Notes on Gaussian processes and majorizing measures

James R. Lee

1 Gaussian processes

Consider a *Gaussian process* $\{X_t\}_{t \in T}$ for some index set T. This is a collection of jointly Gaussian random variables, meaning that every finite linear combination of the variables has a Gaussian distribution. We will additionally assume that the process is *centered*, i.e. $\mathbb{E}(X_t) = 0$ for all $t \in T$.

It is well-known that such a process is completely characterized by the covariances $\{\mathbb{E}(X_s X_t)\}_{s,t \in T}$. For $s, t \in T$, consider the canonical distance,

$$d(s,t) = \sqrt{\mathbb{E}\,|X_s - X_t|^2},$$

which forms a metric on T. (Strictly speaking, this is only a pseudometric since possibly $d(s,t) = 0$ even though X_s and X_t are distinct random variables, but we'll ignore this.) Since the process is centered, it is completely specified by the distance $d(s,t)$, up to translation by a Gaussian (e.g. the process $\{X_t + X_{t_0}\}_{t \in T}$ will induce the same distance for any $t_0 \in T$).

1.1 A concrete perspective

If the index set T is countable, one can describe every such process in the following way. Let $\{g_i\}_{i=1}^{\infty}$ be a sequence of i.i.d. standard Gaussians, let $T \subseteq \ell^2$, and put

$$X_t = \sum_{i \geqslant 1} g_i t_i.$$

In this case, it is easy to check that $d(s,t) = \|s - t\|_2$ for $s, t \in T$. (The fact that this construction is universal follows from the fact that every two separable Hilbert spaces are isomorphic.)

1.2 Random projections

If T is finite, then we can think of $T \subseteq \mathbb{R}^n$ for some $n \in \mathbb{N}$. In this case, if g is a standard n-dimensional Gaussian, then

$$X_t = \langle g, t \rangle,$$

and we can envision the process as the projection of T onto a uniformly random direction (see Figure 1).

1.3 Studying the maxima

We will be concerned primarily with the value:

$$\mathbb{E} \sup_{t \in T} X_t.$$

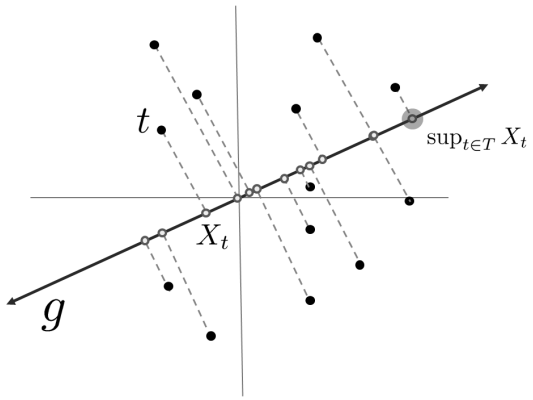

Figure 1: Points in $\mathbb{R}^2$ projected onto a random direction.

(I.e. the expected value of the extremal node circled above.) One may assume that T is finite without losing any essential ingredient of the theory, in which case the supremum can be replaced by a maximum. Note that we are studying the *tails* of the process. Dealing with these extremal values is what makes understanding the above quantity somewhat difficult.

As some motivation for the classical study of this quantity, one has the following.

Theorem 1.1. *For a separable Gaussian process $\{X_t\}_{t\in T}$, the following two facts are equivalent.*

1. *The map $t \mapsto X_t(\omega)$ is uniformly continuous (as a map from (T,d) to $\mathbb{R}$) with probability one.*
2. *As $\varepsilon \to 0$,*
$$\mathbb{E} \sup_{d(s,t)\leqslant\varepsilon} |X_s - X_t| \to 0.$$

However, from our viewpoint, the quantitative study of $\mathbb{E}\sup_{t\in T} X_t$ in terms of the geometry of (T,d) will play the fundamental role.

1.4 Bounding the sup

We will concentrate first on finding good upper bounds for $\mathbb{E}\sup_{t\in T} X_t$. Toward this end, fix some $t_0 \in T$, and observe that

$$\mathbb{E}\sup_{t\in T} X_t = \mathbb{E}\sup_{t\in T}(X_t - X_{t_0}).$$

Since $\sup_{t\in T}(X_t - X_{t_0})$ is a non-negative random variable, we can write

$$\mathbb{E}\sup_{t\in T}(X_t - X_{t_0}) = \int_0^\infty \mathbb{P}\left(\sup_{t\in T} X_t - X_{t_0} > u\right) du,$$

and concentrate on finding upper bounds on the latter probabilities.

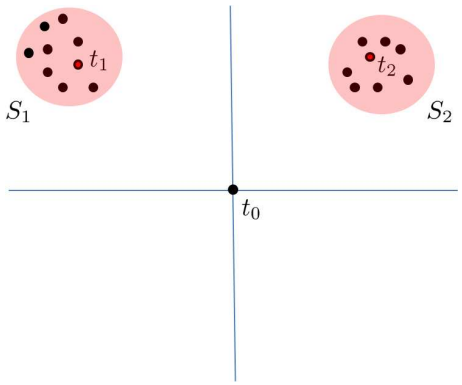

Figure 2: A clustered set of Gaussians.

1.5 Improving the union bound

As a first step, we might write

$$\mathbb{P}\left(\sup_{t\in T} X_t - X_{t_0} > u\right) \leqslant \sum_{t\in T} \mathbb{P}\left(X_t - X_{t_0} > u\right).$$

While this bound is decent if the variables $\{X_t - X_{t_0}\}_{t\in T}$ are somewhat independent, it is rather abysmal if the variables are clustered.

