Majorizing Measures, Codes, and Information

Yifeng Chu and Maxim Raginsky, *Fellow, IEEE*

***Abstract*—The majorizing measure theorem of Fernique and Talagrand is a fundamental result in the theory of random processes. It relates the boundedness of random processes indexed by elements of a metric space to complexity measures arising from certain multiscale combinatorial structures, such as packing and covering trees. This paper builds on the ideas first outlined in a little-noticed preprint of Andreas Maurer to present an information-theoretic perspective on the majorizing measure theorem, according to which the boundedness of random processes is phrased in terms of the existence of efficient variable-length codes for the elements of the indexing metric space.**

***Index Terms*—majorizing measures, Gaussian processes, stochastic processes, variable-length codes on metric spaces.**

I. Introduction

Let (T, d) be a compact metric space, and let $(X_t)_{t\in T}$ be a zero-mean random process indexed by the elements of T and satisfying the increment condition

$$\mathbf{P}\Big[|X_s - X_t| \ge u d(s,t)\Big] \le 2e^{-u^2}, \quad s,t \in T,\, u > 0. \tag{1}$$

The question of boundedness of the expected supremum $\mathbf{E}[\sup_{t\in T} X_t]$ arises in a variety of applications, such as high-dimensional probability, convex geometry, functional analysis, theoretical computer science, mathematical statistics, and learning theory [1]. A fundamental result of Fernique [2] and Talagrand [3] states that, for any fixed but arbitrary $t_0 \in T$,

$$\mathbf{E}\Big[\sup_{t\in T}|X_t - X_{t_0}|\Big] \lesssim \inf_{\mu\in\mathcal{P}(T)} \sup_{t\in T} I_\mu(t), \tag{2}$$

where $\mathcal{P}(T)$ is the set of all Borel probability measures on T,

$$I_\mu(t) := \int_0^{\operatorname{diam}(T)} \sqrt{\log\frac{1}{\mu(B(t,\varepsilon))}}\,\mathrm{d}\varepsilon \tag{3}$$

is the *Fernique–Talagrand functional*, and $a \lesssim b$ indicates that $a \le Cb$ for some absolute constant $C > 0$. Here, $\operatorname{diam}(T) := \sup\{d(s,t) : s,t \in T\}$ is the (finite) diameter of T, $B(t,\varepsilon)$ denotes the closed ball of radius ε centered at t, and the logarithm is taken to base 2. Moreover, if $(X_t)_{t\in T}$ is a Gaussian process and T is endowed with the canonical metric

$$d(s,t) := \sqrt{\mathbf{E}[(X_s - X_t)^2]}, \tag{4}$$

then the bound (2) can be reversed [3]:

$$\mathbf{E}\Big[\sup_{t\in T}|X_t - X_{t_0}|\Big] \gtrsim \inf_{\mu\in\mathcal{P}(T)} \sup_{t\in T} I_\mu(t). \tag{5}$$

The inequality (2) says that the expected supremum of (X_t) is bounded provided there exists at least one $\mu \in \mathcal{P}(T)$, such that $\sup_{t\in T} I_\mu(t) < \infty$ (such probability measures are called *majorizing measures*); moreover, for Gaussian processes, the existence of a majorizing measure is equivalent to boundedness of expected suprema.

While the Fernique–Talagrand functional provides sharp bounds on the expected supremum of (X_t), the machinery of majorizing measures was eventually replaced by an equivalent combinatorial framework involving various multiscale discrete structures, such as covering and packing trees (see, e.g., [4] or [1, Ch. 6]). In a little-noticed preprint [5], A. Maurer has provided an alternative information-theoretic perspective on majorizing measures starting from the observation that the integrand in (3) has an information-theoretic flavor–indeed, $\log\frac{1}{\mu(B(t,\varepsilon))}$ is (roughly) proportional to the number of bits one would use to localize a point $t \in T$ up to a ball of radius ε when given a 'prior' probability measure $\mu \in \mathcal{P}(T)$. Building on this intuition, Maurer used the well-known correspondence between probability measures and uniquely decodable codes on a discrete (finite or countably infinite) alphabet [6, Sec. 5.5] to show that the Fernique–Talagrand functional can be bounded both from above and from below in terms of information-theoretic quantities pertaining to sequences of variable-length lossy codes on T that allow for arbitrarily accurate reconstruction of the points $t \in T$ using sufficiently long codewords, with the metric d playing the role of the fidelity criterion [7].

In this paper, we develop this perspective further by giving a more direct construction of multiscale variable-length codes that allows us to seamlessly recover a number of existing bounds on expected suprema. Moreover, we derive an information-theoretic version of the lower bound (5) for Gaussian processses that matches our upper bounds up to a multiplicative constant. In terms of applications, we expect these results to be of use in analyzing the statistical performance of machine learning algorithms through the lens of data compression and the Minimum Description Length principle [8], [9].

The remainder of the paper is organized as follows. Section II contains preliminary material related to variable-length codes (VLC's) on metric spaces and to the Fernique–Talagrand functional (3). In particular, the key notion of an *admissible sequence* of VLC's is introduced in Section II-A. Section III presents upper bounds on the expected supremum of a centered random process satisfying the subgaussian increment condition (1) in terms of admissible sequences of VLC's. The starting point is Theorem 2, which gives a high-probability upper bound on the supremum of the process (X_t) for all t in a small ball around an arbitrary point $t_0 \in T$ in terms of a certain functional on admissible sequences of VLC's.

The authors are with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois, Urbana, IL 61801, USA. Emails: ychu26@illinois.edu, maxim@illinois.edu.

This work was supported by the Illinois Institute for Data Science and Dynamical Systems (iDS^2), an NSF HDR TRIPODS institute, under award CCF-1934986. A preliminary version of this work was presented at the 2023 IEEE International Symposium on Information Theory.

Compared to an analogous result of Maurer [5, Theorem 6], which was stated for all $t \in T$, this local control allows for a more refined characterization of the expected supremum in Theorem 3. The estimate in Theorem 3 appears to be new. In Section III-A, we show that several existing upper bounds can be recovered using explicitly constructed sequences of VLC's. Then, in Section III-B, we use Theorem 3 to obtain an upper bound on the expected supremum in terms of an arbitrary uniquely decodable code (Theorem 7). In Section IV, we use the machinery of VLC's to obtain matching (up to constants) lower bounds for Gaussian processes with the canonical metric (4). This is a new result that further extends the scope of Maurer's preprint [5], which was only concerned with upper-bounding the expected supremum of a subgaussian process. The key ingredient is a recursive construction of a family of VLC's, which can serve as a coding-theoretic counterpart of the recursive partitioning scheme used by Bednorz [10] and Talagrand [11] in their proof of the lower bound for Gaussian processes. The two key results of that section are Theorem 8 and Theorem 9. Finally, in Section V, we present an alternative analysis using *random* VLC's and, in particular, obtain the counterparts of Theorems 7 and 8.

II. Preliminaries

A. Variable-length codes on metric spaces

Let a compact metric space (T, d) be given, with $\mathrm{diam}(T) = 1$. A *variable-length code* (VLC) on T is a pair (π, f), where $\pi : T \to T$ is a mapping whose image $\pi(T)$ is a discrete subset of T and $f : \pi(T) \to \{0,1\}^*$ is a uniquely decodable representation of the elements of $\pi(T)$ by finite-length binary strings. In particular, it satifsies the Kraft–McMillan inequality

$$\sum_{s \in \pi(T)} 2^{-\ell(f(s))} \le 1, \tag{6}$$

where $\ell(\cdot)$ denotes the length of a binary string. We say that a VLC (π, f) *operates at resolution* $\rho \ge 0$ if $d(t, \pi(t)) \le \rho$ for all $t \in T$ (if $\rho = 0$, then T is necessarily discrete, and (π, f) is lossless). A natural procedure for constructing such VLC's is to fix a ρ-partition $\mathcal{A}$ of T, i.e., a family of disjoint subsets $A \subseteq T$ of diameter at most ρ whose union is all of T; if $\rho > 0$, then it is always possible to find a finite partition by compactness of T. Then, picking a point s_A in each $A \in \mathcal{A}$, we take $\pi(t) := s_{A(t)}$, where $A(t)$ is the unique element of $\mathcal{A}$ that contains t. The mapping π constructed in this way is evidently idempotent, i.e., $\pi \circ \pi = \pi$. Finally, with $\pi(T) = \{s_A : A \in \mathcal{A}\}$, we can choose any mapping $f : \pi(T) \to \{0,1\}^*$ that satisfies (6). For instance, we can pick any $\mu \in \mathcal{P}(T)$ such that $\mu(A) > 0$ for all $A \in \mathcal{A}$ and let f be the Shannon code with codelengths $\ell(f(s_A)) = \lceil \log \frac{1}{\mu(A)} \rceil, A \in \mathcal{A}$.

We will be working with *sequences* of VLC's that operate at multiple scales. We will say that $\mathcal{C} = \{(\pi_k, f_k)\}_{k \in \mathbb{N}}$ is *admissible* if the following conditions are satisfied:

(i) for each $k \in \mathbb{N}$, π_k is idempotent and there exists a mapping $\varphi_{k|k+1} : T \to T$, such that $\pi_k = \varphi_{k|k+1} \circ \pi_{k+1}$;
(ii) (π_k, f_k) operates at resolution ρ_k, and $\rho_k \to 0$ as $k \to \infty$;
(iii) for each t and each k, $\ell(f_k \circ \pi_k(t)) \le \ell(f_{k+1} \circ \pi_{k+1}(t))$.

We also define (π_0, f_0) with $\pi_0(T) = \{t_0\}$ and $\ell(f_0(t_0)) = 0$ for a fixed but arbitrary point $t_0 \in T$. Condition (i) says that, for each t, the encoding $\pi_k(t)$ can be determined from any $\pi_m(t)$, $m > k$, without knowledge of t. We can take $\varphi_{k|k+1} = \pi_k$ without loss of generality since $\pi_k = \varphi_{k|k+1} \circ \pi_{k+1} = \varphi_{k|k+1} \circ \pi_{k+1} \circ \pi_{k+1} = \pi_k \circ \pi_{k+1}$. This implies that the partition $\mathcal{A}_{k+1}$ induced by π_{k+1} is a refinement of $\mathcal{A}_k$, and for each $t \in T$ the sets $A_k(t)$ (of diameter ρ_k) are nested, i.e., $T \equiv A_0(t) \supseteq A_1(t) \supseteq A_2(t) \supseteq \ldots$. Conditions (i) and (ii) imply that $\rho_k \searrow 0$ as $k \to \infty$. Finally, condition (iii) constrains the binary descriptions of each t to become more 'informative' as k increases.

