Notes for High-Dimensional Probability Second Edition by Roman Vershynin

Gallant Tsao

July 31, 2025

Contents

7 Random Processes

This chapter concerns mostly with random processes - collection random variables $(X_t)_{t\in T}$, which may be dependent. In calssical settings like Brownian motion, t represents time so $T \subset \mathbb{R}$. However, in high-dimensional probability T can be any set, and we'll deal with Gaussian processes a lot.
In this chapter, we'll explore powerful comparison inequalities for Gaussian processes - Slepian, Sudakov-Frenique, and Gordon - by using a new trick: Gaussian interpolation. Then we use these tools to prove a sharp bound on the operator norm of $m \times n$ Gaussian random matrices.
How does a Gaussian process $(X_t)_{t\in T}$ capture the geometry of T? We'll prove a lower bound on the Gaussian width using covering numbers, and link it to other ideas like effective dimension. Moreover, we'll also compute the size of a ranodm projection of any bounded set $T \subset \mathbb{R}^n$, which heavily depends on the Gaussian width.

7.1 Basic Concepts and Examples

Definition 7.1.1. A random process is a collection of random variables $(X_t)_{t\in T}$ on the same probability space, which are indexed by elements t of some index set T.

Example 7.1.2 (Discrete time)**.** If $T = \{1, \dots, n\}$ then the random process

$$(X_1, \dots, X_n)$$

can be identifies as a random vector in $\mathbb{R}^n$.

Example 7.1.3 (Random walks)**.** If $T = \mathbb{N}$, a discrete-time random process $(X_n)_{n\in\mathbb{N}}$ is simply a sequence of random variables. An important example is a *random walk* defined as

$$X_n := \sum_{i=1}^{n} Z_i,$$

where the increments Z_i are independent, mean zero random variables. See Figure 7.1 for an illustration:

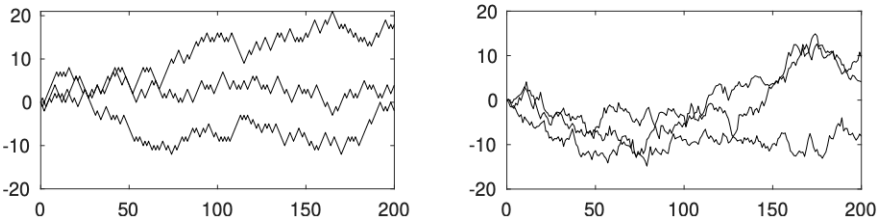


Figure 7.1 A few trials of a random walk (left) and standard Brownian motion (right).

Example 7.1.4 (Brownian motion)**.** The most classical continuous-time random process is the standard *Brownian motion* $(X_t)_{t\geq 0}$, or the *Wiener process*. It can be characterized as follows:

(i) The process has continuous sample paths, i.e. the random function $f(t) := X_t$ is continuous almost surely;

(ii) The increments are independent and satisfy $X_t - X_s \sim N(0, t-s)$ for all $t \geq s$.

Figure 7.1 above also shows some sample paths of a standard Brownian motion.

Example 7.1.5 (Random fields)**.** When the index set T is a subset of $\mathbb{R}^n$, a random process $(X_t)_{t\in T}$ is sometimes called a spatial random process, or *random field*. For example, the water temperature X_t ar the location on Earth that is parameterized by t can be modeled as a spatial random process.

7.1.1 Covariance and Increments

In section 3.2, we introduced the covariance matrix of a random vector. Here we'll define the *covariance function* of a random process $(X_t)_{t\in T}$ in a similar manner. For simplicity, assume the random process has zero mean:

$$\mathbb{E}[X_t] = 0 \text{ for all } t \in T.$$

The covariance function of the process is defined as

$$\Sigma(t,s) := \operatorname{Cov}(X_t, X_s) = \mathbb{E}[X_t X_s], \ t,s \in T.$$

The increments of the random process are defined as

$$d(t,s) := \|X_t - X_s\|_{L^2} = (\mathbb{E}\left[(X_t - X_s)^2\right])^{1/2}, \ t,s \in T.$$

Example 7.1.6. The increments of the standard Brownian motion satisfy

$$d(t,s) = \sqrt{t-s}, \ t \geq s$$

by definition. The increments of a random walk of Example 7.1.3 with $\mathbb{E}\left[Z_i^2\right] = 1$ behave similarly:

$$d(n,m) = \sqrt{n-m}, \ n \geq m.$$

Remark 7.1.7 (The canonical metric)**.** Even if the index set T has no geometric structure, the increments $d(t,s)$ always define a metric on T, thys automatically turning T into a metric space. However, as we see in Example 7.1.6, this metric may not match the Euclidean distance on $\mathbb{R}^n$.

Remark 7.1.8 (Covariance v.s. increments)**.** The covariance and the increments contain roughly the same information about the random process. Increments can be writeen using the covariance: Just expand the square to see that

$$d(t,s)^2 = \Sigma(t,t) - 2\Sigma(t,s) + \Sigma(s,s).$$

Vise versa, if the zero random variable belongs to the process, we can also recover the covariance from the increments (Exercise 7.1).

7.1.2 Gaussian Processes

Definition 7.1.9. A random process $(X_t)_{t\in T}$ is called a Gaussian process if, for any finite subset $T_0 \subset T$, the random vector $(X_t)_{t\in T_0}$ has a normal distribution. Equivalently, $(X_t)_{t\in T}$ is Gaussian if every finite linear combination $\sum_{t\in T_0} a_t X_T$ is a normal random variable (Exercise 3.16).

The notion of Gaussian processes generalized that of Gaussian random vectors in $\mathbb{R}^n$. A classical example of a Gaussian process is the standard Brownian motion.

Remark 7.1.10 (Distribution is determined by covariance, increments)**.** The distribution of a mean-zero Gaussian random vector in $\mathbb{R}^n$ is completely determined by its covariance matrix (Proposition 3.3.5). The same goes for a mean-zero Gaussian process: its distribution is determined by the covariance function $\Sigma(t,s)$, or equivalently by the increments $d(t,s)$, assuming the zero variable is part of the process.

Many tools we learned about random vectors can be applied to random processes. For example, Gaussian concentration (Theorem 5.2.3) applies:

Theorem 7.1.11 (Concentration of Gaussian processes)**.** Let $(X_t)_{t\in T}$ be a Gaussian process with finite T. Then

$$\|\sup_{t\in T} X_t - \mathbb{E}\left[\sup_t X_t\right]\|_{\psi_2} \le C \sup_{t\in T}\sqrt{\mathrm{Var}(X_t)}.$$

Proof. Exercise 5.9(b). □

Let's look at a broad class of Gaussian processes indexed by high-dimensional sets $T \subset \mathbb{R}^n$. Take a standard normal vector $g \sim N(0, I_n)$ and define

$$X_t := \langle g, t\rangle\,,\ t \in T.$$

This guves us a Gaussian process $(X_t)_{t\in T}$ called the *canonical Gaussian process*. The increments match the Euclidean distance:

$$\|X_t - X_s\|_{L^2} = \|t - s\|_2,\ t, s \in T.$$

Actually, one can realize any Gaussian process as the canonical process above because of the lemma below:

Lemma 7.1.12 (Gaussian random vectors)**.** Let X be a mean-zero Gaussian random vector in $\mathbb{R}^n$. Then there exist points $t_1, \dots, t_n$ such that

$$X \sim (\langle g, t_i\rangle)_{i=1}^n,\ \text{where } g \sim N(0, I_n).$$

Proof. IF Σ denotes the covariance matrix of X, then

$$X \equiv \Sigma^{1/2} g \text{ where } g \sim N(0, I_n).$$

The entries of $\Sigma^{1/2}g$ are $\langle t_i, g\rangle$ where the t_i are the rows of $\Sigma^{1/2}$. Done! □

It follows that for any Gaussian process $(X_s)_{x\in S}$, all finite-dimensional marginas $(X_s)_{s\in S_0}$, $|S_0| = n$ can be represented as the canonical Gaussian process indexed in a certain subset $T_0 \subset \mathbb{R}^n$.