Look at Figure 2. Since the variables in, e.g. S_1, are highly correlated (in the "geometric" language, they tend to project close together on a randomly chosen direction), the union bound is overkill. It is natural to choose representatives $t_1 \in S_1$ and $t_2 \in S_2$. We can first bound $X_{t_1} - X_{t_0}$ and $X_{t_2} - X_{t_0}$, and then bound the intra-cluster values $\{X_t - X_{t_1}\}_{t\in S_1}$ and $\{X_t - X_{t_2}\}_{t\in S_2}$. This should yield better bounds as the diameter of S_1 and S_2 are hopefully significantly smaller than the diameter of T.

Formally, we have

$$\begin{aligned}\mathbb{P}\left(\sup_{t\in T} X_t - X_{t_0} > u\right) \quad \leqslant \quad & \mathbb{P}(X_{t_1} - X_{t_0} > u/2) + \sum_{t\in S_1} \mathbb{P}(X_t - X_{t_1} > u/2) \\ & + \mathbb{P}(X_{t_1} - X_{t_0} > u/2) + \sum_{t\in S_2} \mathbb{P}(X_t - X_{t_2} > u/2).\end{aligned}$$

Of course, there is no reason to stop at one level of clustering, and there is no reason that we should split the contribution $u = u/2+u/2$ evenly. In the next post, we'll see the "generic chaining" method which generalizes and formalizes our intuition about improving the union bound.

2 The generic chaining

In the last section, we considered a Gaussian process $\{X_t\}_{t\in T}$ and were trying to find upper bounds on the quantity $\mathbb{E}\sup_{t\in T} X_t$. We saw that one could hope to improve over the union bound by clustering the points and then taking mini union bounds in each cluster.

2.1 Hierarchical clustering

To specify a clustering, we'll take a sequence of progressively finer approximations to our set T. First, recall that we fixed $t_0 \in T$, and we have used the observation that $\mathbb{E}\sup_{t\in T} X_t = \mathbb{E}\sup_{t\in T}(X_t - X_{t_0})$.

Now, assume that T is finite. Write $T_0 = \{t_0\}$, and consider a sequence of subsets $\{T_n\}$ such that $T_0 \subseteq T_1 \subseteq T_2 \subseteq \cdots \subseteq T$. We will assume that for some large enough m, we have $T_n = T$ for $n \geqslant m$. For every $n \geqslant 0$, let $\pi_n : T \to T_n$ denote a "closest point map" which satisfies $d(t, \pi_n(t)) = d(t, T)$ for all $t \in T$.

The main point is that we can now write, for any $t \in T$,

$$X_t - X_{t_0} = \sum_{n\geqslant 1} X_{\pi_n(t)} - X_{\pi_{n-1}(t)}. \tag{1}$$

This decomposition is where the term "chaining" arises, and now the idea is to bound the probability that $X_t - X_{t_0}$ is large in terms of the segments in the chain.

2.2 What should T_n look like?

One question that arises is how we should think about choosing the approximations T_n. We are trading off two measures of quality: The denser T_n is in the set T (or, more precisely, in the set T_{n-1}) the smaller the variances of the segments $X_{\pi_n(t)} - X_{\pi_{n-1}(t)}$ will be. On the other hand, the larger T_n is, the more segments we'll have to take a union bound over.

So far, we haven't used any property of our random variables except for the fact that they are centered. To make a more informed decision about how to choose the sets $\{T_n\}$, let's recall the classical Gaussian concentration bound.

Lemma 2.1. *For every $s, t \in T$ and $\lambda > 0$,*

$$\mathbb{P}(X_s - X_t > \lambda) \leqslant \exp\left(-\frac{\lambda^2}{2\,d(s,t)^2}\right). \tag{2}$$

This should look familiar: $X_s - X_t$ is a mean-zero Gaussian with variance $d(s,t)^2$.

Now, a first instinct might be to choose the sets T_n to be progressively denser in T. In this case, a natural choice would be to insist on something like T_n being a 2^{-n}-net in T. If one continues down this path in the right way, a similar theory would develop. We're going to take a different route and consider the other side of the tradeoff.

Instead of insisting that T_n has a certain level of accuracy, we'll insist that T_n is at most a certain size. Should we require $|T_n| \leqslant n$ or $|T_n| \leqslant 2^n$, or use some other function? To figure out the right bound, we look at (2). Suppose that $g_1, g_2, \ldots, g_m$ are i.i.d. $N(0,1)$ random variables. In that case, applying (2) and a union bound, we see that to achieve

$$\mathbb{P}(\exists i : g_i > B) \leqslant m\mathbb{P}(g_1 > B) < 1,$$

we need to select $B \asymp \sqrt{\log m}$. If we look instead at m^2 points instead of m points, the bound grows to $\sqrt{2\log m}$. Thus we can generally square the number of points before the union bound has to pay a constant factor increase. This suggests that the right scaling is something like $|T_{n+1}| = |T_n|^2$. So we'll require that $|T_n| \leqslant 2^{2^n}$ for all $n \geqslant 1$.

2.3 The generic chaining

This leads us to the generic chaining bound, due to Fernique (though the formulation we state here is from Talagrand).

