr)^2. \tag{12}$$

Let us consider now the case when $\widehat{\lambda} > 1$. In this case

$$p - q \leq \sqrt{2\delta\overline{v}(p)}\,\widehat{\lambda} = 2\delta.$$

In conclusion, applying Proposition 3.3 (page 30) we see that with probability at least $1-\epsilon$, for any posterior distribution ρ,

$$L(\mathbb{P},\rho) \leq p \leq q + \max\Bigl\{2\delta, \sqrt{2\delta\overline{v}(q)}\cosh\bigl(t/m\bigr) + 2\delta(1-q)\cosh\bigl(t/m\bigr)^2\Bigr\},$$

which is precisely the statement to be proved.

In the special case when $m = \lfloor\log(n)^2\rfloor - 1 \geq \log(n)^2 - 2$,

$$\frac{t}{m} \leq \frac{1}{4\bigl[\log(n)^2 - 2\bigr]}\log\left(\frac{n}{8\log\bigl[\log(n)^2 - 1\bigr]}\right) \leq \log(n)^{-1}$$

as soon as the last inequality holds, that is as soon as $n \geq \exp(\sqrt{2}) \simeq 4.11$ to make $\log(n)^2 - 2$ positive and

$$3\log(n)^2 - 8 + \log(n)\log\Bigl\{8\log\bigl[\log(n)^2 - 1\bigr]\Bigr\} \geq 0,$$

which holds true for any $n \geq 5$, as can be checked numerically. $\square$

4. Linear classification and support vector machines

We are going in this section to consider more specifically the case of linear binary classification. In this setting $\mathcal{W} = \mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \{-1, +1\}$, $w = (x, y)$, where $x \in \mathbb{R}^d$ and $y \in \{-1, +1\}$, $\Theta = \mathbb{R}^d$, and

$$L(w, \theta) = \mathbb{1}\big[\langle \theta, x \rangle y \leq 0\big].$$

Although we will stick in this presentation to the case when $\mathcal{X}$ is a vector space of finite dimension, the results also apply to support vector machines, where the pattern space is some arbitrary space mapped to a Hilbert space $\mathcal{H}$ by some implicit mapping $\Psi : \mathcal{X} \to \mathcal{H}$, $\Theta = \mathcal{H}$ and $L(w, \theta) = \mathbb{1}\big(\langle \theta, \Psi(x) \rangle y \leq 0\big)$. It turns out that classification algorithms do not need to manipulate $\mathcal{H}$ itself, but only to compute scalar products of the form $k(x_1, x_2) = \langle \Psi(x_1), \Psi(x_2) \rangle$, defining a symmetric positive kernel k on the original pattern space $\mathcal{X}$. The converse is also true, any positive symmetric kernel k can be represented as a scalar product in some mapped Hilbert space (this is the Moore-Aronszajn theorem). Often used kernels on $\mathbb{R}^d$ are

$$\begin{aligned} k(x_1, x_2) &= \big(1 + \langle x_1, x_2 \rangle\big)^s, \text{ for which } \dim \mathcal{H} < \infty, \\ k(x_1, x_2) &= \exp\big(-\|x_1 - x_2\|^2\big), \text{ for which } \dim \mathcal{H} = +\infty. \end{aligned}$$

In the following, we will work in $\mathbb{R}^d$, which covers only the case when $\dim \mathcal{H} < \infty$, but extensions would be possible.

Let us consider, after [5, 8] as prior probability measure π the centered Gaussian measure with covariance $\beta^{-1}\,\mathbf{Id}$, so that

$$\frac{\mathrm{d}\pi}{\mathrm{d}\theta}(\theta) = \left(\frac{\beta}{2\pi}\right)^{d/2} \exp\left(-\frac{\beta \|\theta\|^2}{2}\right).$$

Let us also consider the function

$$\begin{aligned} \varphi(x) &= \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} \exp\big(-t^2/2\big)\,\mathrm{d}t, \qquad x \in \mathbb{R} \\ &\leq \min\left\{\frac{1}{x\sqrt{2\pi}}, \frac{1}{2}\right\} \exp\left(-\frac{x^2}{2}\right), \quad x \in \mathbb{R}_+. \end{aligned}$$

Let π_θ be the measure π shifted by θ, defined by the identity

$$\int h(\theta')\,\mathrm{d}\pi_\theta(\theta') = \int h(\theta + \theta')\,\mathrm{d}\pi(\theta').$$

