Majorizing codes and measures

Andreas Maurer
Adalbertstr. 55
D-80799 München
am@andreas-maurer.eu

Abstract

An information theoretical interpretation of majorizing and minorizing measures is given. The expression logarithmic in the reciprocal of the measure of a ball is replaced by the number of bits needed to achieve desired precision in some convergent code. We also give a local version of the majorizing bound.

1 Introduction

Consider a compact metric space (T, d) of unit diameter and a random process X_t indexed by T such that

$$\Pr\{|X_{t_1} - X_{t_2}| > s\} \leq 2\exp\left(\frac{-s^2}{d^2(t_1, t_2)}\right) \text{ for } t_1, t_2 \in T \text{ and } s > 0. \quad (1)$$

The boundedness properties of such sub-gaussian processes have been intensively studied. Let t_0 be a fixed member of T. Then the metric entropy inequality ([4], [8])

$$E\sup_{t\in T}|X_t - X_{t_0}| \leq C\int_0^1 \sqrt{N(T, d, \epsilon)}d\epsilon \quad (2)$$

holds, where $N(T, d, \epsilon)$ is the smallest number of open balls of radius ϵ needed to cover T and C is a universal constant.

Now suppose that there is a probability measure m defined on T. Then we have the majorizing measure bound ([4],[6],[7])

$$E\sup_{t\in T}|X_t - X_{t_0}| \leq C'\sup_{t\in T}\int_0^1 \sqrt{\ln\frac{1}{m(B(t, \epsilon))}}d\epsilon, \quad (3)$$

where $B(t, \epsilon)$ denotes the open ball of radius ϵ centered at t and C' is another universal constant. This is an improvement, since for an appropriate choice of the measure m the inequality (2) can be recovered from (3) up to a constant. Michel Talagrand also showed that for Gaussian processes it is possible to choose

m so as to reverse the above inequality, thus providing a complete characterization of boundedness in terms of these majorizing measures ([4], [5],[7]).

Several alternatives equivalent to the right hand side in (3) have been derived, ranging from sums involving nested partitions ([4] and [6]) to approximating sets [6] and various kinds of trees (see the appendix of [7]). The construction of these alternatives was largely motivated by specific analytical problems, such as the proofs of upper or lower bounds, for which one or another formulation is more convenient. The different formulations can also be understood as providing different perspectives on the interpretation of the right hand side of (3), which, according to Talagrand, measures the "size" of the space T, once the infimum over all probability measures m has been taken.

Such an interpretation is the subject of the present paper, which relates the boundedness of the process $(X_t)_{t\in T}$ to the possibility of giving efficient names to the members of T.

The expression under the square-root in the integrand in (3) already suggests an information theoretical interpretation, intuitively something like "the number of bits needed to describe t with precision ϵ". In the following this intuition will be made precise.

A code for T assigns to every $t \in T$ a codeword $c(t)$, which is a finite or infinite sequence of characters from a finite alphabet α. A simple example is $T = [0,1]$, when every $t \in T$ is encoded by binary expansion. Since the cardinality of T can be very large of infinite, most codewords will be very long or infinite, and therefore difficult or impossible to memorize or communicate. We are therefore compelled to tolerate some ambiguity which arises from the truncation of codewords. Given a code c, a point $t \in T$ and a truncation length $k \in \mathbb{N}$, the truncated codeword $p_k(c(t))$ is defined as the string composed of the leading k characters of $c(t)$, if $c(t)$ is longer that k, and $p_k(c(t)) = c(t)$ otherwise.

In the setting of metric spaces the ambiguity caused by truncation can be quantified as the diameter of the set of points whose truncated codewords are identical.

$$D_k(c,t) = \sup\{d(t',t'') : p_k(c(t')) = p_k(c(t'')) = p_k(c(t))\}.$$

The code is called convergent if $D_k(c,t) \to 0$ as $k \to \infty$ for every $t \in T$. For a convergent code and $t \in T$, $\epsilon > 0$ we define

$$len(t,\epsilon) = \min\{k : D_k(c,t) < \epsilon\}.$$

In Section 2 these definitions will be repeated in a more formal setting. In the meantime, to gain some intuition, let us suppose that the code is binary and that we are to receive the codeword of some $t \in T$ bit after bit, in the style of a guessing game. Initially the set of possible candidates for t is all of T. Subsequently every bit we receive corresponds to the answer of a yes-no question, which separates the set of previously possible candidates into an

impossible set and a set of remaining possible candidates for t. After k bits we have received the string $p_k(c(t))$ and the set of possible candidates has diameter $D_k(c,t)$. If the code is convergent this diameter will become arbitrarily small in a finite time, and $len(c,t,\epsilon)$ is the number of bits communicated until the set of remaining candidates has shrunk to a diameter smaller than ϵ. So $len(c,t,\epsilon)$ is a scale sensitive ("ϵ") and local ("t") measure of descriptive complexity incurred through the use of the code c.

The following is the principal contribution of this paper, stated here for binary codes. More general statements are given in Theorems 8 and 14.

Theorem 1 *Let (T,d) be a compact metric space of unit diameter.*

(i) For every convergent binary code c for T there is a probability measure m on T such that for every $t \in T$.

$$\int_0^1 \sqrt{\ln \frac{1}{m(B(t,\epsilon))}} d\epsilon \leq 2 \int_0^1 \sqrt{len(c,t,\epsilon)} d\epsilon + 2.$$

(ii) For every probability measure m on T there is a convergent binary code $c : T \to \{0,1\}^{\mathbb{N}}$ such that for every $t \in T$

$$\sqrt{\ln 2} \int_0^1 \sqrt{len(c,t,\epsilon)} d\epsilon \leq 93 \int_0^1 \sqrt{\ln \frac{1}{m(B(t,\epsilon))}} d\epsilon + 69.$$

The (sample-) boundedness of subgaussian processes can therefore be characterized in terms of the convergence properties of codes for the index set. Such follows from the majorizing measure bound (3) and the minorizing inequality for Gaussian processes [4][5].

The integral over scales can be equivalently (up to constants) expressed as a weighted sum over the questions asked and answered in the above guessing game.