A natural procedure for constructing an admissible sequence of VLC's is to start with an increasing sequence $(\mathcal{A}_k)_{k \ge 1}$ of partitions of T such that each $A \in \mathcal{A}_k$ has diameter at most ρ_k, with $\rho_k \searrow 0$ as $k \to \infty$. The mappings π_k are constructed exactly in the same manner as for a single partition, so that conditions (i)–(ii) will be satisfied automatically. To construct the binary encodings f_k satisfying condition (iii), choose some $\mu_1 \in \mathcal{P}(T)$ such that $\mu_1(A) > 0$ for all $A \in \mathcal{A}_1$; then, for each $k \ge 1$ and each $B \in \mathcal{A}_k$, choose a probability measure $\nu_{k+1}(\cdot|B) \in \mathcal{P}(T)$ such that $\nu_{k+1}(A|B) > 0$ for each $A \in \mathcal{A}_{k+1}(B) := \{A \in \mathcal{A}_{k+1} : A \subseteq B\}$ and 0 otherwise. This defines a sequence of probability measures $(\mu_k)_{k \ge 1}$ by

$$\mu_{k+1}(\cdot) = \sum_{B \in \mathcal{A}_k} \nu_{k+1}(\cdot|B)\mu_k(B), \tag{7}$$

and we can choose each f_k so that (ignoring the rounding issues) for each $t \in T$ and each k we have

$$\begin{aligned}
\ell(f_{k+1} \circ \pi_{k+1}(t)) &\approx \log \frac{1}{\mu_{k+1}(A_{k+1}(t))} \\
&= \log \frac{1}{\mu_k(A_k(t))} + \log \frac{1}{\nu_{k+1}(A_{k+1}(t)|A_k(t))} \\
&\ge \log \frac{1}{\mu_k(A_k(t))} \\
&\approx \ell(f_k \circ \pi_k(t)),
\end{aligned}$$

where in the second line we have used the fact that $A_{k+1}(t) \in \mathcal{A}_{k+1}(A_k(t))$, and therefore $\mu_{k+1}(A_{k+1}(t)) = \nu_{k+1}(A_{k+1}(t)|A_k(t))\mu_k(A_k(t))$.

In fact, there is no loss of generality to consider admissible sequences of VLC's constructed from a single $\mu \in \mathcal{P}(T)$ rather than from a recursively defined sequence $(\mu_k)_{k \ge 1}$. For convenience, we pick an arbitrary measure $\mu_0 \in \mathcal{P}(T)$.

Proposition 1: Given any increasing sequence $(\mathcal{A}_k)_{k \ge 0}$ of partitions of T and any $0 \le l < k$, for a sequence of probability measures $(\mu_k)_{k \ge 0}$ defined above,

$$\mu_k(A) = \mu_l(A), \quad \forall A \in \mathcal{A}_l.$$

Proof: We apply induction on k. When $k = 1$, $\mu_1(T) = \mu_0(T) = 1$. Assume the statement holds for $k - 1$. Then for any $l < k$ and any $A \in \mathcal{A}_l$, by writing $A = \dot{\bigcup}_{B \in \mathcal{A}_{k-1}(A)} B$, we have

$$\mu_k(A) = \sum_{B \in \mathcal{A}_{k-1}(A)} \mu_k(B)$$

$$= \sum_{B \in \mathcal{A}_{k-1}(A)} \mu_{k-1}(B)$$
$$= \sum_{B \in \mathcal{A}_{k-1}(A)} \mu_l(B)$$
$$= \mu_l(A),$$

where the second equality follows from the definition of μ_k in (7) and third equality is due to inductive hypothesis. ■

Thus, for large enough $K \in \mathbb{N}$ such that $\pi_K(t)$ is a singleton for all $t \in T$, if we take $\mu = \mu_K$,

$$\ell(f_k \circ \pi_k(t)) \approx \log \frac{1}{\mu_k(A_k(t))} = \log \frac{1}{\mu(A_k(t))}$$

for all $k \in \mathbb{N}$.

There is also an alternative 'top-down' procedure for constructing admissible sequences of VLC's, suggested by Maurer [5]. In what follows, we assume that T is discrete and let $f : T \to \{0,1\}^*$ be a uniquely decodable code on T. For any $w \in \{0,1\}^*$, let

$$\hat{f}(w) := \{t \in T : w \text{ is a prefix of } f(t)\}$$

and define the *vocabulary* of f as the set $\mathcal{V}(f)$ of all $w \in \{0,1\}^*$ for which $\hat{f}(w)$ is nonempty. Since f is uniquely decodable, $\ell(f(t)) \geq 1$ for all $t \in T$, but we still have $\hat{f}(\mathsf{e}) = T$, where e denotes the empty binary word. It is not hard to show that f is a prefix code if and only if $\hat{f}(f(t)) = \{t\}$ for every $t \in T$. For each $k \geq 0$, we define the truncation map $\mathsf{T}_k : \{0,1\}^* \to \{0,1\}^*$ as follows:

$$\mathsf{T}_k(w) = \mathsf{T}_k(w_1 w_2 \dots w_{\ell(w)}) := \begin{cases} \mathsf{e}, & k = 0 \\ w_1 \dots w_k, & 0 < k < \ell(w) \\ w, & k \geq \ell(w) \end{cases}. \tag{8}$$

We then define, for each $k \geq 0$ and each $t \in T$,

$$\Delta_k(f,t) := \operatorname{diam}\big(\hat{f}(\mathsf{T}_k \circ f(t))\big) = \sup\{d(s,s') : \mathsf{T}_k \circ f(t) \text{ is a prefix of } f(s), f(s')\}, \tag{9}$$

which gives the worst-case distortion that can be incurred by truncating the codeword $f(t)$ to length k. The sequence $(\Delta_k(f,t))_{k \geq 0}$ is evidently nonincreasing, and $\Delta_k(f,t) = 0$ for all $k \geq \ell(t)$ since f is uniquely decodable. Therefore, we can define the functions $\ell(f,\cdot,\cdot) : T \times [0,\infty) \to \mathbb{Z}_+$ by

$$\ell(f,t,\varepsilon) := \min\{k \in \mathbb{Z}_+ : \Delta_k(f,t) < \varepsilon\} \tag{10}$$

and $f : T \times [0,\infty) \to \mathcal{V}(f)$ by

$$f(t,\varepsilon) := \mathsf{T}_{\ell(f,t,\varepsilon)} \circ f(t). \tag{11}$$

In other words, $f(t,\varepsilon)$ is the operation of truncating $f(t)$ at the smallest length needed to localize t within a set of diameter smaller than ε. We then have the following:

Lemma 1 (Maurer [5]): Let $f : T \to \{0,1\}^*$ be a uniquely decodable code on a discrete metric space (T,d), and consider the family of sets

$$C_\varepsilon(f) := f(T,\varepsilon) = \{f(t,\varepsilon) : t \in T\} \subseteq \mathcal{V}(f), \ \varepsilon \geq 0. \tag{12}$$

Then we have the following:

1) For each $\varepsilon \geq 0$, the set $C_\varepsilon(f)$ is prefix-free, i.e., no element of $C_\varepsilon(f)$ is a prefix of another such element, and therefore
$$\sum_{v \in C_\varepsilon(f)} 2^{-\ell(v)} \leq 1. \tag{13}$$
2) For each $\varepsilon \geq 0$, the sets $\hat{f}(v)$, $v \in C_\varepsilon(f)$, form an ε-partition of T.
3) If $0 \leq \varepsilon \leq \varepsilon'$, then the partition induced by $C_\varepsilon(f)$ is a refinement of the partition induced by $C_{\varepsilon'}(f)$.

Proof: Fix some $s,t \in T$ and assume $f(t,\varepsilon)$ is a nontrivial prefix of $f(s,\varepsilon)$, i.e., $f(s,\varepsilon)$ is equal to $f(t,\varepsilon)w$, the concatenation of $f(t,\varepsilon)$ and some $w \in \{0,1\}^*$ with $\ell(w) \geq 1$. Then

$$f(s) = f(s,\varepsilon)v = f(t,\varepsilon)wv$$

for some $v \in \{0,1\}^*$. From this and from the definition of $\ell(f,\cdot,\cdot)$ it follows that $\ell(f,s,\varepsilon) \leq \ell(f,t,\varepsilon)$, i.e., $\ell(w) = 0$. Consequently, for any $s,t \in T$, either $f(s,\varepsilon) = f(t,\varepsilon)$ or neither of $f(s,\varepsilon) = f(t,\varepsilon)$ is a prefix of the other one. Thus, $C_\varepsilon(f)$ is prefix-free.

The prefix-free property of $C_\varepsilon(f)$ implies that the sets $\hat{f}(v)$, $v \in C_\varepsilon(f)$, are disjoint—indeed, if $t \in \hat{f}(v) \cap \hat{f}(w)$, then $f(t) = wu = vu'$ for some $u,u' \in \{0,1\}^*$, so either $w = v$ (and thus $\hat{f}(v) = \hat{f}(w)$) or one of them is a prefix of the other. Since the latter is impossible, $v = w$, and thus the sets $\hat{f}(v)$ are disjoint. Since each t belongs to $\hat{f}(f(t,\varepsilon))$, the sets $\hat{f}(v)$, $v \in C_\varepsilon(f)$ cover T. Therefore, $\{\hat{f}(v) : v \in C_\varepsilon(f)\}$ is a partition of T. Moreover, for any $v \in C_\varepsilon(f)$ there exists some $t \in T$ such that $v = f(t,\varepsilon) = \mathsf{T}_{\ell(f,t,\varepsilon)} \circ f(t)$. Then

$$\operatorname{diam}(\hat{f}(v)) = \operatorname{diam}(\hat{f}(\mathsf{T}_{\ell(f,t,\varepsilon)} \circ f(t)) = \Delta_{\ell(f,t,\varepsilon)} \leq \varepsilon$$

by definition of $\ell(f,t,\varepsilon)$. This shows that the sets $\hat{f}(v)$, $v \in C_\varepsilon(f)$, form an ε-partition of T.