Theorem 2.2. *Let $\{X_t\}_{t\in T}$ be a Gaussian process, and let $T_0 \subseteq T_1 \subseteq \cdots \subseteq T$ be a sequence of subsets such that $|T_0| = 1$ and $|T_n| \leqslant 2^{2^n}$ for $n \geqslant 1$. Then,*

$$\mathbb{E} \sup_{t\in T} X_t \leqslant O(1) \sup_{t\in T} \sum_{n\geqslant 0} 2^{n/2} d(t, T_n). \tag{3}$$

Proof. As before, let $\pi_n : T \to T_n$ denote the closest point map and let $T_0 = \{t_0\}$. Using (2), for any $n \geqslant 1$, $t \in T$, and $u > 0$, we have

$$\mathbb{P}\left(|X_{\pi_n(t)} - X_{\pi_{n-1}(t)}| > u2^{n/2} d(\pi_n(t), \pi_{n-1}(t))\right) \leqslant \exp\left(-\frac{u^2}{2} 2^n\right).$$

Now, the number of pairs $(\pi_n(t), \pi_{n-1}(t))$ can be bounded by $|T_n| \cdot |T_{n-1}| \leqslant 2^{2^{n+1}}$, so we have

$$\mathbb{P}\left(\exists t : |X_{\pi_n(t)} - X_{\pi_{n-1}(t)}| > u2^{n/2} d(\pi_n(t), \pi_{n-1}(t))\right) \leqslant 2^{2^{n+1}} \exp\left(-\frac{u^2}{2} 2^n\right). \tag{4}$$

If we define the event

$$\Omega_u = \left\{\forall n \geqslant 1, t \in T : |X_{\pi_n(t)} - X_{\pi_{n-1}(t)}| \leqslant u2^{n/2} d(\pi_n(t), \pi_{n-1}(t))\right\},$$

then summing (4) yields,

$$\mathbb{P}(\overline{\Omega_u}) \leqslant \sum_{n\geqslant 1} 2^{2^{n+1}} \exp\left(-\frac{u^2}{2} 2^n\right) \leqslant O(1)\, e^{-u^2} \tag{5}$$

for $u \geqslant 4$, since we get geometrically decreasing summands.

Write

$$S = \sup_{t\in T} \sum_{n\geqslant 1} 2^{n/2} d(\pi_n(t), \pi_{n-1}(t)).$$

Note that if Ω_u occurs, then $\sup_{t\in T}(X_t - X_{t_0}) \leqslant uS$. Thus (5) implies that

$$\mathbb{P}(\sup_{t\in T} X_t - X_{t_0} > uS) \leqslant O(1)\, e^{-u^2},$$

which implies that

$$\mathbb{E} \sup_{t\in T} X_t \leqslant O(S) \leqslant O(1) \sup_{t\in T} \sum_{n\geqslant 1} 2^{n/2} d(\pi_n(t), \pi_{n-1}(t)). \tag{6}$$

Finally, by the triangle inequality,

$$d(\pi_n(t), \pi_{n-1}(t)) \leqslant d(t, T_n) + d(t, T_{n-1}) \leqslant 2\, d(t, T_{n-1}).$$

Plugging this into (6) recovers (3). □

Theorem 1.2 gives us a fairly natural way to upper bound the expected supremum using a hierarchical clustering of T. Rather amazingly, we'll soon see that this upper bound is tight. Talagrand's majorizing measure theorem states that if we take the *best* choice of $\{T_n\}$ in Theorem 1.2, then the upper bound in (3) is within a constant factor of $\mathbb{E} \sup_{t\in T} X_t$.

3 Majorizing measures: Some Gaussian tools

In order to prove that the chaining argument is tight, we will need some additional properties of Gaussian processes. For the chaining upper bound, we used a series of union bounds specified by a tree structure. As a first step in producing a good lower bound, we will look at a way in which the union bound is tight.

Theorem 3.1 (Sudakov inequality)**.** *For some constant $C > 0$, the following holds. Let $\{X_t\}_{t\in T}$ be a Gaussian process such that for every distinct $s, t \in T$, we have $d(s,t) \geqslant \alpha$. Then,*

$$\mathbb{E} \sup_{t\in T} X_t \geqslant C\alpha\sqrt{\log |T|}.$$

The claim is an elementary calculation for a sequence of i.i.d. $N(0,1)$ random variables $g_1, g_2, \ldots, g_n$ (i.e. $\mathbb{E}\sup_i g_i \geqslant C\sqrt{\log n}$). We will reduce the general case to this one using Slepian's comparison lemma.

Lemma 3.2 (Slepian's Lemma)**.** *Let $\{X_t\}_{t\in T}$ and $\{Y_t\}_{t\in T}$ be two Gaussian processes such that for all $s, t \in T$,*

$$\mathbb{E}\,|X_s - X_t|^2 \geqslant \mathbb{E}\,|Y_s - Y_t|^2. \tag{7}$$

Then $\mathbb{E}\sup_{t\in T} X_t \geqslant \mathbb{E}\sup_{t\in T} Y_t$.

There is a fairly elementary proof of Slepian's Lemma, if one is satisfied with the weaker conclusion $2\,\mathbb{E}\,|X_s - X_t|^2 \geqslant \mathbb{E}\,|Y_s - Y_t|^2$, which suffices for our purposes.

To see that Lemma 3.2 yields Theorem 3.1, take a family $\{X_t\}_{t\in T}$ with $d(s,t) \geqslant \alpha$ for all $s \neq t \in T$ and consider the associated variables $Y_t = \frac{\alpha}{\sqrt{2}} g_t$ where $\{g_t\}_{t\in T}$ is a family of i.i.d. $N(0,1)$ random variables. It is straightforward to verify that (7) holds, hence by the lemma, $\mathbb{E}\sup_{t\in T} X_t \geqslant \frac{\alpha}{\sqrt{2}}\mathbb{E}\sup_{t\in T} g_t$, and the result follows from the i.i.d. case.