Theorem 2 *Let (T,d) be a compact metric space of unit diameter and c any convergent code for T. Then for every $t \in T$*

$$1 + 2^{-3/2} \sum_{k=1}^{\infty} \frac{D_k(c,t)}{\sqrt{k}} \leq \int_0^1 \sqrt{len(c,t,\epsilon)} d\epsilon \leq 1 + 2^{-1} \sum_{k=1}^{\infty} \frac{D_k(c,t)}{\sqrt{k}}.$$

A consequence of this formula and (3) is, that boundedness of a subgaussian processes is ensured, whenever there exists a code which "comes to the point" as quickly as $O(k^{-p})$ with $p > 1/2$, for every point $t \in T$.

Observe however that the above inequalities are uniform, in the sense that they hold for every $t \in T$, not just for the comparison of the suprema. The majorizing bound (3) can be extended to hold in a similar way. The following result is stated here in terms of binary codes (but in view of Theorem 1 it could have been stated equivalently in terms of probability measures). We also state the result for finite T and refer to [4] for the infinite case and the issues of measurability which arise.

Theorem 3 *Let (T,d) be a finite metric space of unit diameter. Given the sub-gaussian condition (1), a fixed member t_0 of T, a convergent binary code c for T and $\delta \in (0,1)$, we have with probability at least $1-\delta$ that for all $t \in T$*

$$|X_t - X_{t_0}| \leq 4\sqrt{\ln 2}\int_0^1 \sqrt{len\,(c,t,\epsilon)}d\epsilon + 4\sqrt{\ln \frac{2}{\delta}}.$$

Replacing the code by a probability measure, as indicated by Theorem 1, using integration by parts and passing to suprema we can recover the classical inequality (3) up to a constant. Observe however that the right hand side of the majorizing measure bound (3) only gives global information, while Theorem 3 can be used to motivate and justify the selection of some specific $t \in T$, because the values of the integral can be taken into consideration in the evaluation of an optimization candidate. This makes contact to work which applies majorizing measures to statistical learning theory [1].

The author is unaware of other work making explicit the interconnection between majorizing (or minorizing) measures on one side and the encoding of metric spaces on the other side. This is surprising, because of the suggestive presence of the logarithm in the right hand side of (3). Also, majorizing measures are intimately connected with nested partitions, as shown in [6], while on the other hand certain systems of nested partitions are equivalent to codes. That the majorizing bound holds as in Theorem 3, and not just for a comparison of the suprema, is implicit in the proofs of, but never stated in [4],[6] and [7]. It is stated in [1] under more specialized circumstances.

2 Codes in metric spaces

In this section we introduce definitions and some facts pertaining to the encoding of metric spaces.

Let α be a finite set, called the alphabet. Its cardinality is denoted with $|\alpha|$. A *finite string* s over α is a finite sequence

$$s = (s\,(1)\,,...,s\,(len\,(s))) \in \alpha^{len(s)}.$$

The length $len\,(s)$ is a property of the string s. The set α^* of finite strings contains strings of all length. The string of length zero is called the empty string and denoted with 0.

An *infinite string* over α is an infinite sequence

$$s = (s\,(k))_{k\in\mathbb{N}} \in \alpha^{\mathbb{N}}.$$

With α^{**} we denote the set of all strings, finite or infinite, $\alpha^{**} = \alpha^* \cup \alpha^{\mathbb{N}}$. If $s \in \alpha^{\mathbb{N}}$ we write $len\,(s) = \infty$. If $s_1 \in \alpha^*$ and $s_2 \in \alpha^{**}$ we use $s_1 \circ s_2$ to denote their concatenation

$$(s_1 \circ s_2)\,(k) = \begin{cases} s_1\,(k) & \text{if} \quad 1 \leq k \leq len\,(s_1) \\ s_2\,(k - len\,(s_1)) & \text{if} \quad len\,(s_1) < k \leq len\,(s_1) + len\,(s_2) \end{cases}.$$

If $s = s_1 \circ s_2$ then s_1 is called a prefix of s, and s is called an extension of s_1. The set of extensions of a string s is $s \circ \alpha^{**} = \{s \circ s' : s' \in \alpha^{**}\}$, the set of finite extensions is $s \circ \alpha^* = \{s \circ s' : s' \in \alpha^*\}$. For $k \in \mathbb{N}_0$, the truncation of s at k is the finite string $p_k(s)$ defined as

$$p_k(s) = \begin{cases} 0 & \text{if} \quad k = 0 \\ (s(1), ..., s(k)) & \text{if} \quad 1 \leq k < len(s) \\ s & \text{if} \quad len(s) \leq k \end{cases}$$

A *code* for a set T over an alphabet α is a mapping $c : T \to \alpha^{**}$ from T to the set α^{**} of strings over α. The range $c(T)$ of the function c is called the set of codewords, and for $t \in T$ the string $c(t)$ is the codeword representing t. The code is called finite if $c(T) \subseteq \alpha^*$. We use c^{-1} to denote the (set-valued) inverse function. If $s \in \alpha^*$ is any finite string then

$$\hat{c}(s) := c^{-1}(s \circ \alpha^{**}) \subseteq T$$

is the set of points of T which correspond to possible extensions of s, so $\hat{c}(s)$ is the subset of T, which is effectively described by the message string s. If s_1 is a prefix of $s_2 \in \alpha^*$ then $\hat{c}(s_1) \supseteq \hat{c}(s_2)$. Also we have $\hat{c}(0) = T$. Of course, depending on s, the set $\hat{c}(s)$ may well be empty. The vocabulary $\mathcal{V}(c)$ is the set of finite prefixes actually used by the code, that is

$$\mathcal{V}(c) = \{v \in \alpha^* : \hat{c}(s) \neq \emptyset\},$$

so v is in the vocabulary if and only if v is finite and extends to a codeword. The vocabulary consists of finite strings which describe nonempty subsets of T, meaningful messages as we might say. Since $\hat{c}(s) \neq \emptyset$ for every $s \in \mathcal{V}(c)$, there is a prototype function $\pi_c : \mathcal{V}(c) \to T$ such that $\pi(s) \in \hat{c}(s)$ for every $s \in \mathcal{V}(c)$.