7.2 Slepian, Sudakov-Fernique, and Gordon Inequalities

In many applications, it helps to have a *uniform* bound on a random process:

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] =?$$

Remark 7.2.1 (Making T finite)**.** To avoid measurability issues, let's think of

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \text{ as shorthand for } \sup_{T_0\subset T} \mathbb{E}\left[\max_{t\in T_0} X_t\right]$$

where T_0 runs over all finite subsets. The general case usually follows by approximation.

For some processes, this quantity can be computed exactly. For example, if (X_t) is a standard Brownian motion, the so-called reflection principle gives

$$\mathbb{E}\left[\sup_{t\le t_0} X_t\right] = \sqrt{\frac{2t_0}{\pi}} \text{ for every } t_0 \ge 0.$$

For general random processes - evern Gaussian - the problem is nontrivial.
The first general bound we prove is the Slepian comparison inequality for Gaussian processes. It basically says: the faster the process grows (in terms of the increments), the farther it gets.

Theorem 7.2.2 (Slepian inequality)**.** Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be two mean zero Gaussian processes. Assume that for all $t, s \in T$, we have

$$\mathbb{E}\left[X_t^2\right] = \mathbb{E}\left[Y_t^2\right] \text{ and } \mathbb{E}\left[(X_t - X_s)^2\right] \le \mathbb{E}\left[(Y_t - Y_s)^2\right].$$

Then $\sup_{t\in T} X_t$ is stochastically dominated by $\sup_{t\in T} Y_t$: For every $\tau \in \mathbb{R}$,

$$P\left(\sup_{t\in T} X_t \ge \tau\right) \le P\left(\sup_{t\in T} Y_t \ge \tau\right).$$

Consequently,

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le \mathbb{E}\left[\sup_{t\in T} Y_t\right].$$

We'll provide a proof later in the chapter, as we need some preliminary knowledge on Gaussian interpolation.

7.2.1 Gaussian Interpolation

Assume that T is finite; then we can loot at $X = (X_t)_{t\in T}$ and $Y = (Y_t)_{t\in T}$ as Gaussian random vectors in $\mathbb{R}^n$ with $n = |T|$. We may also assume that X and Y are independent.
Define the Gaussian random vector $Z(u)$ in $\mathbb{R}^n$ that continuously interpolates between $Z(0) = Y$ and $Z(1) = X$:

$$Z(u) := \sqrt{u}X + \sqrt{1-u}Y,\ u \in [0,1].$$

Then the covariance matrix of $Z(u)$ continuously interpolates linearly between the covariance matrices of Y and X:

$$\Sigma(Z(u)) = u\Sigma(X) + (1-u)\Sigma(Y).$$

This is because

$$\begin{aligned}
\Sigma(Z(u)) &= \mathbb{E}\left[Z(u)Z(u)^T\right] \\
&= \mathbb{E}\left[(\sqrt{u}X + \sqrt{1-u}Y)(\sqrt{u}X + \sqrt{1-u}Y)^T\right] \\
&= u\mathbb{E}\left[(X-\mu_X)(X-\mu_X)^T\right] + \sqrt{u(1-u)}\mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)^T\right] \\
&\quad + \sqrt{u(1-u)}\mathbb{E}\left[(Y-\mu_Y)(X-\mu_X)^T\right] + (1-u)\mathbb{E}\left[(Y-\mu_Y)(Y-\mu_Y)^T\right] \\
&= u\Sigma(X) + 0 + 0 + (1-u)\Sigma(Y) \quad \text{(Independence)} \\
&= u\Sigma(X) + (1-u)\Sigma(Y).
\end{aligned}$$

For a given function $f : \mathbb{R}^n \to \mathbb{R}$, let's study how $\mathbb{E}[f(Z(u))]$ changes as u increases from 0 to 1. Of special interest to us is the function

$$f(x) = \mathbf{1}_{\{\max_i x_i < \tau\}}.$$

We'll be able to show that in this case, $\mathbb{E}[f(Z(u))]$ increases in u. This would imply the conclusion of Slepian inequality, since then

$$\mathbb{E}[f(Z(1))] \ge \mathbb{E}[f(Z(0))] \implies P\left(\max_i X_i < \tau\right) \ge P\left(\max_i Y_i < \tau\right)$$

as claimed.
Let's start via the following useful identity:

Lemma 7.2.3 (Gaussian integration by parts)**.** Let $X \sim N(0,1)$. Then for any differentiable function $f : \mathbb{R} \to \mathbb{R}$ we have

$$\mathbb{E}[Xf(X)] = \mathbb{E}[f'(X)].$$

Proof. Assume first that f has bounded support. Denoting the Gaussian density by

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

we can express the expectation as an integral, and integrate it by parts:

$$\begin{aligned}\mathbb{E}\left[f'(X)\right] &= \int_{\mathbb{R}} f'(x)p(x)\ dx \\ &= \left[f(x)p(x)\right]_{-\infty}^{\infty} - \int_{\mathbb{R}} f(x)p'(x)\ dx \\ &= 0 - \int_{\mathbb{R}} f(x)p'(x)\ dx \\ &= -\int_{\mathbb{R}} f(x)p'(x)\ dx.\end{aligned}$$

We have already proved before (Exercise 2.3) that $p'(x) = -xp(x)$, hence the integral above equals

$$\int_{\mathbb{R}} f(x)p(x)x\ dx = \mathbb{E}\left[Xf(X)\right],$$

as claimed. The result can be extended to general functions by an approximation argument. The lemma is proved. □

By rescaling, we can extend Gaussian integration by parts for $X \sim N(0, \sigma^2)$:

$$\mathbb{E}\left[Xf(X)\right] = \sigma^2\mathbb{E}\left[f'(X)\right].$$

(Just write $X = \sigma Z$ for $Z \sim N(0,1)$ and apply Lemma 7.2.3). We can also extend it to high dimensions:

Lemma 7.2.4 (Multivariate Gaussian integration by parts)**.** Let $X \sim N(0, \Sigma)$. Then for any differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ we have

$$\mathbb{E}\left[Xf(X)\right] = \Sigma \cdot \mathbb{E}\left[\nabla f(X)\right]$$

assuming both expectations are finite. In other words,

$$\mathbb{E}\left[X_i f(X)\right] = \sum_{j=1}^{n} \Sigma_{ij}\mathbb{E}\left[\frac{\partial f}{\partial x_j}(X)\right],\ i = 1, \ldots, n.$$

Proof. Exercise 7.6. □

Lemma 7.2.5 (Gaussian interpolation)**.** Consider two independent Gaussian random vectors $X \sim N(0, \Sigma^X)$ and $Y \sim N(0, \Sigma^Y)$. Define the interpolation Gaussian vector

$$Z(u) := \sqrt{u}X + \sqrt{1-u}Y,\ u \in [0,1].$$

Then for any twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we have

$$\frac{d}{du}\mathbb{E}\left[f(Z(u))\right] = \frac{1}{2}\sum_{i,j=1}^{n}(\Sigma_{ij}^X - \Sigma_{ij}^Y)\mathbb{E}\left[\frac{\partial^2 f}{\partial x_i \partial x_j}(Z(u))\right],$$

assuming all expectations exist and are finite.