The Sudakov inequality gives us "one level" of a lower bound; the following strengthening will allow us to use it recursively. If we have a Gaussian process $\{X_t\}_{t\in T}$ and $A \subseteq T$, we will use the notation

$$g(A) = \mathbb{E} \sup_{t\in A} X_t.$$

For $t \in T$ and $R \geqslant 0$, we also use the notation

$$B(t, R) = \{s \in T : d(s,t) \leqslant R\}.$$

Here is the main theorem of this post; its statement is all we will require for our proof of the majorizing measures theorem:

Theorem 3.3. *For some constants $C > 0$ and $r > 1$, the following holds. Suppose $\{X_t\}_{t\in T}$ is a Gaussian process, and let $t_1, t_2, \ldots, t_m \in T$ be such that $d(t_i, t_j) \geqslant \alpha$ for $i \neq j$. Then,*

$$g(T) \geqslant C\alpha\sqrt{\log m} + \min_{i=1,2,\ldots,m} g(B(t_i, \alpha/r)). \tag{8}$$

The proof of the preceding theorem relies on the a strong concentration property for Gaussian processes. First, we recall the classical isoperimetric inequality for Gaussian space. We remind the reader that for a function $F : \mathbb{R}^n \to \mathbb{R}$,

$$\|F\|_{\mathrm{Lip}} = \sup_{x\neq y\in\mathbb{R}^n} \frac{|F(x) - F(y)|}{\|x - y\|}.$$

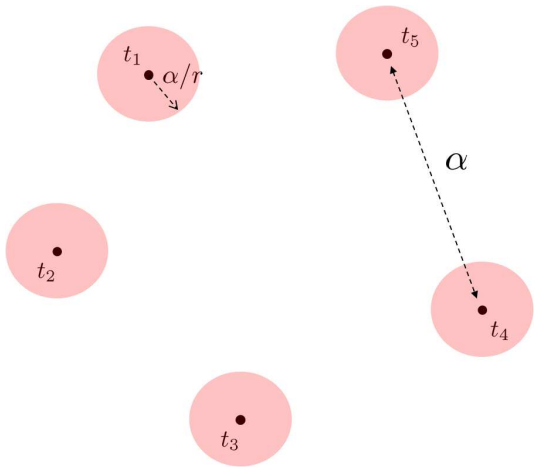

Figure 3: Setup for the Sudakov inequality.

Theorem 3.4. *Let $F : \mathbb{R}^n \to \mathbb{R}$, and let $\mu = \int F \, d\gamma_n$, where γ_n is the standard n-dimensional Gaussian measure. Then,*

$$\gamma_n \left(x \in \mathbb{R}^n : |F(x) - \mu| > \lambda \right) \leqslant 2 \exp\left(\frac{-\lambda^2}{2\|F\|_{\mathrm{Lip}}} \right). \tag{9}$$

Using this, we can prove the following remarkable fact.

Theorem 3.5. *Let $\{X_t\}_{t \in T}$ be a Gaussian processes, then*

$$\mathbb{P}\left(\left| \sup_{t \in T} X_t - \mathbb{E} \sup_{t \in T} X_t \right| > \lambda \right) \leqslant 2 \exp\left(\frac{-\lambda^2}{2 \sup_{t \in T} \mathbb{E}(X_t^2)} \right). \tag{10}$$

A notable aspect of this statement is that only the maximum variance affects the concentration, not the *number* of random variables. We now prove Theorem 3.5 using Theorem 3.4.

Proof. We will prove it in the case $|T| = n$, but of course our bound is independent of n. The idea is that given a Gaussian process $\{X_1, X_2, \ldots, X_n\}$, we can write

$$X_i = a_{i1}\, g_1 + a_{i2}\, g_2 + \cdots + a_{in}\, g_n,$$

for $i = 1, 2, \ldots, n$, where $\{g_i\}_{i=1}^n$ are standard i.i.d. normals, and the matrix $A = (a_{i,j})$ is a matrix of real coefficients. In this case, if $g = (g_1, g_2, \ldots, g_n)$ is a standard n-dimensional Gaussian, then the vector Ag is distributed as $(X_1, X_2, \ldots, X_n)$.

If we put $F(x) = \max\{(Ax)_i : i = 1, \ldots, n\}$, then Theorem 3.4 yields (10) as long as $\|F\|_{\mathrm{Lip}} \leqslant \max_i \sqrt{\mathbb{E}(X_i^2)}$. It is easy to see that

$$\|F\|_{\mathrm{Lip}} = \|A\|_{2 \to \infty} = \sup_{\|x\|_2 = 1} \|Ax\|_\infty.$$

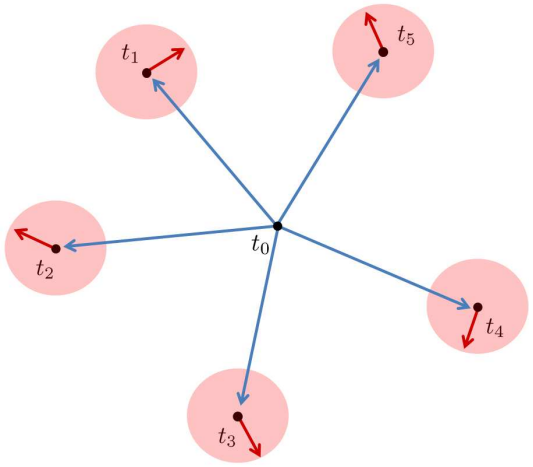

Figure 4: A "chaining" lower bound.