If there is a sigma algebra Σ defined on T, we say that c is Σ-measurable if $\forall k \in \mathbb{N}, a \in \alpha$

$$\{t \in T : len(c(t)) \geq k, (c(t))(k) = a\} \in \Sigma.$$

Note that this implies than $\hat{c}(s) \in \Sigma$, for all $s \in \alpha^*$ and that $len(c(.))$ is a Σ-measurable function.

To say that a finite code is *instantaneous* means that means that for any $t_1, t_2 \in T$, $t_1 \neq t_2$ we have $c(t_1) \notin c(t_2) \circ \alpha^*$. So no codeword can be a prefix to any other, and every codeword can be uniquely decoded as soon as its last character has been received, hence the name. Instantaneous codes are sometimes also called *prefix-free*. A finite code is instantaneous if and only if $\hat{c}(c(t)) = \{t\}$ for every $t \in T$.

Theorem 4 (Kraft inequality [3]) *For any instantaneous code $c : T \to \alpha^*$ we have*

$$\sum_{t \in T} |\alpha|^{-len(c(t))} \leq 1,$$

where $|\alpha|$ *denotes the cardinality of the alphabet* α.

Conversely, if $l : T \to \mathbb{N}$ *satisfies* $\sum_{t\in T} |\alpha|^{-l(t)} \leq 1$ *then there exists an instantaneous code* $c : T \to \alpha^*$ *such that* $l(t) = len(c(t))$.

Suppose now that (T, d) is a metric space. We will always assume that any code $c : T \to \alpha^{**}$ is Borel-measurable. For $A \subseteq T$ we use $D(A)$ to denote the diameter of A, so $D(A) = \sup\{d(x, y) : x, y \in A\}$.

Fix a code $c : T \to \alpha^{**}$. If a string v is in the vocabulary, $v \in \mathcal{V}(c)$, the number $D(\hat{c}(v))$ can be interpreted as the ambiguity associated with v. For $t \in T$ and $k \in \mathbb{N}_0$ we use the shorthand

$$D_k(c, t) = D(\hat{c}(p_k(c(t)))),$$

for the ambiguity incurred by the truncation of $c(t)$ to length at most k. The sequence $D_k(c, t)$ is nonincreasing. A code is convergent if $\lim_{k\to\infty} D_k(c, t) = 0$ for every $t \in T$. A finite code is convergent if and only if $D(\hat{c}(c(t))) = 0$, which happens if and only if $\hat{c}(c(t)) = \{t\}$ which happens if and only if the code is instantaneous.

For any convergent code c, $t \in T$ and $\epsilon > 0$ there exists K such that $D_k(c, t) < \epsilon$ for all $k \geq K$. We can therefore define a function $len(c, ., .) : T \times (0, 1] \to \mathbb{N}_0$ by

$$len(c, t, \epsilon) = \min\{k : D_k(c, t) < \epsilon\}$$

and a two-argument function $c : T \times (0, 1] \to \mathcal{V}(c)$ by

$$c(t, \epsilon) = p_{len(c,t,\epsilon)}(c(t)).$$

So $c(t, \epsilon)$ is the minimum length prefix of $c(t)$ which has an ambiguity bounded by ϵ, and $len(c, t, \epsilon)$ is the length of this prefix. $c(t, \epsilon)$ is the code of t truncated to precision ϵ.

If the ambiguity of any $v \in \mathcal{V}(c)$ string is smaller than $\epsilon > 0$, if $D(\hat{c}(v)) < \epsilon$, we can extend this definition to v by setting

$$\begin{aligned} len(c, v, \epsilon) &= \min\{k : D(\hat{c}(p_k(v))) < \epsilon\} \\ c(v, \epsilon) &= p_{len(c,v,\epsilon)}(v). \end{aligned}$$

Observe also that these definitions imply a projection formula: If $t \in T$, $v \in \mathcal{V}(c)$, $D(\hat{c}(v)) < \epsilon \leq \eta$ then

$$c(t, \eta) = c(c(t, \epsilon), \eta) \text{ and } c(v, \eta) = c(c(v, \epsilon), \eta) \tag{4}$$

For fixed ϵ the range of the function $c_\epsilon : t \in T \mapsto c(t, \epsilon) \in \mathcal{V}(c)$ is denoted by $\mathcal{S}_\epsilon(c)$. This can be regarded as a cross-section of the code at metric resolution ϵ.

Convergent codes, when truncated at any specified precision, produce instantaneous codes on these cross-sections.

Proposition 5 *Fix a convergent $c : T \to \alpha^{**}$ and $\epsilon \geq 0$.*

(i) The identity map on $\mathcal{S}_\epsilon(c)$ is an instantaneous code.

(ii)

$$\sum_{v \in \mathcal{S}_\epsilon(c)} |\alpha|^{-len(v)} \leq 1. \tag{5}$$

(iii) If $0 \leq \epsilon \leq \eta$ the map $v \in \mathcal{S}_\epsilon(c) \mapsto c(v, \eta)$ maps $\mathcal{S}_\epsilon$ onto $\mathcal{S}_\eta$.

Proof. Let $t, t' \in T$ and assume that $c(t, \epsilon) = c(t', \epsilon) \circ s$ for some $s \in \alpha^*$. Then $c(t) \in c(t, \epsilon) \circ \alpha^* = c(t', \epsilon) \circ s \circ \alpha^* \subseteq c(t', \epsilon) \circ \alpha^*$ and $D(c(t', \epsilon)) \leq \epsilon$. By the optimality of $c(t, \epsilon)$ we therefore have $len(c(t, \epsilon)) \leq len(c(t', \epsilon))$. Thus $len(c(t', \epsilon)) + len(s) = len(c(t, \epsilon)) \leq len(c(t', \epsilon))$ so that $len(s) = 0$ and therefore $c(t, \epsilon) = c(t', \epsilon)$.