Finally, if $v \in \mathcal{V}$ is such that $\operatorname{diam}(\hat{f}(v)) < \varepsilon$, we can extend the maps $\ell(f,\cdot,\varepsilon)$ and $f(\cdot,\varepsilon)$ to v by

$$\ell(f,v,\varepsilon) := \min\{k : \operatorname{diam}(\hat{f}(\mathsf{T}_k(v))) < \varepsilon\},$$
$$f(v,\varepsilon) := \mathsf{T}_{\ell(f,v,\varepsilon)}(v).$$

Let $0 \leq \varepsilon \leq \varepsilon'$ be given. Then we claim that the map $f(\cdot,\varepsilon')$ maps $C_\varepsilon(f)$ onto $C_{\varepsilon'}(f)$. This follows from the projection formula

$$f(t,\varepsilon') = f(f(t,\varepsilon),\varepsilon'). \tag{14}$$

To prove the latter, note that, since $f(t,\varepsilon')$ is a prefix of $f(t,\varepsilon)$ which is in turn a prefix of $f(t)$, we have $\ell(f,f(t,\varepsilon),\varepsilon') = \ell(f,t,\varepsilon')$. This implies that

$$f(f(t,\varepsilon),\varepsilon') = \mathsf{T}_{\ell(f,f(t,\varepsilon),\varepsilon')}(f(t,\varepsilon)) = \mathsf{T}_{\ell(f,t,\varepsilon')}(f(t)) = f(t,\varepsilon').$$

Consequently, $f(\cdot,\varepsilon') = f(f(\cdot,\varepsilon),\varepsilon')$, which means that $f(\cdot,\varepsilon') : C_\varepsilon(f) \to C_{\varepsilon'}(f)$ is a surjection. This, in turn, implies that, for each $v \in C_{\varepsilon'}(f)$, the set $\hat{f}(v)$ is a disjoint union of sets of the form $\hat{f}(w)$ for $w \in C_\varepsilon(f)$. ■

Using this lemma, we can construct an admissible sequence of VLC's from any uniquely decodable code and any decreasing sequence of resolution levels:

Theorem 1: Given any uniquely decodable code $f : T \to \{0,1\}^*$ and any sequence $(\rho_k)_{k\geq 1}$ of positive reals monotonically decreasing to 0 with $\rho_1 < \mathrm{diam}(T)$, there exists an admissible sequence of VLC's $\mathcal{C} = \{(\pi_k, f_k)\}_{k\geq 1}$ such that

$$f(t, \rho_k) = f_k \circ \pi_k(t).$$

Proof: Define the sets $C_0 = \{\mathsf{e}\}$ and $C_k := C_{\rho_k}(f)$ for $k \geq 1$. Then, by Lemma 1, the sets $C_1, C_2, \ldots$ are prefix-free and, for each k, there is a surjection $h_k : C_k \to C_{k-1}$ given by $h_k(v) := f(v, \rho_{k-1})$ with $\rho_0 = \mathrm{diam}(T)$, where we have extended the map $f(\cdot, \varepsilon)$ to $v \in \mathcal{V}(f)$ satisfying $\mathrm{diam}(\hat{f}(v)) < \varepsilon$, as in the proof of Lemma 1. For each $t \in T$, let $w_k(t) := f(t, \rho_k) \in C_k$. Then, from the projection formula (14) and from the definition of h_k it follows that

$$w_k(t) = f(f(t, \rho_{k+1}), \rho_k) = h_{k+1} \circ w_{k+1}(t).$$

Now, since $\mathcal{V}(f)$ is countable, by the countable axiom of choice there exists a map $g : \mathcal{V}(f) \to T$ such that $g(v) \in \hat{f}(v)$ for every $v \in \mathcal{V}(f)$.[1] Let $g_k := g|_{C_k}$. Since the sets $\{\hat{f}(v)\colon v \in C_k\}$, are disjoint, g_k is injective and hence invertible. Therefore, if we take $\pi_k := g_k \circ w_k$, then

$$\begin{aligned}
\pi_k(t) &= g_k \circ w_k(t) \\
&= g_k \circ h_{k+1} \circ w_{k+1}(t) \\
&= g_k \circ h_{k+1} \circ g_{k+1}^{-1} \circ \pi_{k+1}(t) \\
&=: \varphi_{k|k+1} \circ \pi_{k+1}(t).
\end{aligned}$$

Finally, define $f_k : \pi_k(T) \to \{0,1\}^*$ by $f_k \circ \pi_k(t) := w_k(t)$. This is a uniquely decodable code since $\{w_k(t) : t \in T\}$ is a prefix-free set, and moreover

$$\begin{aligned}
\ell(f_{k+1} \circ \pi_{k+1}(t)) &= \ell(f(t), \rho_{k+1}) \\
&\geq \ell(f(t), \rho_k) \\
&= \ell(f_k \circ \pi_k(t)).
\end{aligned}$$

Thus, the sequence $\mathcal{C} = \{(\pi_k, f_k)\}_{k\geq 1}$ of VLC's constructed in this manner is admissible. ■

[1] This procedure can be made constructive: Let $p : \{0,1\}^* \to T \cup \{\perp\}$, where $\perp$ is a separate 'false' symbol, be a program that accepts a binary string $v \in \{0,1\}^*$ and either returns the unique $t \in T$ such that $f(t) = v$ or $\perp$ if there is no such t. Indeed, let $\{v_n : n \in \mathbb{N}\}$ be the lexicographic ordering of $f(T)$; for each n, let $t_n \in T$ be such that $f(t_n) = v_n$. Such a t_n is unique because f is uniquely decodable. Given $v \in \{0,1\}^*$, it can be checked in finitely many steps whether $v = v_n$ for some n; if this is the case, $p(v) = t_n$. Otherwise, $p(v) =\perp$. Let $R \subseteq \mathcal{V}(f) \times T$ be the binary relation defined by $(v,t) \in R \Leftrightarrow t \in \hat{f}(v)$. The relation R is total, i.e., for any $v \in \mathcal{V}(f)$ there exists some $t \in T$ such that $(v,t) \in R$. The totality of R is witnessed by p since, for any $v \in \mathcal{V}(f)$, either $p(v) \neq\perp$ or we can keep adding bits to v until $p(w) \neq\perp$ for some $w \in \mathcal{V}(f)$ with prefix v. Since $v \in \mathcal{V}(f)$, this process is guaranteed to terminate in finitely many steps. The countable axiom of choice then says that there exists a function $g : \mathcal{V}(f) \to T$, such that

$$\mathrm{graph}(g) := \{(v, g(v)) : v \in \mathcal{V}(f)\} \subseteq R.$$

The function g can be straightforwardly constructed using the program p.

B. Some facts about the Fernique–Talagrand functional

The Fernique–Talagrand functional (3) can be expressed in terms of increasing sequences of partitions. To that end, following Bednorz [10], we define the following functional on pairs of probability measures $\mu, \nu \in \mathcal{P}(T)$:

$$M(\mu, \nu) := \int_T \int_0^1 \sqrt{\log \frac{1}{\mu(B(t,\varepsilon))}}\, \mathrm{d}\varepsilon\, \nu(\mathrm{d}t).$$

Then $I_\mu(t) = M(\mu, \delta_t)$, where δ_t is the Dirac measure centered at t, and the bounds (2) and (5) can be expressed as

$$\mathbf{E}\Big[\sup_{t\in T} X_t\Big] \lesssim \inf_\mu \sup_t M(\mu, \delta_t) = \inf_\mu \sup_\nu M(\mu, \nu)$$

and (for Gaussian processes on T with the induced metric)

$$\mathbf{E}\Big[\sup_{t\in T} X_t\Big] \gtrsim \inf_\mu \sup_t M(\mu, \delta_t) = \inf_\mu \sup_\nu M(\mu, \nu).$$

We then have the following upper bounds on $I_\mu(t)$ and $M(\mu,\nu)$ [10]: Let $\mathcal{A} = (\mathcal{A}_k)_{k\geq 1}$ be an increasing sequence of partitions of T, such that $\mathrm{diam}(A) \leq r^{-k}$ for all $A \in \mathcal{A}_k$ and all k for some $r > 1$. Then, setting $\mathcal{A}_0 = \{T\}$, we have

$$I_\mu(t) \leq \sum_{k>0} r^{-k+1} \sqrt{\log \frac{\mu(A_{k-1}(t))}{\mu(A_k(t))}}$$

and

$$M(\mu,\nu) \leq \sum_{k>0} r^{-k+1} \sum_{B\in\mathcal{A}_{k-1}} \sum_{A\in\mathcal{A}_k(B)} \nu(A) \sqrt{\log \frac{\mu(B)}{\mu(A)}}.$$

We can express these bounds in information-theoretic terms. Let us define, for any $k \geq 1$ and any $B \in \mathcal{A}_j$ with $j < k$, the entropy and the cross-entropy

$$\begin{aligned}
H_{\mathcal{A}_k|B}(\mu) &:= -\sum_{A\in\mathcal{A}_k} \mu(A|B) \log \mu(A|B), \\
H_{\mathcal{A}_k|B}(\mu,\nu) &:= -\sum_{A\in\mathcal{A}_k} \mu(A|B) \log \nu(A|B).
\end{aligned}$$

Using Jensen's inequality and the bound $\inf_\mu \sup_t M(\mu, \delta_t) \leq \sup_\mu M(\mu,\mu)$ [10], [12], we can obtain the following:

Proposition 2: For any $r > 1$,

$$\mathbf{E}\Big[\sup_{t\in T} X_t\Big] \lesssim \inf_{\mathcal{A}} \sup_\mu \sum_{k>0} r^{-k+1} \sum_{B\in\mathcal{A}_{k-1}} \mu(B) \sqrt{H_{\mathcal{A}_k|B}(\mu)},$$

where the infimum is over all increasing sequences $\mathcal{A} = (\mathcal{A}_k)_{k\geq 1}$ of partitions of T with $\mathrm{diam}(A) \leq r^{-k}$ for all $A \in \mathcal{A}_k, k \geq 1$.

III. The upper bound and some consequences

To keep things simple, we will assume for the remainder of the paper that T is finite. For infinite T, one can define

$$\mathbf{E}\Big[\sup_{t\in T} X_t\Big] := \sup\left\{ \mathbf{E}\Big[\sup_{t\in S} X_t\Big] : S \subset T, S \text{ finite} \right\}$$

following Talagrand [1], thus reducing the analysis to the finite case.