But $\|A\|_{2\to\infty}$ is just the maximum ℓ_2 norm of any row of A, and the ℓ_2 norm of row i is

$$\sqrt{\sum_{j=1}^{n} a_{ij}^2} = \sqrt{\mathbb{E}(X_i^2)}.$$

□

Using this theorem, we are ready to prove Theorem 3.3.

Proof of Theorem 3.3. Assume that the conditions of Theorem 3.3 hold. Pick an arbitrary $t_0 \in T$, and recall that we can write

$$g(T) = \mathbb{E}\sup_{t\in T} X_t = \mathbb{E}\sup_{t\in T}(X_t - X_{t_0})$$

since our gaussians are centered.

Now, for $i = 1, \dots, m$, let $Y_i = \sup_{t\in B(t_i,\alpha/r)} X_t - X_{t_i}$. Then,

$$\begin{aligned}
\mathbb{E}\sup_{t\in T} X_t &\geqslant \mathbb{E}\sup_{i\leqslant m}\left[(Y_i - \mathbb{E}Y_i) + \mathbb{E}Y_i + X_{t_i}\right] \\
&\geqslant \mathbb{E}\max\left(\{X_{t_1}, \dots, X_{t_m}\}\right) + \min_{i\leqslant m}\mathbb{E}(Y_i) - \mathbb{E}\max_{i\leqslant m}|Y_i - \mathbb{E}Y_i| \\
&\geqslant C\alpha\sqrt{\log m} + \min_{i\leqslant m} g(B(t_i,\alpha/r)) - \mathbb{E}\max_{i\leqslant m}\left|\sup_{t\in B(t_i,\alpha/r)}(X_t - X_{t_i}) - g(B(t_i,\alpha/r))\right|,
\end{aligned}$$

where in the last line we have used Theorem 3.1.

Now, for all $t \in B(t_i, \alpha/r)$, the variance of $X_t - X_{t_i}$ is at most $(\alpha/r)^2$, hence Theorem 3.5 implies that the final term is at most $c_0(\alpha/r)\sqrt{\log m}$ for some universal $c_0 > 0$. But this means that by choosing $r = 2CC_0$, we achieve

$$\mathbb{E}\sup_{t\in T} X_t \geqslant \frac{C}{2}\alpha\sqrt{\log m} + \min_{i\leqslant m} g(B(t_i,\alpha/r)),$$

which completes the proof. □

4 The majorizing measures theorem

We will now prove Talagrand's majorizing measures theorem, showing that the generic chaining bound is tight for Gaussian processes. The proof here will be a bit more long-winded than Talagrand's proof, but also (I think), quite a bit more accessible as well. Most importantly, we will highlight the key idea with a simple combinatorial argument.

First, let's recall the bound we proved earlier.

Theorem 4.1. *Let $\{X_t\}_{t\in T}$ be a Gaussian process, and let $T_0 \subseteq T_1 \subseteq \cdots \subseteq T$ be a sequence of subsets such that $|T_0| = 1$ and $|T_n| \leqslant 2^{2^n}$ for $n \geqslant 1$. Then,*

$$\mathbb{E} \sup_{t\in T} X_t \leqslant O(1) \sup_{t\in T} \sum_{n\geqslant 0} 2^{n/2}\, d(t, T_n). \tag{11}$$

In order to make things slightly easier to work with, we look at an essentially equivalent way to state (11). Consider a Gaussian process $\{X_t\}_{t\in T}$ and a sequence of increasing partitions $\{\mathcal{A}_n\}_{n\geqslant 0}$ of T, where increasing means that $\mathcal{A}_{n+1}$ is a refinement of $\mathcal{A}_n$ for $n \geqslant 0$. Say that such a sequence $\{\mathcal{A}_n\}$ is *admissible* if $\mathcal{A}_0 = \{T\}$ and $|\mathcal{A}_n| \leqslant 2^{2^n}$ for all $n \geqslant 1$. Also, for a partition P and a point $t \in T$, we will use the notation $P(t)$ for the unique set in P which contains t.

By choosing T_n to be any set of points with one element in each piece of the partition $\mathcal{A}_n$, (11) yields,

$$\mathbb{E} \sup_{t\in T} X_t \leqslant O(1) \sup_{t\in T} \sum_{n\geqslant 0} 2^{n/2} \operatorname{diam}(\mathcal{A}_n(t)). \tag{12}$$

We can now state our main theorem, which shows that this is essentially the only way to bound $\mathbb{E} \sup_{t\in T} X_t$.

Theorem 4.2. *There is a constant $L > 0$ such that for any Gaussian process $\{X_t\}_{t\in T}$, there exists an admissible sequence $\{\mathcal{A}_n\}$ which satisfies,*

$$\mathbb{E} \sup_{t\in T} X_t \geqslant L \sup_{t\in T} \sum_{n\geqslant 0} 2^{n/2} \operatorname{diam}(\mathcal{A}_n(t)). \tag{13}$$

Recall that for a subset $A \subseteq T$, we defined $g(A) = \mathbb{E} \sup_{t\in A} X_t$, and in the last post, we proved the following "Sudakov inequality."

Theorem 4.3. *For some constants $\kappa > 0$ and $r \geqslant 4$, the following holds. Suppose $\{X_t\}_{t\in T}$ is a Gaussian process, and let $t_1, t_2, \ldots, t_m \in T$ be such that $d(t_i, t_j) \geqslant \alpha$ for $i \neq j$. Then,*

$$g(T) \geqslant \kappa\alpha\sqrt{\log_2 m} + \min_{i=1,2,\ldots,m} g(B(t_i, \alpha/r)). \tag{14}$$

We will use only Theorem 4.3 and the fact that $g(A) \leqslant g(B)$ whenever $A \subseteq B$ to prove Theorem 4.2 (so, in fact, Theorem 4.2 holds with $\mathbb{E} \sup_{t\in T} X_t$ replaced by more general functionals satisfying an inequality like (14)).