Proof. Using the multivariate chain rule,

$$\begin{aligned}\frac{d}{du}\mathbb{E}\left[f(Z(u))\right] &= \sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial f}{\partial x_i}(Z(u))\frac{dZ_i}{du}\right] \\ &= \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial f}{\partial x_i}(Z(u))\left(\frac{X_i}{\sqrt{u}} - \frac{Y_i}{\sqrt{1-u}}\right)\right].\end{aligned}$$

Let's break the sum above into two, and first compute the contribution of the terms containing X_i. To this end, we condition on Y and express

$$\sum_{i=1}^{n} \frac{1}{\sqrt{u}} \mathbb{E}\left[X_i \frac{\partial f}{\partial x_i}(Z(u))\right] = \sum_{i=1}^{n} \frac{1}{\sqrt{u}} \mathbb{E}\left[X_i g_i(X)\right] \quad (*),$$

where

$$g_i(X) = \frac{\partial f}{\partial x_i}(\sqrt{u}X + \sqrt{1-u}Y).$$

Apply the multivariate Gaussian integration by parts (Lemma 7.2.4), we get

$$\begin{aligned}\mathbb{E}\left[X_i g_i(X)\right] &= \sum_{j=1}^{n} \Sigma_{ij}^X \mathbb{E}\left[\frac{\partial g_i}{\partial x_j}(X)\right] \\ &= \sum_{j=1}^{n} \Sigma_{ij}^X \mathbb{E}\left[\frac{\partial^2 f}{\partial x_i \partial x_j}(\sqrt{u}X + \sqrt{1-u}Y)\right] \cdot \sqrt{u}.\end{aligned}$$

Substituting this into $(*)$ to get

$$\sum_{i=1}^{n} \frac{1}{\sqrt{u}} \mathbb{E}\left[X_i \frac{\partial f}{\partial x_i}(Z(u))\right] = \sum_{i,j=1}^{n} \Sigma_{ij}^X \mathbb{E}\left[\frac{\partial^2 f}{\partial x_i \partial x_j}(Z(u))\right].$$

Taking expectations on both sides with respect to Y, we left the conditioning on Y.
We can similarly evaluate the other sum (terms containing Y_i) by conditioning on X. Combining the two sums we complete the proof. □

7.2.2 Proof of Slepian Inequality

We'll establish a preliminary, functional form of Spelian's inequality first:

Lemma 7.2.6 (Slepian inequality, functional form)**.** Consider two mean zero Gaussian random vectors X, Y in $\mathbb{R}^n$. Assume that for all $i, j = 1, \ldots, n$, we have

$$\mathbb{E}\left[X_i^2\right] = \mathbb{E}\left[Y_i^2\right] \text{ and } \mathbb{E}\left[(X_i - X_j)^2\right] \leq \mathbb{E}\left[(Y_i - Y_j)^2\right].$$

Consider a twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$\frac{\partial^2 f}{\partial x_i \partial x_j} \geq 0 \text{ for all } i, j.$$

Then

$$\mathbb{E}\left[f(X)\right] \geq \mathbb{E}\left[f(Y)\right],$$

assuming both expectations exist and are finite.

Proof. The assumptions imply that the entries of the covariance matrices Σ^X and Σ^Y satisfy

$$\Sigma_{ii}^X = \Sigma_{ii}^Y \text{ and } \Sigma_{ij}^X \geq \Sigma_{ij}^Y$$

for all $i, j = 1, \ldots, n$. We can assume that X and Y are independent. Apply Lemma 7.2.5 and using our assumptions, we conclude that

$$\frac{d}{du}\mathbb{E}\left[f(Z(u))\right] \geq 0,$$

so $\mathbb{E}\left[f(Z(u))\right]$ increases in u. Then $\mathbb{E}\left[f(Z(1))\right] = \mathbb{E}\left[f(X)\right]$ is at least as large as $\mathbb{E}\left[f(Z(0))\right] = \mathbb{E}\left[f(Y).\right]$. This completes the proof. □

Now we are ready to prove Slepian's inequality (Theorem 7.2.2). Let's state and prove it in the equivalent form for Gaussian random vectors.

Theorem 7.2.7 (Slepian inequality)**.** Let X, Y be Gaussian random vectors as in Lemma 7.2.6. Then for every $\tau \geq 0$ we have

$$P\left(\max_{i\leq n} X_i \geq \tau\right) \leq P\left(\max_{i\leq n} Y_i \geq \tau\right).$$

Consequently,

$$\mathbb{E}\left[\max_{i\leq n} X_i\right] \leq \mathbb{E}\left[\max_{i\leq n} X_i\right].$$

Proof. Let $h : \mathbb{R} \to [0,1]$ be a twice-differentiable, non-increasing approximation to the indicator function on the interval $(-\infty, \tau)$:

$$h(x) \approx \mathbf{1}_{\{-\infty,\tau\}},$$

like in Figure 7.2 below.

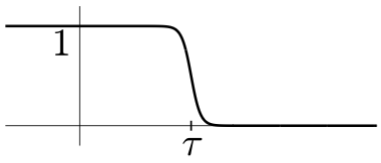


Figure 7.2 The function $h(x)$ is a smooth, non-increasing approximation to the indicator function $\mathbf{1}_{(-\infty,\tau)}$.