We have shown the following: It $t, t' \in T$ then either $c(t, \epsilon) = c(t', \epsilon)$ or neither of $c(t, \epsilon)$ and $c(t', \epsilon)$ is a prefix of the other one. This implies (i) and, by the Kraft inequality (Theorem 4) it implies (ii). (iii) follows from the projection formula (4). ■

To every such cross-section corresponds a suitably fine partition of T. The following facts, which we will not use, are easy to prove.

- The collection of subsets

  $$\mathcal{P}_\epsilon(c) = \{\hat{c}(v) : v \in \mathcal{S}_\epsilon(c)\}$$

  is a partition of T.
- Each member of $\mathcal{P}_\epsilon(c)$ has diameter less than ϵ.
- If $\epsilon \leq \eta$ then $\mathcal{P}_\epsilon(c)$ is a refinement of $\mathcal{P}_\eta(c)$.

3 From codes to random processes

We proof the local majorizing bound (Theorem 3) in the case of finite T and a finite code c. The proof is essentially a union bound over the cross-sections $\mathcal{S}_{2^{-\iota}}(c)$ of the vocabulary at exponentially decreasing scales. It is easier to derive this result from codes than from measures, where an intermediate construction (partitions or ultrametric) is required.

Theorem 6 *Suppose that (T, d) is a finite metric space of unit diameter, $c : T \to \alpha^*$ is an instantaneous code, and X_t is a random process indexed by T such that*

$$\Pr\{|X_t - X_{t'}| > s\} \leq K \exp\left(-\left(\frac{s}{d(t, t')}\right)^p\right) \tag{6}$$

for all $t \geq 0$, $t, t' \in T$ and some $K \geq 1$. Let t_0 be an arbitrary element of T. Then for $\delta \in (0,1)$ with probability at least $1-\delta$ we have that for every $t \in T$

$$|X_t - X_{t_0}| \leq \sum_{l>0} 2^{-l+1} \left(\ln |\alpha| \, len\left(c, t, 2^{-l}\right) + \ln \frac{2^l K}{\delta} \right)^{1/p}.$$

Proof. Recall the prototype function $\pi : \mathcal{V} \to T$ which gives an element of $\hat{c}(v)$ for every $v \in \mathcal{V}$. We now modify it to make $\pi(0) = t_0$, which is possible, since $\hat{c}(0) = T$. Observe that for any $t \in T$ we have $\pi(c(t,1)) = \pi(0) = t_0$, while finiteness of T implies that for sufficiently large l all t' different from t are further than 2^{-l} from t, so that $c\left(t, 2^{-l}\right) = c(t)$ and $\pi\left(c\left(t, 2^{-l}\right)\right) = t$. This implies the chaining inequality: For every $t \in T$

$$|X_t - X_{t_0}| \leq \sum_{l>0} \left| X_{\pi(c(t,2^{-l}))} - X_{\pi(c(t,2^{-l+1}))} \right|.$$

If for $l \in \mathbb{N}$ we have $v \in \mathcal{S}_{2^{-l}}$, then we denote $\hat{v} := c\left(v, 2^{-(l-1)}\right) \in \mathcal{S}_{2^{-(l-1)}}$. The string $\hat{v}$ can be regarded as the parent of v, with resolution coarser by a factor of 2. Both $\pi(v)$ and $\pi(\hat{v})$ are members of $\hat{c}(\hat{v})$, which has diameter bounded by $2^{-(l-1)}$, so

$$d\left(\pi(v), \pi(\hat{v})\right) \leq 2^{-(l-1)}, \text{ for all } v \in \mathcal{S}_{2^{-l}}. \tag{7}$$

For $l \geq 0$ we also define a function $\xi_l : \mathcal{S}_{2^{-l}} \to \mathbb{R}_+$ as follows. For $v \in \mathcal{S}_{2^{-l}}$ we set

$$\xi_l(v) = 2^{-l+1} \left(\ln |\alpha| \, len(v) + \ln \left(\frac{2^l K}{\delta} \right) \right)^{1/p}.$$

We have to show that

$$\Pr\left\{ \exists t \in T : \; |X_t - X_{t_0}| - \sum_{l>0} \xi_l\left(c\left(t, 2^{-l}\right)\right) > 0 \right\} \leq \delta.$$

Denote the left side of this inequality with P. By the chaining inequality

$$P \leq \Pr\left\{ \exists t \in T : \sum_{l>0} \left(\left| X_{\pi(c(t,2^{-l}))} - X_{\pi(c(t,2^{-l+1}))} \right| - \xi_l\left(c\left(t, 2^{-l}\right)\right) \right) > 0 \right\}.$$

If the sum is positive, at least one of the terms has to be positive, so

$$\begin{aligned} P &\leq \Pr\left\{ \exists t \in T, \exists l > 0 : \; \left| X_{\pi(c(t,2^{-l}))} - X_{\pi(c(t,2^{-l+1}))} \right| > \xi_l\left(c\left(t, 2^{-l}\right)\right) \right\} \\ &= \Pr\left\{ \exists l > 0, \exists v \in \mathcal{S}_{2^{-l}} : \left| X_{\pi(v)} - X_{\pi(\hat{v})} \right| > \xi_l(v) \right\} \\ &\leq \sum_{l>0} \sum_{v \in \mathcal{S}_{2^{-l}}} \Pr\left\{ \left| X_{\pi(v)} - X_{\pi(\hat{v})} \right| > \xi_l(v) \right\}. \end{aligned}$$

By (7), for $v \in \mathcal{S}_{2^{-l}}$, we have $d\left(\pi(v), \pi(\hat{v})\right) < 2^{-(l-1)}$. Using (6) we obtain

$$P \leq \sum_{l>0} \sum_{v \in \mathcal{S}_{2^{-l}}} K \exp\left(- \left(2^{l-1} \xi_l(v) \right)^p \right) = \delta \sum_{l>0} \frac{1}{2^l} \sum_{v \in \mathcal{S}_{2^{-l}}} |\alpha|^{-len(v)},$$

where we have used the definition of $\xi_l(v)$. But by the Kraft inequality, (5), the inner sum is at most one, so $P \leq \delta \sum_{l>0} 2^{-l} = \delta$, which completes the proof. ■

We will make repeated use of the following inequalities.