4.1 The partitioning scheme

First, we will specify the partitioning scheme to form an admissible sequence $\{\mathcal{A}_n\}$, and then we will move on to its analysis. As discussed in earlier posts, we may assume that T is finite. Every set $C \in \mathcal{A}_n$ will have a value $\text{rad}(C)$ associated with it, such that $\text{rad}(C)$ is always an *upper bound* on the radius of the set C, i.e. there exists a point $x \in T$ such that $C \subseteq B(x, \text{rad}(C))$.

Initially, we set $\mathcal{A}_0 = \{T\}$ and $\text{rad}(T) = \text{diam}(T)$. Now, we assume that we have constructed $\mathcal{A}_n$, and show how to form the partition $\mathcal{A}_{n+1}$. To do this, we will break every set $C \in \mathcal{A}_n$ into at most 2^{2^n} pieces. This will ensure that

$$|\mathcal{A}_{n+1}| \leqslant 2^{2^n} \cdot |\mathcal{A}_n| \leqslant 2^{2^n} \cdot 2^{2^n} = 2^{2^{n+1}}.$$

Let r be the constant from Theorem 4.3. Put $m = 2^{2^n}$, and let $\Delta = \text{rad}(C)$. We partition C into m pieces as follows. First, choose $t_1 \in C$ which maximizes the value

$$g(B(t_1, \Delta/r^2) \cap C).$$

Then, set $C_1 = B(t_1, \Delta/r) \cap C$. We put $\text{rad}(C_1) = \Delta/r$.

Now we continue in this fashion. Let $D_\ell = C \setminus \bigcup_{i=1}^{\ell-1} C_i$ be the remaining space after we have cut out $\ell - 1$ pieces. For $\ell \leqslant m$, choose $t_\ell \in C$ to maximize the value

$$g(B(t_\ell, \Delta/r^2) \cap D_\ell).$$

For $\ell < m$, set $C_\ell = B(t_\ell, \Delta/r) \cap D_\ell$, and put $\text{rad}(C_\ell) = \Delta/r$.

So far, we have been chopping the space into smaller pieces. If $D_\ell = \emptyset$ for some $\ell \leqslant m$, we have finished our construction of $\mathcal{A}_{n+1}$. But maybe we have already chopped out $m - 1$ pieces, and still some remains. In that case, we put $C_m = D_m$, i.e. we throw everything else into C_m. Since we cannot reduce our estimate on the radius, we also put $\text{rad}(C_m) = \Delta$.

We continue this process until T is exhausted, i.e. eventually for some n large enough, $\mathcal{A}_n$ only contains singletons. This completes our description of the partitioning.

A remark: The whole idea here is that we have chosen the "largest possible piece," (in terms of g-value), but we have done this *with respect to the* Δ/r^2 *ball*, while we cut out the Δ/r ball. The reason for this will not become completely clear until the analysis, but we can offer a short explanation here. Looking at the lower bound (14), observe that the balls $B(t_i, \alpha/3)$ are disjoint under the assumptions, but we only get "credit" for the $B(t_i, \alpha/r)$ balls. When we apply this lower bound, it seems that we are throwing a lot of the space away. At some point, we will have to make sure that this thrown away part doesn't have all the interesting stuff! The reason for our choice of Δ/r vs. Δ/r^2 is essentially this: We want to guarantee that if we miss the interesting stuff at this level, then the *previous* level took care of it. To have this be the case, we will have to *look forward* (a level down), which (sort of) explains our choice of optimizing for the Δ/r^2 ball.

4.2 The tree

For the analysis, it will help to consider our partitioning process as having constructed a tree (in the most natural way). The root of the the tree is the set T, and its children are the sets of $\mathcal{A}_1$, and so on. Let's call this tree $\mathcal{W}$. It will help to draw and describe $\mathcal{W}$ in a specific way. First, we

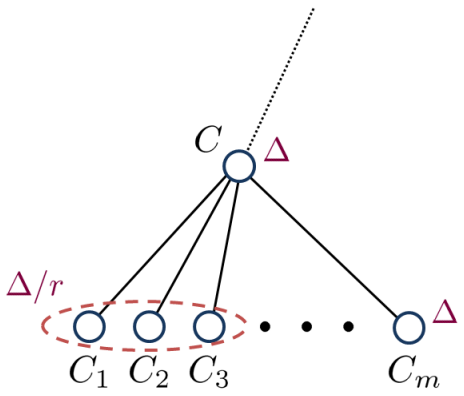

Figure 5: A piece of our tree $\mathcal{W}$.

will assign values to the edges of the tree. If $C \in \mathcal{A}_n$ and C_j is a child of C (i.e., $C_j \in \mathcal{A}_{n+1}$ and $C_j \subseteq C$), then the edge (C, C_j) is given value:

$$\kappa \cdot \frac{\operatorname{rad}(C)}{r} \cdot 2^{n/2}, \tag{15}$$

where κ and r are the constants from Theorem 4.3.

If we define the value of a root-leaf path in $\mathcal{W}$ as the sum of the edge lengths on that path, then for any $t \in T$,

$$\sum_{n \geqslant 0} 2^{n/2} \operatorname{diam}(\mathcal{A}_n(t)) \leqslant 2\frac{\kappa}{r} \left(\text{value of the path from the root to } t\right),$$

simply using $\operatorname{diam}(\mathcal{A}_n(t)) \leqslant 2 \operatorname{rad}(\mathcal{A}_n(t))$.