Let $\mathcal{C} = \{(\pi_k, f_k)\}_{k\in\mathbb{N}}$ be an admissible sequence of VLC's on T, where each (π_k, f_k) operates at resolution ρ_k. Let also $\boldsymbol{p} = (p_k)_{k\geq 1}$ be a positive probability mass function on $\mathbb{N}$ and define the functional

$$\bar{\sigma}_{(\mathcal{C},\boldsymbol{p})}(t) := \sum_{k>0} \rho_{k-1} \sqrt{\ln \frac{2^{\ell(f_k\circ\pi_k(t))+1}}{p_k}}, \tag{15}$$

where we have set $\rho_0 = \mathrm{diam}(T) \equiv 1$. The following is a generalization of [5, Thm. 6]:

Theorem 2: Let $(X_t)_{t\in T}$ be a centered random process satisfying the increment condition (1). Let $(\mathcal{C}, \boldsymbol{p})$ be given, such that $\bar{\sigma}_{(\mathcal{C},\boldsymbol{p})}(t)$ is finite for all $t \in T$. Fix an arbitrary point $t_0 \in T$. Then, for any $u > 0$, $\varepsilon > 0$, and any $A \subset B(t_0, \varepsilon)$,

$$|X_t - X_{t_0}| \leq \bar{\sigma}_{(\mathcal{C},\boldsymbol{p})}(t) + uR(\varepsilon), \qquad \text{for all } t \in A$$

with probability at least $1 - e^{-u^2}$ where $R(\varepsilon) = \sum_{k\geq\bar{k}} \rho_{k-1}$ with $\bar{k} \in \mathbb{N}$ chosen so that $\rho_{\bar{k}} < \varepsilon \leq \rho_{\bar{k}-1}$.

Proof: Since T is finite, for each $t \in T$ there exists some $k = k(t) \in \mathbb{N}$, such that $t = \pi_k(t) = \pi_{k+1}(t) = \dots$. Given $\mathcal{C}$, there is a natural choice of a sequence of admissible VLCs for A. Let $\bar{k} \in \mathbb{N}$ be such that $\rho_{\bar{k}} < \varepsilon \leq \rho_{\bar{k}-1}$. For $k \geq \bar{k}$, since π_k operates at resolution ρ_k, $\mathrm{card}(\pi_k(A))$ is at least two. It is clear that $(\pi^A_k := \pi_k|_A, f_k|_A)_{k\geq\bar{k}}$ is admissible. Let $\pi^A_{\bar{k}-1}(t) = t_0$ for $t \in A$. For $k \geq \bar{k}$, let $S_k := \pi^A_k(A)$ and define $\xi_k : S_k \to \mathbb{R}$ by

$$\xi_k(s) := d(s, \pi^A_{k-1}(s)) \sqrt{\ln \frac{2^{\ell(f_k(s))+1}}{p_k} + u^2}$$

and the event $E_k := \big\{\exists s \in S_k \text{ such that } |X_s - X_{\pi^A_{k-1}(s)}| > \xi_k(s)\big\}$. Then for any $t \in A$ we can form the chaining decomposition [1, p. 19]

$$X_t - X_{t_0} = \sum_{k\geq\bar{k}} (X_{\pi^A_k(t)} - X_{\pi^A_{k-1}(t)}). \tag{16}$$

Using the increment condition (1) and the Kraft–McMillan inequality, we have

$$\begin{aligned}
\mathbf{P}[E_k] &\leq \sum_{s\in S_k} \mathbf{P}[|X_s - X_{\pi^A_{k-1}(s)}| > \xi_k(s)] \\
&\leq 2 \sum_{s\in S_k} \exp\left(- \frac{\xi_k^2(s)}{d^2(s, \pi^A_{k-1}(s))} \right) \\
&= 2 \sum_{s\in S_k} \exp\left(- \ln \frac{2^{\ell(f_k(s))+1}}{p_k} - u^2 \right) \\
&= p_k e^{-u^2} \sum_{s\in S_k} 2^{-\ell(f_k(s))} \\
&\leq p_k e^{-u^2}.
\end{aligned}$$

Then

$$\begin{aligned}
&\mathbf{P}\left[\exists t \in A \text{ such that } |X_t - X_{t_0}| > \sum_{k\geq\bar{k}} \xi_k(\pi^A_k(t))\right] \\
&\leq \mathbf{P}\Bigg[\exists t \in A \text{ such that} \\
&\qquad \sum_{k\geq\bar{k}} \left(|X_{\pi^A_k(t)} - X_{\pi^A_{k-1}(t)}| - \xi_k(\pi^A_k(t)) \right) > 0 \Bigg] \\
&\leq \mathbf{P}\left[\bigcup_k E_k\right] \\
&\leq e^{-u^2},
\end{aligned}$$

where the last step follows from the assumption on the weights p_k. Thus, since $d(\pi^A_k(t), \pi^A_{k-1}(t)) \leq \rho_{k-1}$ for $k > \bar{k}$ and $d(\pi^A_{\bar{k}}(t), \pi^A_{\bar{k}-1}(t)) = d(\pi^A_{\bar{k}}(t), t_0) \leq \varepsilon \leq \rho_{k-1}$,

$$|X_t - X_{t_0}| \leq \bar{\sigma}_{(\mathcal{C},\boldsymbol{p})}(t) + u \sum_{k\geq\bar{k}} \rho_{k-1}, \qquad \text{for all } t \in A$$

with probability at least $1 - e^{-u^2}$. This completes the proof. ■

Corollary 1: Under the same assumptions as in Theorem 2,

$$\mathbf{E}\Big[\sup_{t\in T} X_t\Big] \leq 2 \inf_{(\mathcal{C},\boldsymbol{p})} \sup_{t\in T} \bar{\sigma}_{(\mathcal{C},\boldsymbol{p})}(t).$$

Proof: Applying Thm. 2 with $A = T$, $\varepsilon = \rho_0 = 1$, and $\bar{k} = 1$ yields that, for any $u > 0$,

$$|X_t - X_{t_0}| \leq \bar{\sigma}_{(\mathcal{C},\boldsymbol{p})}(t) + u \sum_{k>0} \rho_{k-1}, \quad \forall t \in T$$

with probability at least $1 - e^{-u^2}$. Since $\ell(f_k \circ \pi_k(t)) \geq 1$ for $k \geq 1$ by the unique decodability of $f_k : S_k \to \{0, 1\}^*$, it follows that $\frac{2^{\ell(f_k\circ\pi_k(t))+1}}{p_k} \geq 4$, and thus

$$\sum_{k>0} \rho_{k-1} \leq \bar{\sigma}_{(\mathcal{C},\boldsymbol{p})}(t) \quad \forall t \in T,$$

which implies that, for any $u > 0$,

$$|X_t - X_{t_0}| \leq \bar{\sigma}_{(\mathcal{C},\boldsymbol{p})}(t)(u + 1), \qquad \forall t \in T$$

with probability at least $1 - e^{-u^2}$. By letting $\bar{a} := \sup_{t\in T} \bar{\sigma}_{(\mathcal{C},\boldsymbol{p})}(t)$ and $\bar{Z} := \sup_{t\in T} |X_t - X_{t_0}|$, the inequality above can be rewritten as $\mathbf{P}[\bar{Z} > \bar{a} + u] \leq e^{-u^2/\bar{a}^2}$ for all $u > 0$, so

$$\begin{aligned}
\mathbf{E}\Big[\sup_{t\in T} X_t\Big] &= \mathbf{E}\Big[\sup_{t\in T} (X_t - X_{t_0})\Big] \\
&\leq \mathbf{E}[\bar{Z}] \\
&= \int_0^\infty \mathbf{P}[\bar{Z} > r]\,\mathrm{d}r \\
&= \int^{\bar{a}} \mathbf{P}[\bar{Z} > r]\,\mathrm{d}r + \int_{\bar{a}}^\infty \mathbf{P}[\bar{Z} > r]\,\mathrm{d}r \\
&\leq \bar{a} + \int_0^\infty e^{-u^2/\bar{a}^2}\,\mathrm{d}u \leq 2\bar{a}.
\end{aligned}$$

Now take the infimum over all $(\mathcal{C}, \boldsymbol{p})$. ■

To further exploit the properties of admissible sequences of VLC's, we consider the setting $p_k = 2^{-k}$ and $\rho_k = r^{-k}$ for some $r \geq 2$. Then

$$\bar{\sigma}_{\mathcal{C},\boldsymbol{p}}(t) \leq \sum_{k>0} \rho_{k-1} \sqrt{\ell(f_k \circ \pi_k(t))} + \sum_{k>0} \rho_{k-1} \sqrt{\ln \frac{2}{p_k}}$$

$$\leq \sum_{k>0} r^{-k+1}\sqrt{\ell(f_k\circ\pi_k(t))}+3$$
$$\lesssim \sum_{k>0} r^{-k+1}\sqrt{\ell(f_k\circ\pi_k(t))} =: \sigma_{\mathcal{C}}(t).$$

Thus, in what follows, we can focus on $\sigma_{\mathcal{C}}$. Moreover, with this choice of (ρ_k), the decaying factor $R(\varepsilon)$ of the tail estimate for $\sup_{t\in A}|X_t - X_{t_0}|$ with $A\subset B(t_0,\varepsilon)$ in Theorem 2 can be controlled by the radius ε up to some constant dependent on r. Specifically:

Corollary 2: For some $r\geq 2$, under the same assumptions as in Theorem 2 with $\rho_k = r^{-k}$ for $\mathcal{C}$,

$$\mathbf{P}\Big[\sup_{t\in A}|X_t - X_{t_0}| \geq c\big(\sup_{t\in A}\sigma_{\mathcal{C}}(t) + ur\varepsilon\big)\Big] \leq e^{-u^2} \tag{17}$$

where $c>0$ is some universal constant.