Define the function $f : \mathbb{R}^n \to \mathbb{R}$ by

$$f(x) = h(x_1) \cdots h(x_n) = \prod_{i=1}^{n} h(x_i).$$

Then $f(x)$ is an approximation to the indicator function

$$f(x) \approx \mathbf{1}_{\{\max_i x_i < \tau\}}.$$

We are looking to apply the functional form of Slepian inequality (Lemma 7.2.6) for $f(x)$.
To check the assumptions of this result, note that for $i \neq j$ we have

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = h'(x_i)h'(x_j) \cdot \prod_{k\notin\{i,j\}} h(x_k).$$

The first two terms are non-positive and the others are nonnegative by assumption, hence the second derivative is nonnegative, as required. It follows that

$$\mathbb{E}\left[f(X)\right] \geq \mathbb{E}\left[f(Y)\right].$$

By approximation, it implies

$$P\left(\max_{i\leq n} X_i < \tau\right) \geq P\left(\max_{i\leq n} Y_i < \tau\right).$$

This proves the first part. The second part follows by using the integrated tail formula in Exercise 1.15 (b):

$$\mathbb{E}\left[f(X)\right] = \int_0^\infty P\left(\max_{i\leq n} X_i \geq \tau\right)\, d\tau \leq \int_0^\infty P\left(\max_{i\leq n} Y_i \geq \tau\right)\, d\tau = \mathbb{E}\left[f(Y)\right].$$

□

7.2.3 Sudakov-Fernique and Gordon Inequalities

Slepian inequality has two assumptions on the processes (X_t) and (Y_t): the equality of varainces and the dominance of increments. We now remove the assumption on the equality of variances:

Theorem 7.2.8 (Sudakov-Fernique inequality)**.** Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be two mean zero Gaussian processes. Assume that for all $t, s \in T$, we have

$$\mathbb{E}\left[(X_t - X_s)^2\right] \le \mathbb{E}\left[(Y_t - Y_s)^2\right].$$

Then

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le \mathbb{E}\left[\sup_{t\in T} Y_t\right].$$

Proof. It is enough to prove this for Gaussian random vectors X and Y in $\mathbb{R}^n$, just like we did for Slepian's inequality in Theorem 7.2.7.
We again deduce the result from Gaussian Interpolation (Lemma 7.2.5). But this time, we'll approximate $f(x) \approx \max_i x_i$. Let $\beta > 0$ be a parameter and define the function

$$f(x) := \frac{1}{\beta} \log \sum_{i=1}^{n} e^{\beta x_i}.$$

We can check that indeed

$$\lim_{\beta\to\infty} f(x) = \max_{i=1,\dots,n} x_i.$$

Substituting $f(x)$ into the Gaussian interpolation formula and simplifying shows that (Exercise 7.7)

$$\frac{d}{du}\mathbb{E}\left[f(Z(u))\right] \le 0 \text{ for all } u \in [0,1].$$

Then we can finish the proof just like in Slepian's inequality. □

Gordon's inequality extends the Slepian and Sudakov-Frenique inequalities to the min-max setting:

Theorem 7.2.9 (Gordon's inequality)**.** Let $(X_{ut})_{u\in U, t\in T}$ and $(Y_{ut})_{u\in U, t\in T}$ be two mean-zero Gaussian processes indexed by pairs of points (u, t) in a product set $U \times T$. Assume that

$$\mathbb{E}\left[(X_{ut} - X_{us})^2\right] \le \mathbb{E}\left[(Y_{ut} - Y_{us})^2\right] \text{ for all } u, t, s;$$
$$\mathbb{E}\left[(X_{ut} - X_{vs})^2\right] \ge \mathbb{E}\left[(Y_{ut} - Y_{vs})^2\right] \text{ for all } u \ne v \text{ and all } t, s.$$

Then for every $\tau \ge 0$,

$$P\left(\inf_{u\in U}\sup_{t\in T} X_{ut} \ge \tau\right) \le P\left(\inf_{u\in U}\sup_{t\in T} Y_{ut} \ge \tau\right).$$

Moreover, by the integrated tail formula,

$$\mathbb{E}\left[\inf_{u\in U}\sup_{t\in T} X_{ut}\right] \le \mathbb{E}\left[\inf_{u\in U}\sup_{t\in T} Y_{ut}\right].$$

Proof. The proof under the additional assumption of equal variances is in Exercise 7.9. The proof for this statement is much harder. □

7.3 Application: Sharp Bounds for Gaussian Matrices

Let's pply the Gaussian comparison inequalities to random matrices. In Section 4.6, we used the ε -net argument to bound the expected operator norm like this:

$$\mathbb{E}\left[\|A\|\right] \le \sqrt{m} + C\sqrt{n}$$

where C is a constant (Exercise 4.41). Now, using the the Sudakov-Fernique inequality, we will tighten this bound for *Gaussian* random matrices and make $C = 1$.

Theorem 7.3.1 (Norms of Gaussian random matrices)**.** Let A be an $m \times n$ matrix with independent $N(0,1)$ entries. Then

$$\mathbb{E}\left[\|A\|\right] \leq \sqrt{m} + \sqrt{n}.$$

Proof. Let's write the norm of A as a supremum of Gaussian processes: By Definition 4.1.8,

$$\|A\| = \max_{u \in S^{n-1}, v \in S^{m-1}} \langle Au, v \rangle = \max_{(u,v) \in T} X_{uv}$$

where

$$T = S^{n-1} \times S^{m-1} \text{ and } X_{uv} := \langle Au, v \rangle \sim N(0,1).$$

To apply the Sudakov-Fernique comparison inequality (Theorem 7.2.8), let us compute the increments of the process (X_{uv}). For any $(u,v), (w,z) \in T$, we have

$$\begin{aligned}
\mathbb{E}\left[(X_{uv} - X_{wz})^2\right] &= \mathbb{E}\left[(\langle Au, v \rangle - \langle Au, z \rangle)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i,j} A_{ij}(u_j v_i - w_j z_i)\right)^2\right] \\
&= \sum_{i,j} (u_j v_i - w_j z_i)^2 \quad \text{(By independence, mean zero, variance 1)} \\
&= \|uv^T - wz^T\|_F^2 \\
&\leq \|u - w\|_2^2 - \|v - z\|_2^2 \quad \text{(By Exercise 7.10)}.
\end{aligned}$$

Now, let's define a simpler Gaussian process (Y_{uv}) with similar increments:

$$Y_{uv} := \langle g, u \rangle + \langle h, v \rangle, \ (u,v) \in T,$$

where $g \sim N(0, I_n)$ and $h \sim N(0, I_m)$ are independent Gaussian vectors. The increments of this process are

$$\begin{aligned}
\mathbb{E}\left[(Y_{uv} - Y_{wz})^2\right] &= \mathbb{E}\left[(\langle g, u - w \rangle + \langle h, v - z \rangle)^2\right] \\
&= \mathbb{E}\left[\langle g, u - w \rangle^2\right] + \mathbb{E}\left[\langle h, v - z \rangle^2\right] \quad \text{(By independence, mean 0)} \\
&= \|u - w\|_2^2 + \|v - z\|_2^2 \quad \text{(By normality of } g \text{ and } h).
\end{aligned}$$

Comparing the increments of the two processes, we see that

$$\mathbb{E}\left[(X_{uv} - X_{wz})^2\right] \leq \mathbb{E}\left[(Y_{uv} - Y_{wz})^2\right] \text{ for all } (u,v), (w,z) \in T,$$

as required in the Sudakov-Fernique inequality. Applying Theorem 7.2.8, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|A\|\right] &= \mathbb{E}\left[\sup_{(u,v) \in T} X_{uv}\right] \\
&\leq \mathbb{E}\left[\sup_{(u,v) \in T} Y_{uv}\right] \\
&= \mathbb{E}\left[\sup_{u \in S^{n-1}} \langle g, u \rangle\right] + \mathbb{E}\left[\sup_{v \in S^{m-1}} \langle h, v \rangle\right] \\
&= \mathbb{E}\left[\|g\|_2\right] + \mathbb{E}\left[\|h\|_2\right] \\
&\leq \left(\mathbb{E}\left[\|g\|_2^2\right]\right)^{1/2} + \left(\mathbb{E}\left[\|h\|_2^2\right]\right)^{1/2} \quad \text{(By Exercise 1.11)} \\
&= \sqrt{n} + \sqrt{m} \quad (\textit{ByProposition } 3.2.1(b)).
\end{aligned}$$