Thus in order to prove Theorem 4.2, which states that for some $L > 0$,

$$g(T) \geqslant L \sup_{t \in T} \sum_{n \geqslant 0} 2^{n/2} \operatorname{diam}(\mathcal{A}_n(t)),$$

it will suffice to show that for some (other) constant $L > 0$, for any root-leaf path P in $\mathcal{W}$, we have

$$g(T) \geqslant L \cdot \operatorname{value}(P). \tag{16}$$

Before doing this, we will fix a convention for drawing parts of $\mathcal{W}$ (see Figure 5). If a node $C \in \mathcal{A}_n$ has children $C_1, C_2, \ldots, C_m$, we will draw them from left to right. We will call an edge (C, C_m) a *right turn* and every other edge will be referred to as a *left turn*. Note that some node C may not have any right turn coming out of it (if the partitioning finished before the last step). Also, observe that along a left turn, the radius always drops by a factor of r, while along a right turn, it remains the same.

We now make two observations about computing the value $\operatorname{value}(P)$ up to a universal constant.

Observation (1): In computing the value of a root-leaf path P, we only need to consider right turns.

To see this, suppose that we have a right turn followed by a sequence of left turns. If the value of the right turn is $\frac{\kappa}{r}\Delta 2^{n/2}$, then the value of the following sequence of left turns is, in total, at most

$$\frac{\kappa}{r}\sum_{j=1}^{\infty} 2^{(n+j)/2}\frac{\Delta}{r^j} \leqslant O(1)\frac{\kappa}{r}\Delta 2^{n/2}.$$

In other words, because the radius decreases by a factor of r along every left turn, their values decrease geometrically, making the whole sum comparable to the preceding right turn. (Recall that $r \geqslant 4$, so indeed the sum is geometric.)

If the problem of possibly of having no right turn in the path P bothers you, note that we could artificially add an initial right turn into the root with value $\mathrm{diam}(T)$. This is justified since $g(T) \geqslant \frac{1}{2}\mathrm{diam}(T)$ always holds. A different way of saying this is that if the path really contained no right turn, then its value is $O(\mathrm{diam}(T))$, and we can easily prove (16).

Observation (2): In computing the value of a root-leaf path P, we need only consider the *last* right turn in any consecutive sequence of right turns.

Consider a sequence of right turns, and the fact that the radius does not decrease. The values (taking away the κ/r factor) look like $\Delta 2^{n/2}, \Delta 2^{(n+1)/2}, \Delta 2^{(n+2)/2}, \ldots$. In other words, they are geometrically increasing, and thus using only the last right turn in every sequence, we only lose a constant factor.

We will abbreviate last right turn to LRT, and write $\mathrm{value}_{\mathrm{LRT}}(P)$ to denote the value of P, just counting last right turns. By the two observations, to show (16) (and hence finish the proof), it suffices to show that, for every root-leaf path P in $\mathcal{W}$,

$$2 \cdot g(T) \geqslant \mathrm{value}_{\mathrm{LRT}}(P). \tag{17}$$

4.3 The analysis

Recall that our tree $\mathcal{W}$ has values on the edges, defined in (15). We will also put some natural values on the nodes. For a node C (which, recall, is just a subset $C \subseteq T$), we put $\mathrm{value}(C) = g(C)$. So the edges have values and the nodes have values. Thus given any subset of nodes and edges in $\mathcal{W}$, we can talk about the value of the subset, which will be the sum of the values of the objects it contains. We will prove (17) by a sequence of inequalities on subsets.

Fix a root-leaf path P, for which we will prove (17). Let's prove the fundamental inequality now. We will consider two consecutive LRTs along P. (If there is only one LRT in P, then we are done by the preceding remarks.) See the figure below. The dashed lines represent a (possibly empty) sequence of left turns and then right turns. The two LRTs are marked in Figure 6.

We will prove the following inequality, which is the heart of the proof. One should understand that the inequality is on the values of the subsets marked in red. The first subset contains two nodes, and the second contains two nodes an edge (see Figure 7).

With this inequality proved, the proof is complete. Let's see why. We start with the first LRT, and we have the inequality of Figure 8.

This gets us started. Now we apply the inequality of (Figure 7) repeatedly to each pair of consecutive LRTs in the path P. What do we have when we've exhausted the tree? Well, precisely all the LRTs in P are marked, yielding $2 \cdot g(T) \geqslant \mathrm{value}_{\mathrm{LRT}}(P)$, as desired.

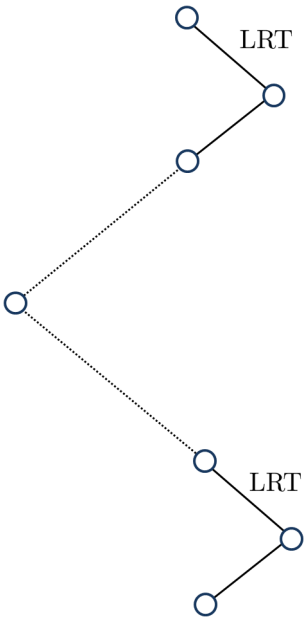

Figure 6: Two LRTs marked in a segment of a root-leaf path.