Lemma 7 *Let $f : (0,1] \to \mathbb{R}_+$ be measurable and nonincreasing and $r > 1$. Then*

$$(r-1)\sum_{k=1}^{\infty} r^{-k} f\left(r^{-k+1}\right) \leq \int_0^1 f(\epsilon)\, d\epsilon \leq (r-1)\sum_{k=1}^{\infty} r^{-k} f\left(r^{-k}\right).$$

Convergence of the right hand sum implies existence of the integral, which in turn implies convergence of the left hand sum.

Proof. We sum, over $k \in \mathbb{N}_0$, the inequalities

$$\left(r^{-k+1} - r^{-k}\right) f\left(r^{-k+1}\right) \leq \int_{r^{-k}}^{r^{-k+1}} f(\epsilon)\, d\epsilon \leq \left(r^{-k+1} - r^{-k}\right) f\left(r^{-k}\right).$$

■

Proof of Theorem 3. We use Theorem 6 and Lemma 7.

$$\begin{aligned}
&\sum_{l>0} 2^{-l+1} \sqrt{(\ln 2)\, len\,(t, 2^{-l}) + \ln \frac{2^{l+1}}{\delta}} \\
&\leq 4\sqrt{\ln 2} \left(\sum_{l>1} 2^{-l} \sqrt{len\,(t, 2^{-l+1})} \right) + \left(\sum_{l>0} 2^{-l+1} \sqrt{(l+1)} \right) \sqrt{\ln \frac{2}{\delta}} \\
&\leq 4\sqrt{\ln 2} \int_0^1 \sqrt{len\,(t,\epsilon)} d\epsilon + 4\sqrt{\ln \frac{2}{\delta}}.
\end{aligned}$$

■

4 From codes to probability measures

Given an instantaneous code, the construction of an appropriate probability measure is a rather direct consequence of the Kraft inequality.

Theorem 8 *Let (T,d) be a compact metric space of unit diameter and let $p \geq 1$. Then for any convergent code $c : T \to \alpha^{**}$ there exists a probability measure m of T such that for all $t \in T$*

$$\int_0^1 \left(\ln \frac{1}{m\,(B\,(t,\epsilon))} \right)^{1/p} d\epsilon \leq 2\,(\ln |\alpha|)^{1/p} \int_0^1 (len\,(c,t,\epsilon))^{1/p}\, d\epsilon + 2.$$

Proof. Define a measure m' on T by

$$m' = \sum_{l>0} 2^{-l} \sum_{v \in \mathcal{S}_{2^{-l}}} |\alpha|^{-len(v)} \delta_{\pi(v)},$$

where $\delta_{\pi(v)}$ is the unit mass (Dirac measure) concentrated at $\pi(v)$, the representative of v in $\hat{c}(v)$. By the Kraft inequality (5) we have $m'(T) \leq 1$, so there is a probability measure $m \geq m'$. Now for any $l \in \mathbb{N}$ and any $t \in T$ we have

$$\begin{aligned} m'\left(\hat{c}\left(c\left(t, 2^{-l}\right)\right)\right) &= \sum_{t' \in \hat{c}(c(t,2^{-l}))} \sum_{k>0} 2^{-k} \sum_{v \in \mathcal{S}_{2^{-k}}} |\alpha|^{-len(v)} \delta_{\pi(v)}(t') \\ &\geq 2^{-l} \sum_{t' \in \hat{c}(c(t,2^{-l}))} \sum_{v \in \mathcal{S}_{2^{-l}}} |\alpha|^{-len(v)} \delta_{\pi(v)}(t') \\ &= 2^{-l} |\alpha|^{-len\left(c\left(t,2^{-l}\right)\right)}. \end{aligned}$$

Also $\hat{c}\left(c\left(t, 2^{-l}\right)\right) \subseteq B\left(t, 2^{-l}\right)$, so

$$\begin{aligned} \ln \frac{1}{m\left(B\left(t, 2^{-l}\right)\right)} &\leq \ln \frac{1}{m\left(\hat{c}\left(c\left(t, 2^{-l}\right)\right)\right)} \leq \ln \frac{1}{m'\left(\hat{c}\left(c\left(t, 2^{-l}\right)\right)\right)} \\ &\leq (\ln |\alpha|)\, len\left(t, 2^{-l}\right) + l \ln 2. \end{aligned}$$

It follows that for any $t \in T$, using $\ln 2 \leq 1$ and $l^{1/p} \leq l$ and Lemma 7,

$$\begin{aligned} \int_0^1 \left(\ln \frac{1}{m\left(B\left(t,\epsilon\right)\right)}\right)^{1/p} d\epsilon &\leq \sum_{l>0} 2^{-l} \left(\ln \frac{1}{m\left(B\left(t, 2^{-l}\right)\right)}\right)^{1/p} \\ &\leq \sum_{l>0} 2^{-l} \left((\ln |\alpha|)\, len\left(t, 2^{-l}\right) + l\right)^{1/p} \\ &\leq 2 (\ln |\alpha|)^{1/p} \sum_{l>0} 2^{-l-1} \left(len\left(t, 2^{-l}\right)\right)^{1/p} + \sum_{l>0} 2^{-l} l^{1/p} \\ &\leq 2 (\ln |\alpha|)^{1/p} \int_0^1 \left(len\left(t,\epsilon\right)\right)^{1/p} d\epsilon + 2. \end{aligned}$$

■

5 An alternative formula

The result which we now prove includes Theorem 2 as the special case $p = 2$.