□

Theorem 7.3.1 is an expectation bound, but we can boost it to a high-probability bound using the concentration tools from Section 5.2:

Corollary 7.3.2 (Norms of Gaussian random matrices: tails)**.** Let A be an $m \infty n$ matrix with independent $N(0,1)$ entries. Then for every $t \geq 0$, we have

$$P(\|A\| \geq \sqrt{m} + \sqrt{n} + t) \leq 2\exp(-ct^2).$$

Proof. Let's combine the bound (Theorem 7.3.1) with Gaussian concentration (Theorem 5.2.3). Think of A as a long random vector in $\mathbb{R}^{n \infty n}$ by concatonating the rows. This makes A a standard normal random vector: $A \sim N(0, I_{nm})$. Consider the function

$$f(A) := \|A\|$$

that maps the vectorized matrix to the matrix's operator norm. Since the operator norm is bounded by the Frobenius norm, and the Frobenius norm is just the Euclidean norm on $\mathbb{R}^{m \times n}$, f is a Lipschitz function on $\mathbb{R}^{m \times n}$ with Lipschitz norm bounded by 1. Then Theorem 5.2.3 yields

$$P(\|A\| \geq \mathbb{E}[\|A\|] + t) \leq 2\exp(ct^2).$$

The bound on $\mathbb{E}[\|A\|]$ from Theorem 7.3.1 completes the proof. □

Aside from the result above, we have that:
A symmetric Gaussian matric satisfies (Exercise 7.11)

$$\mathbb{E}[\|A\|] \leq 2\sqrt{n},$$

and the smallest singular value of an $m \times n$ Gaussian matrix A satisfies (Exercise 7.13)

$$\mathbb{E}[\sigma_n(A)] \geq \sqrt{m} - \sqrt{n}.$$

7.4 Sudakov Inequality

Recall that for a general mean-zero Gaussian process $(X_t)_{t \in T}$ on some index set T, the increments

$$d(t,s) := \|X_t - X_s\|_{L^2} = (\mathbb{E}\left[(X_t - X_s)^2\right])^{1/2}$$

define a metric on T, called the *canonical metric*. This metric determines the covariance function $\Sigma(t,s)$, which in turn determines the distribution of the proces $(X_t)_{t \in T}$ (Remark 7.1.10). So, in theory, we can ask any question about the distribution of the process by understanding the geometry of the metric space (T,d) - studying probability via geometry!
Now the question comes: How can we estimate

$$\mathbb{E}\left[\sup_{t \in T} X_t\right]$$

in terms of the geometry of (T,d)? This is a hard problem we will study from now well into Chapter 8. We'll start with a lower bound in terms of the *metric entropy*, which was introduced in Chapter 4. Recall that for any $\varepsilon > 0$, the *covering number*

$$\mathcal{N}(T,d,\varepsilon)$$

is the samllest cardinality of an ε-net of T in the metric d, or equivalently the smallest number of closed balls of radius ε whose union covers T. The logarithm of the of the covering number, $\log_2 \mathcal{N}(T,d,\varepsilon)$, is called the *metric entropy* ot T.

Theorem 7.4.1 (Sudakov's inequality)**.** Let $(X_t)_{t \in T}$ be a mean-zero Gaussian process. Then, for any $\varepsilon \geq 0$, we have

$$\mathbb{E}\left[\sup_{t \in T} X_t\right] \geq c\varepsilon\sqrt{\log \mathcal{N}(T,d,\varepsilon)}$$

where d is the canonical metric defined above.

Proof. We'll deduce the result from the Sudakov-Frenique comparison inequality (Theorem 7.2.8). Assume that

$$N := \mathcal{N}(T, d, \varepsilon)$$

is finite; the infinite case is in Exercise 7.14. Let $\mathcal{N}$ be a maximal ε-seperated subset of T. Then $\mathcal{N}$ is an ε-net of T (Lemma 4.2.6), and thus

$$|\mathcal{N}| \geq N.$$

Restricting the process to $\mathcal{N}$, we see that it suffices to show that

$$\mathbb{E}\left[\sup_{t\in\mathcal{N}} X_t\right] \geq c\varepsilon\sqrt{\log N}.$$

Let's do it by comparing $(X_t)_{t\in\mathcal{N}}$ to a simpler Gaussian process $(Y_t)_{t\in\mathcal{N}}$, defined as follows:

$$Y_t := \frac{\varepsilon}{\sqrt{2}} g_t \text{ where } g_t \sim_{i.i.d.} N(0,1).$$

To use the Sudakov-Fernique comparison inequality (Theorem 7.2.8), we need to compare the increments of the two processes. Fix two different points $t, s \in \mathcal{N}$. By definition,

$$\mathbb{E}\left[(X_t - X_s)^2\right] = d(t,s)^2 \geq \varepsilon^2$$

while

$$\mathbb{E}\left[(Y_t - Y_s)^2\right] = \frac{\varepsilon^2}{2}\mathbb{E}\left[(g_t - g_s)^2\right] = \varepsilon^2 \quad (g_t - g_s \sim N(0,2)).$$

This implies that

$$\mathbb{E}\left[(X_t - X_s)^2\right] \geq \mathbb{E}\left[(Y_t - Y_s)^2\right] \text{ for all } t, s \in \mathcal{N}.$$

By applying Theorem 7.2.8, we obtain

$$\mathbb{E}\left[\sup_{t\in\mathcal{N}} X_t\right] \geq \mathbb{E}\left[\sup_{t\in\mathcal{N}} X_t\right] = \frac{\varepsilon}{2}\mathbb{E}\left[\max_{t\in\mathcal{N}} g_t\right] \geq c\varepsilon\sqrt{\log N}.$$

In the last step, we used that the expected maximum of N i.i.d $N(0,1)$ random variables is at least $c\sqrt{\log N}$ (Exercise 2.38 (b)). The proof is complete. □

7.4.1 Application for covering numbers in $\mathbb{R}^n$

Sudakov's inequality can be used to bound the covering numbers of an arbitrary set $T \subset \mathbb{R}^n$:

Corollary 7.4.2 (Sudakov inequality in $\mathbb{R}^n$)**.** Let $T \subset \mathbb{R}^n$. Then for any $\varepsilon > 0$,

$$\mathbb{E}\left[\sup_{t\in T} \langle g, t\rangle\right] \geq c\varepsilon\sqrt{\log \mathcal{N}(T,\varepsilon)},$$

where $\mathcal{N}(T,\varepsilon)$ just the covering number of T.