Proof: By the choice of $\bar{k}$, $R(\varepsilon) = \sum_{k\geq\bar{k}} r^{-k+1} \lesssim r\varepsilon$ and, by the choice of (ρ_k, p_k), $\bar{\sigma}_{\mathcal{C},\boldsymbol{p}}(t) \lesssim \sigma_{\mathcal{C}}(t)$. ■

This result is the key element for obtaining an upper bound on $\mathbf{E}[\sup_{t\in T} X_t]$ for X_t satisfying (1) in terms of an *expectation* of $\sigma_{\mathcal{C}}$ w.r.t. an appropriately chosen probability measure on T:

Theorem 3: Let $(X_t)_{t\in T}$ be a centered random process satisfying the increment condition (1). Let $\nu\in\mathcal{P}(T)$ be the probability measure, under which

$$\nu(t) = \mathbf{P}\Big[X_t = \sup_{s\in T} X_s\Big], \qquad \text{for all } t\in T.$$

Given $\mathcal{C}$ such that $\sigma_{\mathcal{C}}(t)<\infty$ for all $t\in T$,

$$\mathbf{E}\Big[\sup_{t\in T} X_t\Big] \lesssim r\int_T \sigma_{\mathcal{C}}(t)\nu(\mathrm{d}t). \tag{18}$$

Proof: Fix $t_0\in T$. Let $Y_t := |X_t - X_{t_0}|$, $\alpha(t) := d(t,t_0)$, and

$$J := \int_T \alpha(t)\nu(\mathrm{d}t),$$
$$I := J + \int_T \alpha(t)\sqrt{\log(1+\alpha(t)/J)}\nu(\mathrm{d}t).$$

Moreover, define the sets

$$U_0 := \big\{t\in T : \alpha(t)\leq 2J\big\}$$
$$U_n := \Big\{t\in T : 2^nJ < \alpha(t) \leq 2^{n+1}J\Big\} \text{ for } n\geq 1$$
$$V_{0,n} := U_n\cap\big\{t\in T : \sigma_{\mathcal{C}}(t) < J\big\}$$
$$V_{i,n} := U_n\cap\Big\{t\in T : J2^{i-1}\leq\sigma_{\mathcal{C}}(t) < J2^i\Big\} \text{ for } i\geq 1.$$

Let the sets $\Omega_{i,n} = \big\{\arg\sup_{t\in T} X_t \in V_{i,n}\big\}$ be such that $(\Omega_{i,n})_{i,n\geq 0}$ forms a partition of Ω by breaking ties suitably. Then

$$\mathbf{E}\Big[\sup_{t\in T} Y_t\Big] = \sum_{i,n\geq 0}\mathbf{E}[\mathbf{1}_{\Omega_{i,n}}\sup_{t\in T} Y_t]$$
$$= \sum_{i,n\geq 0}\mathbf{E}\Big[\mathbf{1}_{\Omega_{i,n}}\sup_{t\in V_{i,n}} Y_t\Big].$$

For any $v>0$,

$$\mathbf{E}\Big[\mathbf{1}_{\Omega_{i,n}}\sup_{t\in V_{i,n}} Y_t\Big]$$
$$= \int_0^\infty \mathbf{P}\Big[\Omega_{i,n}\cap\{\sup_{t\in V_{i,n}} Y_t > u\}\Big]\,\mathrm{d}u$$
$$\leq v\nu(V_{i,n}) + \int_v^\infty \mathbf{P}\Big[\sup_{t\in V_{i,n}} Y_t > u\Big]\,\mathrm{d}u.$$

Taking $v = c\big(\sup_{t\in V_{i,n}}\sigma_{\mathcal{C}}(t) + r2^{n+1}J\phi(\nu(V_{i,n}))\big)$ with c being the same constant as in Corollary 2, we have

$$\int_v^\infty \mathbf{P}\Big[\sup_{t\in V_{i,n}} Y_t > u\Big]\,\mathrm{d}u \leq cr2^{n+1}J\int_{\phi(\nu(V_{i,n}))}^\infty e^{-u^2}\,\mathrm{d}u$$
$$\lesssim r2^nJ\nu(V_{i,n}),$$

where the first inequality uses Corollary 2 with $\varepsilon = 2^{n+1}J$ and $A = V_{i,n}\subset B(t_0, 2^{n+1}J)$. Therefore,

$$\mathbf{E}\Big[\mathbf{1}_{\Omega_{i,n}}\sup_{t\in V_{i,n}} Y_t\Big]$$
$$\lesssim \big(\sup_{t\in V_{i,n}}\sigma_{\mathcal{C}}(t) + r2^{n+1}J\phi(\nu(V_{i,n}))\big)\nu(V_{i,n}) + r2^nJ\nu(V_{i,n})$$
$$\lesssim 2^iJ\nu(V_{i,n}) + r2^nJ\big(\nu(V_{i,n})\phi(\nu(V_{i,n})) + \nu(V_{i,n})\big).$$

The function $\phi(x) = \sqrt{\log(1/x)}$ satisfies the following inequality for $0<a,b\leq 1$ [3]:

$$a\phi(a) \leq a\phi(b) + (2e)^{-1/2}b. \tag{19}$$

Using this with $a = \nu(V_{i,n})$ and $b = e^{-i-n}$, we have

$$\nu(V_{i,n})\phi(\nu(V_{i,n}))$$
$$\leq \nu(V_{i,n})(i+n)^{1/2} + (2e)^{-1/2}e^{-i-n}$$
$$\leq 2\nu(V_{i,n})n^{1/2} + \nu(V_{i,n})2^{i-n} + (2e)^{-1/2}e^{-i-n},$$

where the last inequality is due to the fact that $(i+n)^{1/2}\leq 2n^{1/2} + 2^{i-n}$ for $i,n\geq 0$. Hence,

$$\mathbf{E}\Big[\mathbf{1}_{\Omega_{i,n}}\sup_{t\in V_{i,n}} Y_t\Big] \lesssim rJ(2^i\nu(V_{i,n}) + 2^nn^{1/2}\nu(V_{i,n}) + e^{-i-n}).$$

Using definition of $U_{i,n}, V_{i,n}$ again, we have

$$J\sum_{i,n\geq 0} 2^i\nu(V_{i,n}) \lesssim \int\sigma_{\mathcal{C}}(t)\nu(\mathrm{d}t)$$
$$J\sum_{i,n\geq 0} 2^nn^{1/2}\nu(V_{i,n}) \lesssim \int\alpha(t)\sqrt{\log(1+\alpha(t)/J)}\nu(\mathrm{d}t).$$

Combining the inequalities above, we obtain

$$\mathbf{E}\Big[\sup_{t\in T} Y_t\Big] \lesssim r\bigg(I + \int\sigma_{\mathcal{C}}(t)\nu(\mathrm{d}t)\bigg).$$

Notice that, by $\log(1+x)\leq x$ and Cauchy-Schwarz,

$$\int\alpha(t)\sqrt{\log(1+\alpha(t)/J)}\nu(\mathrm{d}t) \leq \int\alpha(t)\sqrt{\alpha(t)/J}\nu(\mathrm{d}t)$$
$$\leq \sqrt{\int\alpha^2(t)\nu(\mathrm{d}t)}$$
$$\leq \mathrm{diam}(T,d).$$

Then $I\lesssim\int\sigma_{\mathcal{C}}(t)\nu(\mathrm{d}t)$, which concludes the proof. ■

A. *Some consequences*

We start by giving the following estimate on $\sigma_{\mathcal{C}}$:

Theorem 4: Let $\mathcal{C} = \{(\pi_k, f_k)\}_{k\geq 1}$ be an admissible sequence of VLC's with $\rho_k = r^{-k}$ for some $r \geq 2$. Then

$$\sigma_{\mathcal{C}}(t) \lesssim \sum_{k>0} r^{-k+1}\sqrt{\ell(f_k \circ \pi_k(t)) - \ell(f_{k-1} \circ \pi_{k-1}(t))}. \tag{20}$$

Proof: Let $\ell_k(t) := \ell(f_k \circ \pi_k(t))$, with $\ell_0(t) \equiv 0$. By admissibility of $\mathcal{C}$, $(\ell_k(t))_{k\geq 1}$ is a nondecreasing sequence of positive integers. Therefore,

$$\begin{aligned}\sigma_{\mathcal{C}}(t) &= \sum_{k>0} r^{-k+1}\sqrt{\ell_k(t)} \\ &\leq \sum_{k>0} r^{-k+1}\sqrt{\ell_k(t) - \ell_{k-1}(t)} + \sum_{k>0} r^{-k+1}\sqrt{\ell_{k-1}(t)} \\ &= \sum_{k>0} r^{-k+1}\sqrt{\ell_k(t) - \ell_{k-1}(t)} + r^{-1}\sigma_{\mathcal{C}}(t),\end{aligned}$$

which, together with $(1 - r^{-1})^{-1} \leq 2$, gives (20). ∎

For each $k > 0$, the quantity under the square root in the upper bound (20) is the extra codelength needed to *refine* the description $\pi_{k-1}(t)$ at resolution r^{-k+1} to a more accurate one $\pi_k(t)$ at resolution r^{-k}. Several existing results from the literature can now be obtained as special cases.

First, we can recover the upper bound in terms of *labeled nets* [4], [13]. A labeled net on (T, d) is a pair $(\mathcal{A}, L)$, where $\mathcal{A} = (\mathcal{A}_k)_{k\geq 1}$ is an increasing sequence of partitions of T, such that $\operatorname{diam}(A) \leq r^{-k}$ for all $A \in \mathcal{A}_k$ and all $k \geq 1$, and $L : \mathcal{A} \to \mathbb{N}$ is a *labeling function* satisfying the condition $\{L(A) : A \in \mathcal{A}_{k+1}(B)\} = \{1, \dots, |\mathcal{A}_{k+1}(B)|\}$ for all $B \in \mathcal{A}_k$ and all k.

Theorem 5: For each labeled net $(\mathcal{A}, L)$, there exists an admissible sequence of VLC's, such that

$$\sigma_{\mathcal{C}}(t) \leq c_1 \sum_{k>0} r^{-k+1}\sqrt{\log L(A_k(t))} + c_2, \tag{21}$$

where $c_1, c_2 > 0$ are universal constants.

Proof: We will use the construction described in Section II-A. To that end, define a subprobability measure

$$\mu_1'(\cdot) = \frac{6}{\pi^2}\sum_{A\in\mathcal{A}_1}\frac{1}{L^2(A)|A|}\mathbf{1}_A(\cdot)$$

so that $\mu_1'(A) = \frac{6}{\pi^2 L^2(A)}$ for each $A \in \mathcal{A}_1$, and, for each k and $B \in \mathcal{A}_k$, a subprobability measure

$$\nu_{k+1}'(\cdot|B) := \frac{6}{\pi^2}\sum_{A\in\mathcal{A}_{k+1}(B)}\frac{1}{L^2(A)|A|}\mathbf{1}_A(\cdot),$$

so that $\nu_{k+1}'(A|B) = \frac{6}{\pi^2 L^2(A)}$ for every $A \in \mathcal{A}_{k+1}(B)$. Consequently, there exist probability measures $\mu_1(\cdot) \geq \mu_1'(\cdot)$ and $\nu_{k+1}(\cdot|B) \geq \nu_{k+1}'(\cdot|B)$. Then, with μ_k defined recursively via (7), we have

$$\begin{aligned}&\ell(f_{k+1}\circ\pi_{k+1}(t)) - \ell(f_k\circ\pi_k(t)) \\ &\leq \log\frac{1}{\mu_{k+1}(A_{k+1}(t))} - \log\frac{1}{\mu_k(A_k(t))} + 1 \\ &= \log\frac{1}{\nu_{k+1}(A_{k+1}(t))|A_k(t))} + 1 \\ &\leq \log\frac{1}{\nu_{k+1}'(A_{k+1}(t)|A_k(t))} + 1 \\ &\leq 2\log L(A_{k+1}(t)) + \log\frac{\pi^2}{6} + 1.\end{aligned}$$