4.4 The LRT inequality

Now we are left to prove the inequality in Figure A. First, let's label some of the nodes; see Figure 9. Let $\Delta = \mathrm{rad}(C)$, and suppose that $C \in \mathcal{A}_n$. The purple values are not the radii of the corresponding nodes, but they are *upper bounds* on the radii, recalling that along every left turn, the radius decreases by a factor of r. Since there are at least two left turns in the picture, we get a Δ/r^2 upper bound on the radius of J.

Part of the inequality is easy: We have $g(A) \geqslant g(B)$ since $B \subseteq A$. So we can transfer the red mark from A to B. We are thus left to prove that

$$g(C) \geqslant \frac{\kappa}{r}\Delta 2^{n/2} + g(J). \tag{18}$$

This will allow us to transfer the red mark from C to the LRT coming out of C *and* to J.

When C was partitioned into $m = 2^{2^n}$ pieces, this was by our greedy partitioning algorithm using centers $t_1, t_2, \ldots, t_m$. Since we cut out the Δ/r ball around each center, we have $d(t_i, t_j) \geqslant \Delta/r$ for all $i \neq j$. Applying the Sudakov inequality (Theorem 4.3), we have

$$\begin{aligned} g(C) &\geqslant \kappa\frac{\Delta}{r}\sqrt{\log_2 m} + \min_{i=1,\ldots,m} g(B(t_i, \Delta/r^2)) \\ &= \frac{\kappa}{r}\Delta 2^{n/2} + \min_{i=1,\ldots,m} g(B(t_i, \Delta/r^2)) \\ &\geqslant \frac{\kappa}{r}\Delta 2^{n/2} + \min_{i=1,\ldots,m} g(B(t_i, \Delta/r^2) \cap D_i) \\ &= \frac{\kappa}{r}\Delta 2^{n/2} + g(B(t_m, \Delta/r^2) \cap D_m), \end{aligned}$$

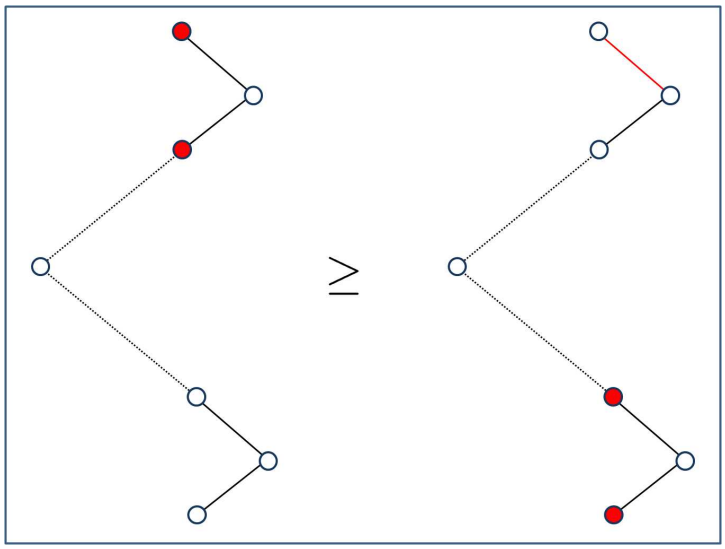

Figure 7: The main inequality on LRTs.

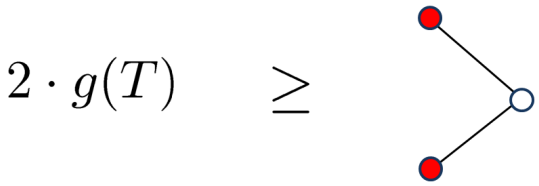

Figure 8: An inequality that holds for any two nodes.

where the last line follows from the greedy manner in which the t_i's were chosen.

But now we claim that

$$g(B(t_m, \Delta/r^2) \cap D_m) \geqslant g(J). \tag{19}$$

This follows from two facts. First, $J \subseteq D_m$ (since $D_m = C_m$ actually). Secondly, the radius of J is at most Δ/r^2! But t_m was chosen to *maximize* the value of $g(B(t_m, \Delta/r^2) \cap D_m)$ over all balls of radius Δ/r^2, so in particular its g-value is at least that of the Δ/r^2 ball containing J.

Combining (19) and the preceding inequality, we prove (18), and thus that the inequality of Figure A is valid. This completes the proof.

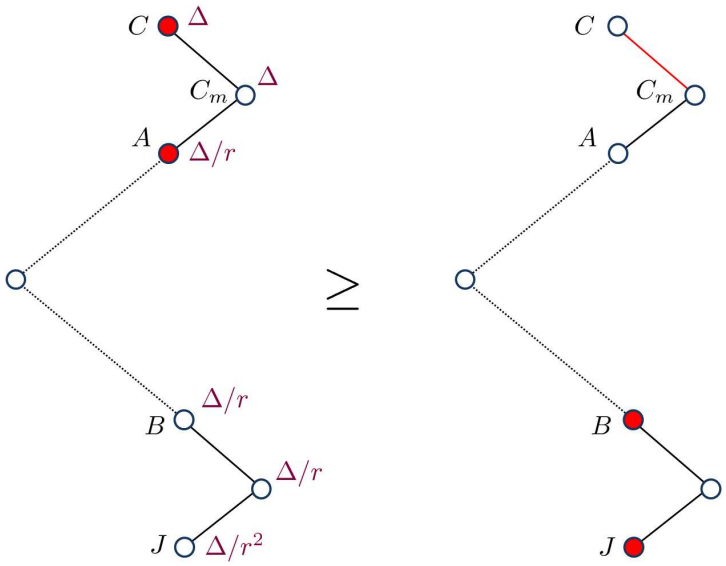

Figure 9: Proving the LRT inequality.