Proof. Consider the canonical Gaussian process $X_t := \langle g, t\rangle$ where $g \sim N(0, I_n)$. As we noted in Section 7.1.2, the canonical distance for this process is the Euclidean distance in $\mathbb{R}^n$, i.e.

$$d(t,s) = \|X_t - X_s\|_{L^2} = \|t - s\|_2 \text{ for any } t, s \in T.$$

Then the corollary directly follows from Sudakov's inequality (Theorem 7.4.1). □

Aside from the bound above, Corollary 7.4.2 is also sharp up to a log factor (Exercise 8.5):

$$\mathbb{E}\left[\sup_{t\in T} \langle g, t\rangle\right] \leq C\log(n) \cdot \varepsilon\sqrt{\log \mathcal{N}(T,\varepsilon)}.$$

For a quick application of Sudakov's inequality, let's (roughly) re-derive the boudne on covering numbers of polytopes in $\mathbb{R}^n$ from Corollary 0.1.1:

Corollary 7.4.3 (Covering numbers of polytopes)**.** Let P be a polytope in $\mathbb{R}^n$ with N vertices, contained in the unit Euclidean ball. Then for every $\varepsilon > 0$, we have

$$\mathcal{N}(P, \varepsilon) \leq N^{C/\varepsilon^2}.$$

Proof. If $x_1, \ldots, X_N$ are the vertices of P, then

$$\mathbb{E}\left[\sup_{t \in P} \langle g, t \rangle\right] \leq \mathbb{E}\left[\sup_{i=1,\ldots,N} \langle g, x_i \rangle\right] \leq C\sqrt{\log N}.$$

The first bound follows from the maximal principle (Exercise 1.4): Since P lies the convex hull of its vertices, for each fixed g, the linear (and thus convex) function $t \mapsto \langle g, t \rangle$ attains its maximum at a vertex. The second bound is due to the maximal inequality from Proposition 2.7.6, as $\langle g, x \rangle \sim N(0, \|x\|_2^2)$ and $\|x\|_2 \leq 1$. Substitute this into Corollary 7.4.2 and simplify completes the proof. □

7.5 Gaussian Width

From the previous subsection, we saw an important quantity associated with any set $T \subset \mathbb{R}^n$: the size of the canonical Gaussian process on T. It shows up a lot in high-dimensional probability, so let's give it a name and look at its basic properties.

Definition 7.5.1. The Gaussian width of a subset $T \subset \mathbb{R}^n$ is defined as

$$w(T) := \mathbb{E}\left[\sup_{t \in T} \langle g, x \rangle\right] \text{ where } g \sim N(0, I_n).$$

Try to think of Gaussian width as a fundamental geometric measure of a set $T \subset \mathbb{R}^n$, like volume or surface area.

Proposition 7.5.2 (Simple properties of Gaussian width)**.** (a) (Finiteness) $w(T)$ is finite if and only if T is bounded.

(b) (Invariance) $w(UT + y) = w(T)$ for any orthogonal matrix U and vector y.

(c) (Convex hulls) $w(\text{conv}(T)) = w(T)$.

(d) (Minkowski addition and scaling) $w(T + S) = w(T) + w(S)$ and $w(aT) = aw(T)$ for any $T, S \subset \mathbb{R}^n$ and $a \in \mathbb{R}$.

(e) (Symmetry)

$$w(T) = \frac{1}{2} w(T - T) = \frac{1}{2}\mathbb{E}\left[\sup_{x,y \in T} \langle g, x - y \rangle\right].$$

(f) (Width and diameter)

$$\frac{1}{\sqrt{2\pi}} \cdot \text{diam}(T) \leq w(T) \leq \frac{\sqrt{n}}{2} \cdot \text{diam}(T).$$

(g) (Linear maps) For any $m \times n$ matrix A, $w(AT) \leq \|A\| w(T)$.

Proof. Only the proof for (f) is demonstrated here, with the rest left to Exercise 7.15.
For the lower bound, fix any $x, y \in T$. Since both $x - y$ and $y - x$ are in $T - T$, property (e) gives

$$w(T) \geq \frac{1}{2}\mathbb{E}\left[\max(\langle x - y, g \rangle, \langle y - x, g \rangle)\right] = \frac{1}{2}\mathbb{E}\left[|\langle x - y, g \rangle|\right] = \frac{1}{2}\sqrt{\frac{2}{\pi}}\|x - y\|_2.$$

The last equality holds since $\langle x - y, g \rangle \sim N(0, \|x - y\|_2)$ and $\mathbb{E}[|X|] = \sqrt{2/\pi}$ for $X \sim N(0, 1)$. Taking the supremum over all $x, y \in T$ gives the result.

For the upper bound, use property (e) again to get

$$w(T) \le \frac{1}{2}\mathbb{E}\left[\sup_{x,y\in T}\langle g, x-y\rangle\right] \le \frac{1}{2}\mathbb{E}\left[\sup_{x,y\in T}\|g\|_2\|x-y\|_2\right] \le \frac{1}{2}\mathbb{E}\left[\|g\|_2\right]\cdot \operatorname{diam}(T).$$

Since $\mathbb{E}\left[\|g\|_2\right] \le \mathbb{E}\left[\|g\|_2^2\right]^{1/2} = \sqrt{n}$, the proof is complete. □

Remark 7.5.3 (Width and diameter)**.** Both upper and lower bounds in Proposition 7.5.2 (f) are optimal and the $O(\sqrt{n})$ gap between them cannot be improved (Exercise 7.16). So, diameter is not a great way to capture Gaussian width.

7.5.1 Geometric Meaning of Width

Gaussian width has a nice geometric meaning: it's about how wide the set $T \subset \mathbb{R}^n$ looks in random directions. The width of T in the direction $\theta \in S^{n-1}$ is the width of the smallest slab (between parallel hyperplanes orthogonal to θ) that contains T (See Figure 7.3 below), which can be expressed as $\sup_{x,y\in T}\langle \theta, x-y\rangle$.

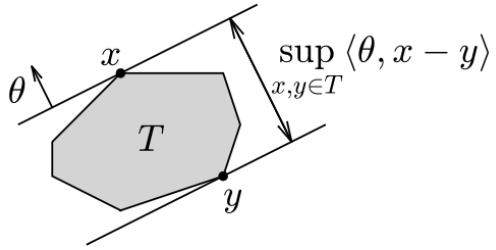


Figure 7.3 The width of a set $T \subset \mathbb{R}^n$ in the direction of a unit vector θ.

If we average the width over all unit directions θ, we get the following definition:

Definition 7.5.4. The spherical width of a set $T \subset \mathbb{R}^n$ is

$$w_s(T) := \mathbb{E}\left[\sup_{x\in T}\langle \theta, x\rangle\right] \text{ where } \theta \sim \operatorname{Unif}(S^{n-1}).$$

The only difference between the Gaussian and spherical widths is in the random vectors we average over: $g \sim N(0, I_n)$ versus $\theta \sim \operatorname{Unif}(S^{n-1})$. Both are rotation invariant, but g is approximately $\sqrt{n}$ times longer than θ. Thus we get

Lemma 7.5.5 (Gaussian v.s. spherical widths)**.** Ths Gaussian width is approximately $\sqrt{n}$ times the spherical width:

$$\left(\sqrt{n} - \frac{C}{\sqrt{n}}\right) w_s(T) \le w(T) \le \sqrt{n}\, w_s(T).$$

Proof. Express the Gaussian vector g throughits length and direction: $g = r\theta$, where $r = \|g\|_2$ and $\theta = g/\|g\|_2$. Now, $\theta \sim \operatorname{Unif}(S^{n-1})$ is independent of r (Exercise 3.22). Thus

$$w(T) = \mathbb{E}\left[\sup_{x\in T}\langle r\theta, x\rangle\right] = \mathbb{E}\left[r\right]\cdot \mathbb{E}\left[\sup_{x\in T}\langle \theta, x\rangle\right] = \mathbb{E}\left[\|g\|_2\right]\cdot w_s(T).$$

Then by using concentration of the norm (Exercise 3.2), this gives

$$\sqrt{n} - \frac{C}{\sqrt{n}} \le \mathbb{E}\left[\|g\|_2\right] \le \sqrt{n},$$

which completes the proof. □

7.5.2 Examples

Example 7.5.6 (Euclidean ball and sphere)**.** The Gaussian widths of the unit ball and sphere are

$$w(S^{n-1}) = w(B_2^n) = \mathbb{E}\,[\|g\|_2] = \sqrt{n} \pm \frac{C}{\sqrt{n}},$$

where we used concentration of the norm (Exercise 3.2) for the last step. The spherical width of these sets of course equal to 1.