The inequality (21) then follows from Theorem 4. ∎

In a similar way, we can recover a bound of Bednorz [10] that involves an increasing partition just like above:

Theorem 6: For each sequence of partitions as in Theorem 5 and each $\mu \in \mathcal{P}(T)$, there exists an admissible sequence $\mathcal{C}$ of VLC's, such that

$$\sigma_{\mathcal{C}}(t) \leq \sum_{k>0} r^{-k+1}\sqrt{\log\frac{\mu(A_{k-1}(t))}{\mu(A_k(t))} + 2}. \tag{22}$$

Proof: Without loss of generality, we may assume that $\mu(A) > 0$ for all $A \in \mathcal{A}_k$ and all k. We follow the same steps as in the proof of Theorem 4, but take $\mu_1 = \mu$ and $\nu_{k+1}(\cdot|B) = \mu(\cdot|B)$ for all k and all $B \in \mathcal{A}_k$. Then, for every $B \in \mathcal{A}_k$ and $A \in \mathcal{A}_{k+1}(B)$, $\nu_{k+1}(A|B) = \mu(A|B) = \frac{\mu(A)}{\mu(B)}$, and

$$\begin{aligned}\ell(f_{k+1}\circ\pi_{k+1}(t)) - \ell(f_k\circ\pi_k(t)) &\leq \log\frac{\mu(A_k(t))}{\mu(A_{k+1}(t))} + 1 \\ &= \log\frac{\mu(A_k(t))}{\mu(A_{k+1}(t))} + 1.\end{aligned}$$

Summing over k, we get (22). ∎

B. Upper bounds via uniquely decodable codes

In II-A, we described the procedure for constructing an admissible sequence of VLC's from a single uniquely decodable code on T. Using this procedure, we can translate previously stated bounds into ones that are based on a single prescribed code. For any uniquely decodable code $f : T \to \{0,1\}^*$, let

$$\sigma_f(t) := \int_0^{\operatorname{diam}(T)}\sqrt{\ell(f,t,\varepsilon)}\,\mathrm{d}\varepsilon,$$

where $\ell(f,t,\varepsilon)$ is defined in (10).

Proposition 3: Given a uniquely decodable code f and $r \geq 2$, there exists an admissible sequence $\mathcal{C}$ of VLC's with $\rho_k = r^{-k}$ such that

$$r^{-2}\sigma_{\mathcal{C}}(t) \leq \sigma_f(t) \leq \sigma_{\mathcal{C}}(t), \qquad \forall\, t \in T.$$

Proof: By definitions of $\ell(f,t,\varepsilon)$ and $f(t,\varepsilon)$ and from Theorem 1, $\ell(f,t,r^{-k}) = \ell(f(t,r^{-k})) = \ell(f_k\circ\pi_k(t))$. Also,

$$\begin{aligned}r^{-2}\sigma_{\mathcal{C}}(t) &= \sum_{k>0} r^{-k}\sqrt{\ell(f,t,r^{-k+1})} \\ &\leq \sum_{k>0}(r-1)r^{-k}\sqrt{\ell(f,t,r^{-k+1})} \\ &\leq \sigma_f(t) = \sum_{k>0}\int_{r^{-k}}^{r^{-k+1}}\sqrt{\ell(f,t,\varepsilon)}\,\mathrm{d}\varepsilon \\ &\leq \sum_{k>0}(r-1)r^{-k}\sqrt{\ell(f,t,r^{-k})} \leq \sigma_{\mathcal{C}}(t).\end{aligned}$$

∎

Combining Theorem 3 and Proposition 3 gives the following bound on the expected supremum of X_t based on uniquely decodable codes:

Theorem 7: Given a uniquely decodable code f such that $\sigma_f(t) < \infty$ for all $t \in T$,

$$\mathbf{E}\Big[\sup_{t\in T} X_t\Big] \lesssim r^3 \int_T \sigma_f(t)\nu(\mathrm{d}t) \\ \leq r^3 \sup_{t\in T} \sigma_f(t).$$

where $\nu \in \mathcal{P}(T)$ is defined in Theorem 3.

IV. The lower bound for Gaussian processes

We now consider the case when the process $(X_t)_{t\in T}$ is Gaussian and T is endowed with the canonical metric (4). We will use $L, L_1, \dots$ to denote various absolute constants. Note that their values are not necessarily the same in different appearances. The main technical tool for the lower bound is the following improvement of Sudakov minoration [1, Prop. 2.4.9]:

Proposition 4 (): Let $t_1, \dots, t_m \in T$ be such that $d(t_\ell, t_{\ell'}) \geq a$ if $\ell \neq \ell'$. Consider $b > 0$ and, for each $\ell \leq m$, a finite set $H_\ell \subset B(t_\ell, b)$. Then for $H = \bigcup_{\ell\leq m} H_\ell$ we have

$$\mathbf{E}\Big[\sup_{t\in H} X_t\Big] \geq \frac{a}{L_1}\sqrt{\log m} - L_2 b\sqrt{\log m} + \min_{\ell\leq m}\mathbf{E}\Big[\sup_{t\in H_\ell} X_t\Big].$$

When $b \leq a/(2L_1L_2)$, we have

$$\mathbf{E}\Big[\sup_{t\in H} X_t\Big] \geq \frac{a}{2L_1}\sqrt{\log m} + \min_{\ell\leq m}\mathbf{E}\Big[\sup_{t\in H_\ell} X_t\Big].$$

Another ingredient is the construction of an increasing sequence of partitions on T. We follow Bednorz [10] and Talagrand [11] and use a greedy procedure to construct the partitions $(\mathcal{A}_k)_{k\geq 0}$ for a fixed $r > 1$. For any $A \subseteq T$, let $G(A) := \mathbf{E}[\sup_{t\in A} X_t]$. Let $\mathcal{A}_0 = \{T\}$. To construct $\mathcal{A}_k$ for $k \geq 1$ from $\mathcal{A}_{k-1}$, we partition each $B \in \mathcal{A}_{k-1}$ into sets $A_1, \dots, A_m$ as follows: Let $B_0 := B$ and choose $t_1 \in B_0$ so that

$$G(B(t_1, r^{-k-1}/2) \cap B_0) = \max_{t\in B_0} G(B(t, r^{-k-1}/2) \cap B_0).$$

Let $A_1 := B(t_1, r^{-k}/2) \cap B_0$ and $B_1 := B_0 \setminus A_1$. Now iterate this: If, for $i \geq 1$, $B_{i-1} \neq \varnothing$, choose $t_i \in B_{i-1}$ so that

$$G(B(t_i, r^{-k-1}/2) \cap B_{i-1}) \\ = \max_{t\in B_{i-1}} G(B(t, r^{-k-1}/2) \cap B_{i-1}).$$

Let $A_i := B(t_i, r^{-k}/2) \cap B_{i-1}$ and $B_i := B_{i-1} \setminus A_i$ and continue. This procedure is guaranteed to terminate for some $m \leq N(T, d, r^{-k}/2)$, where $N(T, d, \cdot)$ are the covering numbers of (T, d).

We are now ready to state the majorizing measure theorem via codes:

Theorem 8: Let $(X_t)_{t\in T}$ be a centered Gaussian process. For any $r \geq \max\{2, L\}$, there exists an admissible sequence $\mathcal{C}$ of VLC's with $\rho_k \leq r^{-k}$, such that

$$\mathbf{E}\Big[\sup_{t\in T} X_t\Big] \gtrsim \frac{1}{r}\sigma_{\mathcal{C}}(t), \qquad \forall t \in T. \tag{23}$$

Proof: Let $L = L_1L_2$, where L_1, L_2 are the contants from Prop. 4. Given $r \geq L$, let $(\mathcal{A})_{k\geq 0}$ be an increasing sequence of partitions as described above. For each k and each $A \in \mathcal{A}_k$, pick a unique point $s_A \in A$ and then let $\pi_k(t) := s_{A_k(t)}$, where $A_k(t)$ is the unique element of $\mathcal{A}_k$ containing t.

We construct the codes f_k inductively. Set $f_0(\pi_0(t)) = \mathsf{e}$, the empty string in $\{0,1\}^*$. Assuming $f_0, \dots, f_{k-1}$ have been constructed, we assign code lengths $\ell_k(s)$ to the points $s \in \pi_k(T)$ and show that this assignment satisfies the Kraft–McMillan inequality. Thus, there exists a prefix code $f_k : \pi_k(T) \to \{0,1\}^*$ with $\ell(f_k(s)) = \ell_k(s)$ for all $s \in \pi_k(T)$.

For each $s \in \pi_k(T)$, let $i_k(s) \in \mathbb{N}$ be the index of the set $A_k(s) \equiv \pi_k^{-1}(s)$ as a partition element in $A_{k-1}(s) \equiv \pi_{k-1}^{-1}(s)$ inherited from the above greedy construction procedure. Then, with $\ell_k(s) = \ell(f_{k-1} \circ \pi_{k-1}(s)) + \log(i_k(s)+1)^2$, we have

$$\sum_{s\in\pi_k(T)} 2^{-\ell_k(s)} \\ = \sum_{t\in\pi_{k-1}(T)} \sum_{s:\pi_{k-1}(s)=t} 2^{-\ell(f_{k-1}(s)) - \log(i_k(s)+1))^2} \\ \leq \sum_{t\in\pi_{k-1}(T)} 2^{-\ell(f_{k-1}(t))} \sum_{i\geq 1} \frac{1}{(i+1)^2} \leq 1.$$

Hence the existence of a uniquely decodable (prefix) code f_k is guaranteed by Kraft–McMillan inequality. Then, by the construction procedure and Prop. 4, for each $t \in T$,

$$G(A_{k-1}(t)) \geq G(A_{k+1}(t)) + Lr^{-k}\sqrt{\log(i_k(\pi_k(t)))}.$$

Since $1 + \sqrt{\log i} \geq \sqrt{1 + \log i} \geq \sqrt{\log(i+1)}$ for $i \geq 1$,

$$G(A_{k-1}(t)) + Lr^{-k} \\ \geq G(A_{k+1}(t)) + Lr^{-k}\sqrt{\log(i_k(\pi_k(t))+1)^2} \\ = G(A_{k+1}(t)) + Lr^{-k}\sqrt{\ell_k(t) - \ell_{k-1}(t)},$$

where $\ell_k(t) = \ell(f_k \circ \pi_k(t))$. Applying induction on $G(A_0(t))$ and $G(A_1(t))$, we obtain

$$G(T) + L \gtrsim \sum_k r^{-k}\sqrt{\ell_k(t) - \ell_{k-1}(t)}, \qquad \forall t \in T \tag{24}$$