Theorem 9 *Let (T, d) be a finite metric space of unit diameter and $c : T \to \alpha^*$ an instantaneous code and $p \geq 1$. Then for every $t \in T$*

$$1 + \frac{2^{1/p-1}}{p} \sum_{k=1}^{\infty} k^{1/p-1} D_k(c, t) \leq \int_0^1 \left(len\left(c, t, \epsilon\right)\right)^{1/p} d\epsilon \leq 1 + \frac{1}{p} \sum_{k=1}^{\infty} k^{1/p-1} D_k(c, t).$$

Proof. Let λ denote the Lebesgue measure. We have

$$\int_0^1 \left(len\left(\epsilon,t\right)\right)^{1/p} d\epsilon = \sum_{k=1}^{\infty} k^{1/p} \lambda\left(\left\{\epsilon \leq 1 : len\left(\epsilon,t\right) = k\right\}\right).$$

The set $\left\{\epsilon \leq 1 : len\left(\epsilon,t\right) = k\right\}$ is an interval. Now $len\left(\epsilon,t\right) = k$ if and only if $D_k\left(c,t\right) < \epsilon \leq D_{k-1}\left(c,t\right)$, so

$$\lambda\left(\left\{\epsilon \leq 1 : len\left(\epsilon,t\right) = k\right\}\right) = D_{k-1}\left(c,t\right) - D_k\left(c,t\right)$$

and a summation by parts gives

$$\begin{aligned}
\int_0^1 \left(len\left(\epsilon,t\right)\right)^{1/p} d\epsilon &= \sum_{k=1}^{\infty} k^{1/p} \left(D_{k-1}\left(c,t\right) - D_k\left(c,t\right)\right) \\
&= 1 + \sum_{k=1}^{\infty} \left(\left(k+1\right)^{1/p} - k^{1/p}\right) D_k\left(c,t\right). \qquad (8)
\end{aligned}$$

Using concavity and smoothness of the function $s \mapsto s^{1/p}$ we get for $k \geq 1$

$$\begin{aligned}
\frac{2^{1/p-1}}{p} k^{1/p-1} &\leq \frac{1}{p}\left(\frac{k+1}{k}\right)^{1/p-1} k^{1/p-1} = \frac{1}{p}\left(k+1\right)^{1/p-1} \\
&\leq \left(k+1\right)^{1/p} - k^{1/p} \\
&\leq \frac{1}{p} k^{1/p-1} = \frac{1}{p} k^{-1/q}.
\end{aligned}$$

The result follows if we substitute these inequalities in (8). ■

6 From probability measures to codes

In [6] Michel Talagrand proves the following result:

Theorem 10 *Let m be a probability measure on T. There exists a sequence of nested partitions $\left(\mathcal{A}_k\right)_{k\geq 0}$ of T such that $\left|\mathcal{A}_0\right| = 1$, $\left|\mathcal{A}_k\right| \leq 2^{2^k}$ and*

$$\sup_{t\in T} \sum_{k\geq 0} 2^{k/p} D\left(A_k\left(t\right)\right) \leq C\left(p\right) \sup_{t\in T} \int_0^1 \left(\ln \frac{1}{m\left(B\left(t,\epsilon\right)\right)}\right)^{1/p} d\epsilon,$$

where $C\left(p\right)$ is a constant depending on p only.

We will use this result in a slightly modified form:

Theorem 11 *Let m be a probability measure on T. There exists a sequence of partitions $(\mathcal{A}_k)_{k\geq 0}$ of T such that $|\mathcal{A}_0| = 1$, $|\mathcal{A}_k| \leq 2^{2^k}$ and*

$$\forall t \in T,\ \sum_{k\geq 0} 2^{k/p} D\left(A_k\left(t\right)\right) \leq C_1\left(p\right) \int_0^1 \left(\ln \frac{1}{m\left(B\left(t,\epsilon\right)\right)}\right)^{1/p} d\epsilon + C_2\left(p\right),$$

where

$$\begin{aligned} C_1\left(p\right) &= \frac{16}{1-2^{-1/p}} \left(\frac{1}{\ln 2}\right)^{1/p} \\ C_2\left(p\right) &= \frac{4}{1-2^{-1/p}} \sum_{l\in\mathbb{N}_0} 2^{-l} \left(l+2\right)^{1/p}. \end{aligned}$$

The only strengthening here is that we require the inequality to be valid for every $t \in T$, not just for the suprema. But this is already implied by Talagrands proof in [6], which we essentially reproduce in the following. We do so without any claim of originality, for the benefit to the reader, to demonstrate that the strengthening doesn't cause any difficulties, and also to make the constants accessible.

On the other hand we weaken the result, because we do not require the partitions to be nested. In view of the equivalences demonstrated in this paper this shows that the requirement of nested partitions in [6] is superfluous: If the partitions are not nested, but as in Theorem 11 then by the proof of Theorem 14 we can construct an appropriate code, Theorem 8 gives us a corresponding probability measure, and Talagrands Theorem 10 produces the nested partitions.

Lemma 12 *If (T,d) is a compact metric space with probability measure m, $\epsilon > 0$ and $A \subseteq T$, then*

$$N\left(A, 2\epsilon\right) \inf_{t\in A} m\left(B\left(t,\epsilon\right)\right) \leq 1,$$

where $N\left(A,2\epsilon\right)$ is the smallest number of open balls of radius 2ϵ needed to cover A.

Proof. Let M be the 2ϵ-packing number

$$M = \sup\left\{n \in \mathbb{N} : \exists P \subseteq A, |P| = n, \forall x, y \in P, d\left(x,y\right) \geq 2\epsilon\right\}.$$

By compactness the set on the right hand side is bounded, so the supremum is attained for some $P^* \subseteq A$. If $z \in A$ then there exists $x \in P^*$ such that $d\left(x,z\right) < 2\epsilon$, otherwise we could enlarge P^* with z, contradicting its maximality. It follows that A can be covered by M open balls of radius 2ϵ centered in P, so that $N\left(A,2\epsilon\right) \leq M$. On the other hand, the open balls centered at the points in P^* with radius ϵ are disjoint, whence

$$1 \geq m\left(\bigcup_{t\in P^*} B\left(t,\epsilon\right)\right) = \sum_{x\in P^*} m\left(B\left(t,\epsilon\right)\right) \geq M \inf_{t\in A} m\left(B\left(t,\epsilon\right)\right).$$