Example 7.5.7 (Cube)**.** The unit ball of the ℓ^∞ norm in $\mathbb{R}^n$ is the cube $[-1,1]^n$. So, by using the duality formula

$$\max\{\langle x, y\rangle : \ y \in B_{p'}^n\} = \|x\|_p,$$

we get

$$w(B_\infty^n) = \mathbb{E}\,[\|g\|_1] = \mathbb{E}\,[|g_1|] \cdot n = \sqrt{\frac{2}{\pi}} \cdot n.$$

Example 7.5.8 (Cross-polytope)**.** The unit ball of the ℓ^1 norm in $\mathbb{R}^n$ is the cross-polytope

$$B_1^n = \{x \in \mathbb{R}^n : \ \|x\|_1 \le 1\}.$$

Its Gaussian width satisfies

$$w(B_1^n) \asymp \sqrt{\log N}$$

where the notation $\asymp$ hides the absolute constant factors. This is because

$$w(B_1^n) = \mathbb{E}\,[\|g\|_\infty] = \mathbb{E}\left[\max_{i=1,\dots,n} |g_i|\right],$$

where the first equation uses duality. Then the result follows from Exercise 2.38 (b).

Example 7.5.9 (Finite point sets)**.** Any finite set of points $T \subset \mathbb{R}^n$ satisfies

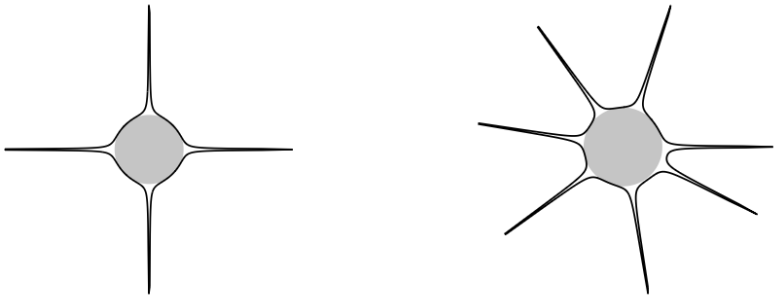
$$w(T) \le C\sqrt{\log |T|} \cdot \operatorname{diam}(T).$$

To prove this, we can assume that $\operatorname{diam}(T) = 1/2$ (by rescaling), and that T lies in the unit Euclidean ball (by translation). Then the result follows from the bound provided in Corollary 7.4.3.

Remark 7.5.10 (Surprising behavior of width in high dimensions)**.** As we can see from Example 7.5.6 to Example 7.5.8, the Gaussian width of the cube B_∞^n is roughly (up to a constant factor) the same as that of its *circumscribed ball* $\sqrt{n}B_2^n$. But for the cross-polytope B_1^n, the width is roughly (up to a log factor) like that of its *inscribed ball* $\frac{1}{\sqrt{n}}B_2^n$, whiich is tiny! Why?

The cube B_∞^n has so many vertices (2^n) that in most directions it sticks out to roughly the circumscribed ball, which drives the width. But the cross-polytops B_1^n only has $2n$ vertices, so a random direction $g \sim N(0, I_n)$ is likely to be far from all of them. The width is not only driven by those lonely $2n$ "spikes" - it's driven by the "bulk". which is roughly the inscribed ball.

Figure 7.4a shows Milmen's *hyperbolic sketch* of B_1^n, highlighting how the bulk (the inscribed ball) dominates since the set has few vertices (spikes). We can make similar sketches to general convex sets too (Figure 7.4b) - they are great for building high-dimensional intuition, even if we lose convexity in the picture.



(a) The octahedron B_1^n (b) General convex set

Figure 7.4 V. Milman's hyperbolic sketch of high-dimensional convex sets

7.5.3 Gaussian Complexity and Effective Dimension

There are also a number of helpful cousins of the Gaussian width $w(T)$. Normally, we would take the expected max of $\langle g, t\rangle$, but sometime it's easier to work with L^1 or L^2 averages:

$$w(T) = \mathbb{E}\left[\sup_{x\in T}\langle g, x\rangle\right],\ \gamma(T) := \mathbb{E}\left[\sup_{x\in T}|\langle g, x\rangle|\right],\ h(T) := \left(\mathbb{E}\left[\sup_{x\in T}\langle g, x\rangle^2\right]\right)^{1/2}.$$

where $g \sim N(0, I_n)$. We call $\gamma(T)$ the *Gaussian complexity* of T. Clearly,

$$w(T) \le (T) \le h(T).$$

The reverse bounds are basically true too:

Lemma 7.5.11 (ALmost equivalent versions of Gaussian width)**.** For any bounded set $T \subset \mathbb{R}^n$, we have:

(a) $\gamma(T - T) = 2w(T)$.

(b) $h(t) \asymp \gamma(T) \asymp w(T) + \|y\|_2$ for any point $y \in T$.

In particular, if T contains the origin, all three versions are equivalent:

$$h(t) \asymp \gamma(T) \asymp w(T).$$

Proof. (a) follows from Proposition 7.5.2 (e), since $T - T$ is origin-symmetric.
(b) Let's prove the first equivalence here, and we'll leave the second equivalence to Exercise 7.20. We trivially have $\gamma(T) \le h(T)$. For the reverse, look at the function $z \mapsto \sup_{x\in T}|\langle z, x\rangle|$ on $\mathbb{R}^n$. Its LIpschitz norm is bounded by the radius

$$\sup_{x\in T}\|x\|_2 = r(T).$$

Then by Gaussian concentration (Theorem 5.2.3),

$$\|\sup_{x\in T}|\langle g, x\rangle| - \gamma(T)\|_{\psi_2} \lesssim r(T).$$

So by the triangle inequality and Proposition 2.6.6 (ii), we get

$$h(T) = \|\sup_{x\in T}|\langle g, x\rangle|\|_{L^2} \lesssim \|\sup_{x\in T}|\langle g, x\rangle|\|_{\psi_2} \lesssim \gamma(T) + r(T) \lesssim \gamma(T)$$

where in the last step, we used the fact that $\gamma(T) \gtrsim r(T)$, which comes from the second part of (b) - just take the supremum over $y \in T$. □

The Gaussian width helps us define a robust version of the notion of dimension. The usual linear-algebraic dimension of a set $T \subset \mathbb{R}^n$, which is the dimension of the smallest affine space containing it, can be susceptible to tiny perturbations of T. Here is a more robust alternative:

Definition 7.5.12. The effective dimension of a bounded set $T \subset \mathbb{R}^n$ is

$$d(T) := \frac{h(T-T)^2}{\operatorname{diam}(T)^2} \asymp \frac{w(T)^2}{\operatorname{diam}(T)^2}.$$

The equivalence follows from Lemma 7.5.11. The effective dimensions is bounded above by the linear-algebraic one:

$$d(T) \leq \dim(T),$$

with equality when T is a Euclidean ball in some subspace (Exercise 7.21). Unlike the usual dimension, the effective dimension is stable - small perturbations to T only slightly change its width and diameter.