In this case

$$\mathcal{K}(\pi_\theta, \pi) = \frac{\beta}{2}\|\theta\|^2,$$

and

$$L(w, \pi_\theta) = \varphi\big[\sqrt{\beta}\|x\|^{-1}\langle\theta, x\rangle y\big].$$

Thus the randomized loss function has an explicit expression : randomization replaces the indicator function of the negative real line by a smooth approximation. As we are eventually interested in $L(w, \theta)$, we will shift things a little bit, considering along with the classification error function L some *error with margin*

$$M(w, \theta) = \mathbb{1}\big[y\|x\|^{-1}\langle\theta, x\rangle \leq 1\big].$$

Unlike $L(w, \theta)$ which is independent of the norm of θ, the margin error $M(w, \theta)$ depends on $\|\theta\|$, counting a classification error each time x is at distance less than $\|x\|/\|\theta\|$ from the boundary $\{x' : \langle\theta, x'\rangle = 0\}$, so that the error with margin region is the complement of the open cone $\big\{x \in \mathbb{R}^d\,;\, y\langle\theta, x\rangle > \|x\|\big\}$.

Let us compute the randomized margin error

$$M(w, \pi_\theta) = \varphi\Big\{\sqrt{\beta}\big[y\|x\|^{-1}\langle\theta, x\rangle - 1\big]\Big\}.$$

It satisfies the inequality

$$M(w, \pi_\theta) \geq \varphi(-\sqrt{\beta}\,)L(w, \theta) = \big[1 - \varphi\big(\sqrt{\beta}\,\big)\big]L(w, \theta).$$

Applying previous results we obtain

Proposition 4.1 *With probability at least $1 - \epsilon$, for any $\theta \in \mathbb{R}^d$,*

$$L(\mathbb{P}, \theta) \leq \big[1 - \varphi(\sqrt{\beta})\big]^{-1} M(\mathbb{P}, \pi_\theta) \leq C_1(\theta),$$

where

$$C_1(\theta) = \big[1 - \varphi\big(\sqrt{\beta}\,\big)\big]^{-1} B\left(M(\,\overline{\mathbb{P}}, \pi_\theta), \frac{\beta\|\theta\|^2}{2}, \epsilon\right),$$

the bound B being defined by equation (11, page 32)*.*

We can now minimize this empirical upper-bound to define an estimator. Let us consider some estimator $\widehat{\theta}$ such that

$$C_1(\widehat{\theta}) \leq \inf_{\theta\in\mathbb{R}^d} C_1(\theta) + \zeta.$$

Then for any fixed parameter $\theta_\star$, $C_1(\theta) \leq C_1(\theta_\star) + \zeta$. On the other hand, with probability at least $1-\epsilon$

$$M(\overline{\mathbb{P}}, \pi_{\theta_\star}) \leq B_-\bigg(M(\mathbb{P}, \pi_{\theta_\star}), \frac{\log(\epsilon^{-1})}{n}\bigg).$$

Indeed

$$\int \exp\Big\{n\lambda\big[M(\overline{\mathbb{P}}, \pi_{\theta_\star}) - \Phi_{-\lambda}\big[M(\mathbb{P}, \pi_{\theta_\star})\big]\Big\}\, \mathrm{d}\mathbb{P}^{\otimes n}$$
$$\leq \int \exp\bigg\{n\lambda \int \Big\{M(\overline{\mathbb{P}}, \theta) - \Phi_{-\lambda}\big[M(\mathbb{P}, \theta)\Big\}\, \mathrm{d}\pi_{\theta_\star}(\theta)\bigg\}\, \mathrm{d}\mathbb{P}^{\otimes n} \leq 1,$$

because $p \mapsto -\Phi_{-\lambda}(p)$ is convex. As a consequence

Proposition 4.2 *With probability at least* $1-2\epsilon$,

$$L(\mathbb{P}, \widehat{\theta}) \leq$$
$$\inf_{\theta_\star \in \Theta} \big[1 - \varphi\big(\sqrt{\beta}\,\big)\big]^{-1} B\bigg(B_-\bigg(M(\mathbb{P}, \pi_{\theta_\star}), \frac{\log(\epsilon^{-1})}{n}\bigg), \frac{\beta\|\theta_\star\|^2}{2}, \epsilon\bigg) + \zeta.$$