■

Proof of Theorem 11. For any $t \in T$ and $k \in \mathbb{N}$ the set of integers l such that $m\left(B\left(t, 2^{-l}\right)\right) \geq 2^{l+1}2^{-2^k}$ is bounded and contains 0, so that a function $g_k : T \to \mathbb{N}_0$ is well defined by

$$g_k(t) = \max\left\{l \in \mathbb{Z} : m\left(B\left(t, 2^{-l}\right)\right) \geq 2^{l+1}2^{-2^k}\right\},$$

and $\left\{g_k^{-1}(\{l\})\right\}_{l\in\mathbb{N}_0}$ is a partition of T. Every member $t \in g_k^{-1}(\{l\})$ satisfies the two inequalities

$$m\left(B\left(t, 2^{-l}\right)\right) \geq 2^{l+1}2^{-2^k} \tag{9}$$

$$m\left(B\left(t, 2^{-l-1}\right)\right) < 2^{l+2}2^{-2^k}. \tag{10}$$

By (9) and Lemma 12 we have

$$N\left(g_k^{-1}(\{l\}), 2 \times 2^{-l}\right) \leq 2^{-l-1}2^{2^k},$$

so that $g_k^{-1}(\{l\})$ has a disjoint partition $\mathcal{A}_{kl}$, satisfying $|\mathcal{A}_{kl}| \leq 2^{-l-1}2^{2^k}$ and $D(A) \leq 2^{-l+2}$ for all $A \in \mathcal{A}_{kl}$. Then $\mathcal{A}_k = \bigcup_{l\in\mathbb{N}_0} \mathcal{A}_{kl}$ is a disjoint partition of T, of cardinality

$$|\mathcal{A}_k| \leq 2^{2^k} \sum_{l\in\mathbb{N}_0} 2^{-l-1} = 2^{2^k}.$$

For any $k \in \mathbb{N}$ and $t \in T$ we use $A_k(t)$ to denote the unique member of $\mathcal{A}_k$ containing t.

Now fix $t \in T$. We may thus assume $m(B(t, \epsilon)) > 0$ for every $\epsilon > 0$, since the inequality to be proved is otherwise trivial. Observe that $D(A_k(t)) \leq 2^{-g_k(t)+2}$, so that

$$\sum_{k=1}^{\infty} D(A_k(t))\, 2^{\frac{k}{p}} \leq 4\sum_{k=1}^{\infty} 2^{\frac{k}{p}} 2^{-g_k(t)} \tag{11}$$

Now consider for $l \in \mathbb{N}_0$ and $t \in T$ the set $K(t,l) = \{k \in \mathbb{N} : g_k(t) = l\}$ and observe that $K(t,l)$ may be empty, but that it is always bounded above since $m\left(B\left(t, 2^{-l-1}\right)\right) > 0$. We can therefore define

$$k^*(t,l) = \begin{cases} 0 & \text{if} \quad K(t,l) = \emptyset \\ \max K(t,l) & \text{if} \quad K(t,l) \neq \emptyset \end{cases}.$$

With these definitions the summation of a geometric series gives

$$\sum_{k=1}^{\infty} 2^{\frac{k}{p}} 2^{-g_k(t)} = \sum_{l:K(t,l)\neq\emptyset} 2^{-l} \sum_{k\in K(t,l)} 2^{\frac{k}{p}} \leq \frac{1}{1-2^{-1/p}} \sum_{l:K(t,l)\neq\emptyset} 2^{-l} 2^{\frac{k^*(t,l)}{p}}. \tag{12}$$

But if $K(t,l)$ is nonempty, then, since $g_{k^*(t,l)}(t) = l$, we get from (10) that

$$2^{k^*(t,l)} - (l+2) \leq \max\left\{0, 2^{k^*(t,l)} - l - 2\right\} \leq \frac{1}{\ln 2} \ln \frac{1}{m\left(B\left(t, 2^{-l-1}\right)\right)},$$

which implies that

$$\sum_{l:K(t,l)\neq\emptyset} 2^{-l} 2^{\frac{k^*(t,l)}{p}} \leq \sum_{l\in\mathbb{N}_0} 2^{-l} \left(\frac{1}{\ln 2} \ln \frac{1}{m\left(B\left(t, 2^{-l-1}\right)\right)} + (l+2) \right)^{1/p}$$

$$\leq \left(\frac{1}{\ln 2}\right)^{1/p} \sum_{l\in\mathbb{N}_0} 2^{-l} \left(\ln \frac{1}{m\left(B\left(t, 2^{-l-1}\right)\right)} \right)^{1/p} + \sum_{l\in\mathbb{N}_0} 2^{-l} (l+2)^{1/p}. \qquad (13)$$

Finally the estimates of Lemma 7 give

$$\sum_{l\in\mathbb{N}_0} 2^{-l} \left(\ln \frac{1}{m\left(B\left(t, 2^{-l-1}\right)\right)} \right)^{1/p} \leq 4 \int_0^1 \left(\ln \frac{1}{m\left(B\left(t,\epsilon\right)\right)} \right)^{1/p} d\epsilon. \qquad (14)$$

The result follows from combining the inequalities, (11), (12), (13) and (14). ■

We will construct a code from a probability measure by concatenating instantaneous codes for each partition and use the following Lemma to compare the representation of Theorem 9 to the sum in Theorem 11.