Thm. 4 and the fact that $G(T) \gtrsim \operatorname{diam}(T)$, so that the universal constant can be absorbed by $G(T)$, give the result. ∎

Note that for any $k \in \mathbb{N}$ and any $s \in \pi_{k-1}(T)$, in the construction of codes $f_k(s)$ given $f_{k-1}(s)$, we can append prefix codes with length $\log(i_k(s)+1)^2$ to the code $f_k(s)$ so that $f_k(s) \in f_{k-1}(s) \circ \{0,1\}^*$. Let $f = f_K$ where $K \in \mathbb{N}$ is such that $\pi_K^{-1}(t) = t$ for all $t \in T$. Then

$$\ell(f, t, r^{-k}) \leq \ell(f_k \circ \pi_k(t)), \quad t \in T, k \in \mathbb{N}$$

since $\Delta(\pi_k^{-1} \circ \pi_k(t)) \leq r^{-k}$. Thus, Thm. 8 can be stated as follows:

Theorem 9: Under the same assumption as in Theorem 8, there exists a uniquely decodable code $f : T \to \{0,1\}^*$ such that

$$\mathbf{E}\Big[\sup_{t\in T} X_t\Big] \gtrsim \frac{1}{r}\sigma_f(t), \qquad \forall t \in T.$$

Combining upper bounds in the Theorems 3 and 7 and lower bounds in Theorems 8 and 9, expected supremum of centered

Gaussian processes $(X_t)_{t\in T}$ can be characterized as

$$\mathbf{E}\Big[\sup_{t\in T} X_t\Big] \asymp \inf \sup_{t\in T} \sigma(t) \asymp \inf \sup_{\mu\in\mathcal{P}(T)} \int \sigma \,\mathrm{d}\mu \asymp \inf \int \sigma \,\mathrm{d}\nu,$$

where σ is either $\sigma_{\mathcal{C}}$ or σ_f, the probability measure ν is same as in Theorem 3, and the infimum is over either admissible sequence of codes $\mathcal{C}$ or uniquely decodable codes f, depending on the expression of σ.

V. Bounds on expected suprema via random codes

In this section, we present an alternative characterization of the expected supremum $\mathbf{E}[\sup_{t\in T} X_t]$ via random codes. Let (Π, F) be a random variable-length code on T that operates at resolution level ρ almost surely and whose randomness is independent of the σ-field induced by $(X_t)_{t\in T}$. Given any probability measure $\mu \in \mathcal{P}(T)$, let τ be a random variable taking values in T with law μ and let

$$M^*_\mu(\rho) := \inf \mathbf{E}\sqrt{\ell(F\circ\Pi(\tau))} \tag{25}$$

where the infimum is over all random codes (Π, F) that operate at resolution level ρ and are independent of τ. In the following, we use the quantity defined in (25) to obtain an upper bound on the expected suprema of subgaussian processes and a matching lower bound for Gaussian processes.

A. The upper bound for subgaussian processes

The argument for the upper bound is very similar to that in Section III, except here the sequence of VLC's on T is not required to be admissible. More specifically, the (random) codes are chosen independently at each scale, so that the optimization over the random codes in (25) can be performed separately at each scale.

We first present a modification of Theorem 2 that does not require admissibility. Let $\mathcal{C} = \{(\pi_k, f_k)\}_{k\in\mathbb{N}}$ be an arbitrary (not necessarily admissible) sequence of VLC's, where each (π_k, f_k) operates at resolution ρ_k and the sequence $(\rho_k)_{k\in\mathbb{N}}$ is nonincreasing. Let $\boldsymbol{p} = (p_k)_{k\geq 1}$ be a positive probability mass function on $\mathbb{N}$.

Theorem 10: Let $(X_t)_{t\in T}$ be a centered random process satisfying the increment condition (1). Fix an arbitrary point $t_0 \in T$. Then, for any $u > 0$, $\varepsilon > 0$, and any $A \subset B(t_0, \varepsilon)$,

$$|X_t - X_{t_0}| \leq 4\bar{\sigma}_{(\mathcal{C},\boldsymbol{p})}(t) + 2uR(\varepsilon), \qquad \text{for all } t\in A$$

with probability at least $1 - e^{-u^2}$ where $\bar{\sigma}_{(\mathcal{C},\boldsymbol{p})}(t)$ is defined in (15) and where $R(\varepsilon) = \sum_{k\geq\bar{k}} \rho_{k-1}$ with $\bar{k}\in\mathbb{N}$ chosen so that $\rho_{\bar{k}} < \varepsilon \leq \rho_{\bar{k}-1}$.

Proof: Let k and $\bar{k}$ be defined as in the proof of Theorem 2. Let $\pi^A_k := \pi_k|_A$. For $k \geq \bar{k}$, since π_k operates at resolution ρ_k, $\mathrm{card}(\pi^A_k(A))$ is at least two. Let $\pi^A_{\bar{k}-1}(t) = t_0$ for $t \in A$. For $k \geq \bar{k}$, let $S_k := \pi^A_k(A)$ and $U_k := \{(\pi^A_k(t), \pi^A_{k-1}(t)) : t \in T\} \subset S_k \times S_{k-1}$ and define $\xi_k : U_k \to \mathbb{R}$ by

$$\xi_k(s,s') := d(s,s')\sqrt{\ln\frac{2^{\ell(f_k(s))+\ell(f_{k-1}(s'))+1}}{p_k} + u^2}$$

and the event $E_k := \big\{\exists (s,s')\in U_k \text{ such that } |X_s - X_{s'}| > \xi_k(s,s')\big\}$. Using the increment condition (1) and the Kraft–McMillan inequality, we have

$$\begin{aligned}
\mathbf{P}[E_k] &\leq \sum_{(s,s')\in U_k} \mathbf{P}[|X_s - X_{s'}| > \xi_k(s,s')] \\
&\leq 2\sum_{s\in S_k}\sum_{s'\in S_{k-1}} \exp\left(-\frac{\xi_k^2(s,s')}{d^2(s,s')}\right) \\
&= 2\sum_{s\in S_k}\sum_{s\in S_{k-1}} \exp\left(-\ln\frac{2^{\ell(f_k(s))+\ell(f_{k-1}(s'))+1}}{p_k} - u^2\right) \\
&= p_k e^{-u^2}\sum_{s\in S_k}\sum_{s\in S_{k-1}} 2^{-\ell(f_k(s))-\ell(f_{k-1}(s'))} \\
&\leq p_k e^{-u^2}.
\end{aligned}$$

For any $t \in A$, combining the above estimate with the chaining decomposition (16) gives

$$\begin{aligned}
&\mathbf{P}\left[\exists t\in A \text{ such that } |X_t - X_{t_0}| > \sum_{k\geq\bar{k}} \xi_k(\pi^A_k(t), \pi^A_{k-1}(t))\right] \\
&\leq \mathbf{P}\Bigg[\exists t\in A \text{ such that} \\
&\qquad \sum_{k\geq\bar{k}}\left(|X_{\pi^A_k(t)} - X_{\pi^A_{k-1}(t)}| - \xi_k(\pi^A_k(t), \pi^A_{k-1}(t))\right) > 0\Bigg] \\
&\leq \mathbf{P}\left[\bigcup_k E_k\right] \\
&\leq e^{-u^2},
\end{aligned}$$

where the last step follows from the assumption on the weights p_k. Thus, since

$$\begin{aligned}
d(\pi^A_k(t), \pi^A_{k-1}(t)) &\leq d(\pi^A_k(t), t) + d(t, \pi^A_{k-1}(t)) \\
&\leq \rho_k + \rho_{k-1} \\
&\leq 2\rho_{k-1}
\end{aligned}$$

for $k > \bar{k}$ and $d(\pi^A_{\bar{k}}(t), \pi^A_{\bar{k}-1}(t)) = d(\pi^A_{\bar{k}}(t), t_0) \leq \varepsilon \leq \rho_{k-1}$,

$$\begin{aligned}
&|X_t - X_{t_0}| \\
&\leq \sum_{k\geq\bar{k}} \xi_k(\pi^A_k(t), \pi^A_{k-1}(t)) \\
&\leq 2\sum_{k>0}\rho_{k-1}\sqrt{\ln\frac{2^{\ell(f_k\circ\pi_k(t))+1}}{p_k}} \\
&\qquad + 2\sum_{k>0}\rho_{k-1}\sqrt{\ln 2^{\ell(f_{k-1}\circ\pi_{k-1}(t))}} + 2u\sum_{k\geq\bar{k}}\rho_{k-1} \\
&\leq 4\bar{\sigma}_{(\mathcal{C},\boldsymbol{p})}(t) + 2u\sum_{k\geq\bar{k}}\rho_{k-1}, \qquad \text{for all } t\in A
\end{aligned}$$

with probability at least $1 - e^{-u^2}$. This completes the proof. ■

Corollary 3: With the same assumptions in Theorem 3,

$$\mathbf{E}\Big[\sup_{t\in T} X_t\Big] \lesssim \int_0^1 M^*_\nu(\varepsilon)\,\mathrm{d}\varepsilon,$$

where $\nu \in \mathcal{P}(T)$ is for the probability measure $\nu \in \mathcal{P}(T)$ defined in Theorem 3.

Proof: Let $\rho_k = r^{-k}$ for some $r > 1$ and $p_k = 2^{-k}$. Let $\{(\Pi_k, F_k)\}_{k\in\mathbb{N}}$ be a sequence of independent random VLC's on T, where each (Π_k, F_k) operates at resolution $\rho_k = r^{-k}$. Applying Theorem 10 and repeating the argument in the proof of Theorem 3, we have for any realization $\{(F_k, \Pi_k)\}_{k\in\mathbb{N}}$,

$$\mathbf{E}\Big[\sup_{t\in T} X_t\Big] \lesssim \sum_k r^{-k+1} \int_T \sqrt{\ell(F_k \circ \Pi_k(t))}\nu(\mathrm{d}t).$$

Taking the expectation with respect to the random codes at each level k yields

$$\mathbf{E}\Big[\sup_{t\in T} X_t\Big] \lesssim \sum_k r^{-k+1} \int_T \mathbf{E}\sqrt{\ell(F_k \circ \Pi_k(t))}\,\mathrm{d}\nu(t).$$

We can now minimize both sides over all independent sequences of random codes $\{(\Pi_k, F_k)\}_{k\in\mathbb{N}}$ under the constraint that each (Π_k, F_k) operates at resolution r^{-k}. Since the codes for each k are independent, the minimization can be carried out separately for each k, which gives

$$\mathbf{E}\Big[\sup_{t\in T} X_t\Big] \lesssim \sum_k r^{-k+1} M^*_\nu(r^{-k}).$$

The integral form in the statement follows by recognizing the function $\varepsilon \mapsto M^*_\nu(\varepsilon)$ is nonincreasing. ■

B. The lower bound for Gaussian processes

As in the formulation via deterministic codes, we obtain a matching lower bound in Corollary 3 for Gaussian processes.