It is also possible to state a result in terms of empirical margins. Indeed

$$M(w, \pi_\theta) \leq M(w, \theta/2) + \varphi(\sqrt{\beta}).$$

Thus with probability at least $1-\epsilon$, for any $\theta \in \mathbb{R}^d$,

$$L(\mathbb{P}, \theta) \leq C_2(\theta),$$

where

$$C_2(\theta) = \big[1 - \varphi\big(\sqrt{\beta}\,\big)\big]^{-1} B\bigg(M(\overline{\mathbb{P}}, \theta/2) + \varphi\big(\sqrt{\beta}\,\big), \frac{\beta\|\theta\|^2}{2}, \epsilon\bigg).$$

However, C_1 and C_2 are non-convex criterions, faster minimization algorithms are available for the usual SVN loss function, for which it is also possible to derive some generalization bound. Indeed

$$M(w, \pi_\theta) = \varphi\big[\sqrt{\beta}\big(\langle\theta, x\rangle y - 1\big)\big] \leq \big(2 - \langle\theta, x\rangle y\big)_+ + \varphi(\sqrt{\beta}).$$

Thus we also have, using this time Proposition 3.3 (page 30)

Proposition 4.3 *With probability at least* $1-\epsilon$*, for any* $\theta \in \mathbb{R}^d$*,*

$$L(\mathbb{P},\theta) \leq \big[1-\varphi\big(\sqrt{\beta}\big)\big]^{-1} B_\Lambda\bigg(\int \big(2-\langle\theta,x\rangle y\big)_+ d\overline{\mathbb{P}}(x,y) + \varphi(\sqrt{\beta}), \\ \frac{\beta\|\theta\|^2 + 2\log\big(|\Lambda|/\epsilon\big)}{2n}\bigg) \\ = \big[1-\varphi\big(\sqrt{\beta}\big)\big]^{-1} \inf_{\lambda\in\Lambda} \Phi_\lambda^{-1}\bigg[C_3(\lambda,\theta) + \varphi\big(\sqrt{\beta}\,\big) + \frac{\log\big(|\Lambda|/\epsilon\big)}{n\lambda}\bigg],$$

where

$$C_3(\lambda,\theta) = \int \big(2-\langle\theta,x\rangle y\big)_+ \,\mathrm{d}\overline{\mathbb{P}}(x,y) + \frac{\beta\|\theta\|^2}{2n\lambda}.$$

The loss function $C_3(\lambda,\theta)$ is the most employed learning criterion for support vector machines, and is called the box constraint. It is convex in θ. There are fast algorithms to compute $\inf_\theta C_3(\lambda,\theta)$ for any fixed value of λ. Here we get an empirical criterion which could be used to optimize also the value of λ.

References

[1] O. Catoni. *Statistical Learning Theory and Stochastic Optimization, Lectures on Probability Theory and Statistics, École d'Été de Probabilités de Saint-Flour XXXI – 2001*, volume 1851 of *Lecture Notes in Mathematics*. Springer, 2004. Pages 1–269.

[2] O. Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes Monograph Series*. Institute of Mathematical Statistics, 2007. Pages i-xii, 1-163.

[3] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley and Sons, New York, second edition, 2006.

[4] Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. Pac-bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 353–360, New York, NY, USA, 2009. ACM.

[5] J. Langford and J. Shawe-Taylor. PAC-bayes & margins. In *Advances in Neural Information Processing Systems*, pages 423–430, 2002.

[6] D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th annual conference on Computational Learning Theory*. Morgan Kaufmann, 1999.

[7] D. A. McAllester. PAC-Bayesian stochastic model selection. *Mach. Learn.*, 51(1):5–21, April 2003.

[8] David Mcallester. Simplified pac-bayesian margin bounds. In *In COLT*, pages 203–215, 2003.

[9] M. Seeger. PAC-Bayesian generalization error bounds for gaussian process classification. Informatics report series EDI-INF-RR-0094, Division of Informatics, University of Edinburgh, 2002.

[10] T. van Erven, P. D. Grünwald, and S. de Rooij. Catching up faster by switching sooner : A predictive approach to adaptive estimation with an application to the AIC-BIC dilemna. *JRSS B*, 2011.

[11] F.M.J. Willems, Y.M. Shtarkov, and T.J. Tjalkens. The context-tree weighting method: basic properties. *IEEE Trans. Inform. Theory*, 41(3):653–664, 1995.