Lemma 13 *Let $k \in \mathbb{N}$ and $p \geq 1$. Then*

$$\sum_{l=2^{k-1}}^{2^k-1} l^{1/p-1} \leq 2^{\frac{k-1}{p}} p.$$

Proof. We have

$$\sum_{l=2^{k-1}}^{2^k-1} l^{1/p-1} \leq \int_{2^{k-1}-1}^{2^k-1} x^{1/p-1} dx = p\left(\left(2^k-1\right)^{1/p} - \left(2^{k-1}-1\right)^{1/p} \right).$$

Now for $a, b \geq 0$ we have $(a+b)^{1/p} \leq a^{1/p} + b^{1/p}$, so

$$\left(2^k-1\right)^{1/p} - \left(2^{k-1}-1\right)^{1/p} = \left(\left(2^{k-1}\right) + \left(2^{k-1}-1\right) \right)^{1/p} - \left(2^{k-1}-1\right)^{1/p} \leq 2^{\frac{k-1}{p}}.$$

■

Theorem 14 *Let (T,d) be a compact metric space of unit diameter and let $p \geq 1$. Then for every probability measure m on T and every finite alphabet α there exists a code $c : T \to \{0,1\}^{**}$ such that for all $t \in T$*

$$\int_0^1 \left(len\left(t,\epsilon\right)\right)^{1/p} d\epsilon \leq C_1(p) \int_0^1 \left(\ln \frac{1}{m\left(B\left(t,\epsilon\right)\right)} \right)^{1/p} d\epsilon + C_2(p),$$

where

$$\begin{aligned} C_1(p) &= \frac{2^{4+1/p}}{1-2^{-1/p}} \left(\frac{1}{\ln 2}\right)^{1/p} \\ C_2(p) &= \frac{p+1}{p} + 2^{1/p} + \frac{2^{2+1/p}}{1-2^{-1/p}} \sum_{l\in\mathbb{N}_0} 2^{-l} (l+2)^{1/p}. \end{aligned}$$

Proof. Let the sequence of partitions $(\mathcal{A}_k)_{k\geq 0}$ be chosen as in Theorem 11. Since $|\mathcal{A}_k| \leq 2^{2^k}$ we can, by simple binary counting, find a binary instantaneous code $c^{(k)} : \mathcal{A}_k \to \{0,1\}^*$ with $len\left(c^{(k)}(A)\right) \leq 2^k$ for all $k \geq 0$ and $A \in \mathcal{A}_k$. We now define our code $c : T \to \{0,1\}^{\mathbb{N}}$ to be the infinite concatenation

$$c(t) = c^{(1)}(A_1(t)) \circ c^{(2)}(A_2(t)) \circ ... \circ c^{(k)}(A_k(t)) \circ ..., \forall t \in T.$$

Fix $t \in T$. Let $k \in \mathbb{N}$ and

$$N = \sum_{l=1}^{k} len\left(c^{(l)} A_l(t)\right) \leq \sum_{l=1}^{k} 2^l \leq 2^{k+1}.$$

If $t' \in \hat{c}(p_N(c(t)))$ then the first N characters of $c(t')$ are the same as those of $c(t)$, so in particular $c^{(k)}(A_k(t')) = c^{(k)}(A_k(t))$. Since $c^{(k)}$ is instantaneous $A_k(t') = A_k(t)$ and therefore $t' \in A_k(t)$. So $\hat{c}(p_N(c(t))) \subseteq A_k(t)$ and, since $N \leq 2^{k+1}$,

$$D_{2^{k+1}}(c,t) \leq D_N(c,t) = D(\hat{c}(p_N(c(t)))) \leq D(A_k(t)).$$

It follows that

$$\begin{aligned}
\frac{1}{p}\sum_{k=1}^{\infty} k^{1/p-1} D_k(c,t) &\leq \frac{1}{p}\left(1 + \sum_{k=2}^{\infty}\sum_{l=2^{k-1}}^{2^k-1} l^{1/p-1} D_l(c,t)\right) \\
&\leq \frac{1}{p}\left(1 + \sum_{k=2}^{\infty} D(A_{k-2}(t)) \sum_{l=2^{k-1}}^{2^k-1} l^{1/p-1}\right) \\
&\leq \frac{1}{p} + 2^{1/p}\sum_{k=0}^{\infty} D(A_k(t))\, 2^{\frac{k}{p}} \\
&= \left(\frac{1}{p} + 2^{1/p}\right) + 2^{1/p}\sum_{k=1}^{\infty} D(A_k(t))\, 2^{\frac{k}{p}},
\end{aligned}$$

where we used Lemma 13 in the last inequality. To obtain the conclusion we combine this inequality with the conclusions of Theorem 9 and Theorem 11. ■

7 Glossary of notation

(T, d)	metric space
$D(A)$ for $A \subseteq T$	diameter of A
α	finite alphabet with cardinality $\|\alpha\|$
α^*	finite strings over α
$\alpha^{**} = \alpha^* \cup \alpha^{\mathbb{N}}$	finite or infinite strings over α
$p_k(v), v \in \alpha^{**}$	truncation of v to length at most k
$v \circ w$, for $v \in \alpha^*, w \in \alpha^{**}$	concatenation
$v \circ \alpha^{**} = \{v \circ w : w \in \alpha^{**}\}$	set of extensions of the finite string $v \in \alpha^*$
$c : T \to \alpha^{**}$	a code for T over α
$\hat{c}(v) = c^{-1}(v \circ \alpha^{**})$	subset of T described by $v \in \alpha^*$
$\mathcal{V}(c) = \{v : \hat{c}(v) \neq \emptyset\}$	vocabulary used by the code
$D_k(c,t) = D(\hat{c}(p_k(c(t))))$	ambiguity incurred by truncation of $c(t)$ to k
$len(c,t,\epsilon) = \min\{k : D_k(c,t) < \epsilon\}$	length required to describe t with precision ϵ
$c(t,\epsilon)$	the corresponding string
$\mathcal{S}_\epsilon(c) = \{c(t,\epsilon) : t \in T\}$	cross-section of vocabulary at resolution ϵ

References

[1] Audibert, J.-Y and Bousquet, O. Combining PAC-Bayesian and generic chaining bounds. *Journal of Machine Learning Research*, 8, p863-889, 2007.

[2] W. Bednorz. A theorem on majorizing measures. *The Annals of Probability*, Vol. 34, No.5, 1771-1781, 2006.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*, Wiley, 1991.

[4] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*, Springer, 1991.

[5] M. Talagrand. A simple proof of the majorizing measure theorem. *Geometric and Functional Analysis*. Vol 2, No.1, 1992.

[6] Talagrand, M. (2001). Majorizing measures without measures. *Ann. Probab.* 29 411–417, 2001.

[7] M. Talagrand. *The Generic Chaining. Upper and Lower Bounds for Stochastic Processes.* Springer, Berlin, 2005.

[8] A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*, Springer Verlag, 1996.