Theorem 11: Let $(X_t)_{t\in T}$ be a centered Gaussian process. Then

$$\mathbf{E}\Big[\sup_{t\in T} X_t\Big] \gtrsim \sup_{\mu\in\mathcal{P}(T)} \int_0^1 M^*_\mu(\varepsilon)\,\mathrm{d}\varepsilon.$$

Before proving this lower bound, we first establish two auxiliary lemmas. Recall the definition of the functional $M(\mu,\nu)$ from Section II-B. The first auxiliary lemma is a restatement of [10, Theorem 4]. The original argument relies on a distributional version of super-Sudakov inequality obtained from careful bi-level grouping of constructed partitions. Below we provide a simpler argument based on an elementary algebraic fact.

Lemma 2: For a centered Gaussian process $(X_t)_{t\in T}$,

$$\mathbf{E}\sup_{t\in T} X_t \gtrsim \sup_\mu M(\mu,\mu).$$

Proof: Notice that, for any $\mu \in \mathcal{P}(T)$, any admissible sequence $\mathcal{C}$ of VLCs, and for each $k > 0$,

$$\begin{aligned}
&1 + \int_T \sqrt{\ell(f_k \circ \pi_k(t))}\mu(\mathrm{d}t)\\
&\geq \sum_{s\in\pi_k(T)} \Big(2^{-\ell(f_k(s))} + \mu(\pi_k^{-1}(s))\sqrt{\ell(f_k(s))}\Big)\\
&\geq \sum_{s\in\pi_k(T)} \mu(\pi_k^{-1}(s))\sqrt{\log\frac{1}{\mu(\pi_k^{-1}(s))}}\\
&\geq \int_T \sqrt{\log\frac{1}{\mu(B(t,r^{-k}))}}\mu(\mathrm{d}t).
\end{aligned}$$

The second inequality uses (19) and the last inequality is due to the fact that $\pi_k^{-1} \circ \pi_k(t) \subset B(t, r^{-k})$ for any $t \in T$. Thus, for some universal constant $L > 0$,

$$L + \int_T \sigma_{\mathcal{C}}\,\mathrm{d}\mu \geq \sum_k \int_T \sqrt{\log\frac{1}{\mu(B(t,r^{-k}))}}\mu(\mathrm{d}t)$$

for all $\mu \in \mathcal{P}(T)$. Together with Theorem 8, it implies that

$$\mathbf{E}\Big[\sup_{t\in T} X_t\Big] \gtrsim \int_0^1 \int_T \sqrt{\log\frac{1}{\mu(B(t,\epsilon))}}\mu(\mathrm{d}t)\,\mathrm{d}\epsilon \quad \forall\mu\in\mathcal{P}(T).$$

Taking the supremum over μ completes the proof. ■

The second auxiliary lemma shows that the integrand in the Fernique–Talagrand functional (3) can be lower-bounded in terms of a particular random VLC.

Lemma 3: For any $\rho > 0$ and any $\mu \in \mathcal{P}(T)$, there exists a random code $(\Pi^\mu_\rho, f_{\mathsf{e}})$ that operates at resolution level ρ, so that

$$\mathbf{E}\sqrt{\ell(f_{\mathsf{e}} \circ \Pi^\mu_\rho(t))} \leq \sqrt{3\log\frac{1}{\mu(B(t,\rho))}} + 2 \quad \forall t\in T$$

Proof: We make use of the following random code construction due to Kontoyiannis [14]. Let $U = (U_1, U_2, \ldots)$ be i.i.d. samples drawn from μ and for each $t \in T$ define the waiting time

$$\tau_t = \inf\{i \geq 1 : d(t, U_i) \leq \rho\}.$$

Then $\Pi^\mu_\rho(t) = U_{\tau_t}$. For the binary code, we use Elias' prefix-free code for the positive integers [15], so that

$$\begin{aligned}
&\ell(f_{\mathsf{e}} \circ \Pi^\mu_\rho(t))\\
&\leq \lceil\log(\tau_t+1)\rceil 2\lceil\log(\lceil\log(\tau_t+1)\rceil+1)\log(\tau_t+1)\rceil\\
&\leq 2\log\tau_t + 4
\end{aligned}$$

Since $\mathbf{E}\tau_t = \frac{1}{\mu(B(t,\rho))}$, taking expectations and applying Jensen's inequality gives

$$\begin{aligned}
\mathbf{E}\sqrt{\ell(f_{\mathsf{e}} \circ \Pi^\mu_\rho(t))} &\leq \sqrt{\mathbf{E}\ell(f_{\mathsf{e}} \circ \Pi^\mu_\rho(t))}\\
&\leq \sqrt{3\log\frac{1}{\mu(B(t,\rho))}} + 2.
\end{aligned}$$

This completes the proof. ■

With these two results at hand, we can prove Theorem 11.

Proof of Theorem 11: Combining the results from Lemmas 2 and 3 yields

$$\begin{aligned}
\mathbf{E}\Big[\sup_{t\in T} X_t\Big] &\gtrsim \sup_{\mu\in\mathcal{P}(T)} \int_0^1 \int_T \sqrt{\log\frac{1}{\mu(B(t,\epsilon))}}\mu(\mathrm{d}t)\,\mathrm{d}\epsilon\\
&\gtrsim \sup_{\mu\in\mathcal{P}(T)} \int_0^1 \int_T \mathbf{E}\sqrt{\ell(f_{\mathsf{e}} \circ \Pi^\mu_\varepsilon(t))}\mu(\mathrm{d}t)\,\mathrm{d}\varepsilon\\
&\gtrsim \sup_{\mu\in\mathcal{P}(T)} \int_0^1 M^*_\mu(\varepsilon)\,\mathrm{d}\epsilon.
\end{aligned}$$

■

Putting together the results of Corollary 3 and Theorem 11, we see that, for centered Gaussian processes $(X_t)_{t\in T}$,

$$\mathbf{E}\Big[\sup_{t\in T} X_t\Big] \asymp \sup_{\mu}\int_0^1 M^*_\mu(\varepsilon)\,\mathrm{d}\varepsilon \asymp \int_0^1 M^*_\nu(\varepsilon)\,\mathrm{d}\varepsilon$$

where $\nu \in \mathcal{P}(T)$ is the probability measure defined in Theorem 3.

Acknowledgment

We would like to thank the Associate Editor and the anonymous reviewer for their careful reading of the paper and for several constructive suggestions that helped significantly improve the presentation.

References

[1] M. Talagrand, *Upper and Lower Bounds for Stochastic Processes*. Springer, 2014.

[2] X. Fernique, "Regularité des trajectoires des fonctions aléatoires gaussiennes," in *Ecole d'Eté de Probabilités de Saint-Flour IV—1974*. Springer, 1975, pp. 1–96.

[3] M. Talagrand, "Regularity of gaussian processes," *Acta Mathematica*, vol. 159, pp. 99–149, 1987.

[4] O. Guédon and A. Zvavitch, "Supremum of a process in terms of trees," in *Geometric Aspects of Functional Analysis*. Springer, 2003, pp. 136–147.

[5] A. Maurer, "Majorizing codes and measures," preprint, available at http://www.andreas-maurer.eu/mmnotes.pdf, 2010.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2006.

[7] I. Kontoyiannis and J. Zhang, "Arbitrary source models and Bayesian codebooks in rate-distortion theory," *IEEE Trans. Inform. Theory*, vol. 48, no. 8, pp. 2276–2290, August 2002.

[8] P. Massart, *Concentration Inequalities and Model Selection*. Springer, 2007.

[9] J.-Y. Audibert and O. Bousquet, "Combining PAC-Bayesian and generic chaining bounds," *Journal of Machine Learning Research*, vol. 8, no. 32, pp. 863–889, 2007.

[10] W. Bednorz, "The majorizing measure approach to sample boundedness," *Colloquium Mathematicae*, vol. 139, no. 2, pp. 205–227, 2015.

[11] M. Talagrand, "A simple proof of the majorizing measure theorem," *Geometric and Functional Analysis*, vol. 2, no. 1, pp. 118–125, March 1992.

[12] X. Fernique, "Caractérisation de processus à trajectoires majorées ou continues," *Séminaire de probabilités de Strasbourg*, vol. 12, pp. 691–706, 1978.

[13] R. van Handel, "Probability in high dimension," Princeton University lecture notes, 2016.

[14] I. Kontoyiannis, "Pointwise redundancy in lossy data compression and universal lossy data compression," *IEEE Transactions on Information Theory*, vol. 46, no. 1, pp. 136–152, 2000.

[15] P. Elias, "Universal codeword sets and representations of the integers," *IEEE Transactions on Information Theory*, vol. 21, no. 2, pp. 194–203, 1975.

Yifeng Chu received B.S. and M.S. degrees in electrical engineering from the University of Illinois Urbana-Champaign, Urbana, IL, USA (UIUC) in 2018 and 2021 respectively, and is currently pursuing the Ph.D. degree at UIUC. His research interests include machine learning, probability theory, and information theory.

Maxim Raginsky (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Northwestern University, Evanston, IL, USA, in 2000, 2000, and 2002, respectively. He has held research positions with Northwestern University, the University of Illinois at Urbana-Champaign, Urbana, IL, USA, where he was a Beckman Foundation Fellow from 2004 to 2007, and Duke University, Durham, NC, USA. In 2012, he returned to UIUC, where he is currently a Professor and William L. Everitt Fellow with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory. His interests cover probability and stochastic processes, deterministic and stochastic control, machine learning, optimization, and information theory. Much of his recent research is motivated by fundamental questions in modeling, learning, and simulation of nonlinear dynamical systems, with applications to advanced electronics, autonomy, and artificial intelligence. Dr. Raginsky received the Faculty Early Career Development (CAREER) Award from the National Science Foundation in 2013 and was the recipient of the 2024 IEEE Control Systems Society Roberto Tempo CDC Best Paper Award (with Belinda Tzen, Anant Raj, and Francis Bach). He is a member of the editorial boards of *Journal of Machine Learning Research*, *Foundations and Trends in Machine Learning*, *SIAM Journal on Mathematics of Data Science*, and *Mathematics of Control, Signals, and Systems*.