7.6 Application: Random Projection of Sets

What happens if we project a set $T \subset \mathbb{R}^n$ onto a random m-dimensional subspace in $\mathbb{R}^n$ (picked uniformly from the Grassmannian $G_{m,n}$)? We might view this like dimensionality redution, like in the Johnson-Lindenstrauss lemma. What can we say about the size (diameter) of the projected set PT where P is the random projection?
For a finite set T, the Johnsen-Lindenstrauss lamm (Theorem 5.3.1) says that if $m \gtrsim \log|T|$, the random projection P acts essentially as a scaling of T:

$$\operatorname{diam}(PT) \approx \sqrt{\frac{m}{n}} \operatorname{diam}(T).$$

But if the cardinality of T is too large or infinite, the above may fail. For instance, if $T = B_2^n$ is the unit Euclidean ball, no projection can shrink its size:

$$\operatorname{diam}(PT) = \operatorname{diam}(T).$$

What about for general sets T? The next result states that a random projection cannor shrink T beyond its spherical width $w_s(T)$:

Theorem 7.6.1 (Sizes of random projections of sets)**.** Let $T \subset \mathbb{R}^n$ be a bounded set, and P be the orthogonal projection in $\mathbb{R}^n$ onto a random m-dimensional subspace $E \sim \operatorname{Unif}(G_{n,m})$. Then

$$\mathbb{E}\left[\operatorname{diam}(PT)\right] \asymp w_s(T) + \sqrt{\frac{m}{n}} \operatorname{diam}(T),$$

where the notation $\asymp$ hides positive absolute constant factors.

Proof. We'll prove the upper bound here. The lower bound is in Exercise 7.26.
Step 1: Change the model. Let's switch the view just like in the proof of Lemma 5.3.2. A random subspace $E \subset \mathbb{R}^n$ can be obtained by randomly rotating some fixed subspace, such as $\mathbb{R}^m$. But instead of fixing T and randomly rotating $\mathbb{R}^m$, we can fix $E = \mathbb{R}^m$ and randomly rotate T. A random rotation of a vector $x \in T$ is Ux where $U \sim \operatorname{Unif}(O(n))$ is a random orthogonal matrix. Projecting Ux onto $E = \mathbb{R}^m$ means keeping the first m coordinates, i.e. Qx where Q is the $m \times n$ matrix consisting of the first m columns of U. So, we can work with Q instead of P.
Step 2: Approximation. Without loss of generality, assume $\operatorname{diam}(T) \leq 1$. We need to bound

$$\operatorname{diam}(QT) = \sup_{x \in T-T} \|Qx\|_2 = \sup_{x \in T-T} \max_{z \in S^{m-1}} \langle Qx, z \rangle .$$

We will proceed with an ε-net argument as in the proof of Theorem 4.4.3. Choose an $(1/2)$-net $\mathcal{N}$ of the sphere S^{m-1} so that

$$|\mathcal{N}| \leq 5^m$$

using Corollary 4.2.11. We can replace the supremum over the sphere S^{m-1} by the supremum over the net $\mathcal{N}$ paying a factor of 2 (Execise 4.35):

$$\operatorname{diam}(QT) \leq 2 \sup_{x \in T-T} \max_{z \in \mathcal{N}} \langle Qx, z \rangle = 2 \max_{z \in \mathcal{N}} \sup_{x \in T-T} \left\langle Q^T z, x \right\rangle .$$

Now, here is the plan: we will first bound

$$\sup_{x \in T-T} \langle Q^T z, x \rangle \quad (*)$$

for a fixed $z \in \mathcal{N}$, and then take the union bound over all z.

Step 3: Concentration. Fix $z \in \mathcal{N}$. By construction, $Q^T z$ is uniformly distributed on the sphere: $Q^T z \sim \text{Unif}(S^{n-1})$ (Exercise 7.24). The expectation can be expressed as the spherical width:

$$\mathbb{E}\left[\sup_{x \in T-T} \langle Q^T z, x \rangle\right] = w_s(T-T) = 2w_s(T).$$

(The last equality if just a spherical version of Proposition 7.5.2 (e)). To check that $(*)$ concentrates around its mean, we use the concentration inequality on the sphere (Theorem 5.1.3). Since $\text{diam}(T) \le 1$ by assumption, the function $z \mapsto \sup_{x \in T-T} \langle z, x \rangle$ on the sphere has Lipschitz norm at most 1. So we get

$$P\left(\sup_{x \in T-T} \langle Q^T z, x \rangle \ge 2w_s(T) + t\right) \le 2\exp(-cnt^2).$$

Step 4: Union bound. Now we unfix $z \in \mathcal{N}$ by taking the union bound:

$$P\left(\max_{z \in \mathcal{N}} \sup_{x \in T-T} \langle Q^T z, x \rangle \ge 2w_s(T) + t\right) \le |\mathcal{N}| \cdot 2\exp(-cnt^2).$$

Recall that $|\mathcal{N}| \le 5^m$. Choosing $t = Cs\sqrt{m/n}$ with constant C large enough, the probability above is bounded by $2e^{-ms^2}$ for any $s \ge 1$. Therefore, we get

$$P\left(\frac{1}{2}\text{diam}(QT) \ge 2w_s(T) + Cs\sqrt{\frac{m}{n}}\right) \le e^{-ms^2} \text{ for any } s \ge 1.$$

From this, we can bound the expected value of $\text{diam}(QT)$ using the integrated tail formula Lemma 1.6.1, and the proof is complete. □

Remark 7.6.2 (Phase transition)**.** Let's get more insight from Theorem 7.6.1. Since the sum of two terms is equivalent to maximum (up to a factor of 2), we can write:

$$\text{diam}(PT) \asymp \max\left[w_s(T), \sqrt{\frac{m}{n}}\text{diam}(T)\right].$$

Let's find the "phase transition" point where these two terms are equal. Set them to be equal and solving for m, we get

$$m = \frac{(\sqrt{n}w_s(T))^2}{\text{diam}(T)^2} \asymp \frac{w(T)^2}{\text{diam}(T)^2} \asymp d(T),$$

using Lemma 7.5.5 and the definition of effective dimension $d(T)$ (Definition 7.5.12). So the take-away:

$$\text{diam}(PT) \asymp \begin{cases} \sqrt{\frac{m}{n}}\text{diam}(T) & \text{if } m \ge d(T) \\ w_s(T) & \text{if } m \le d(T). \end{cases}$$

See figure 7.5 below.

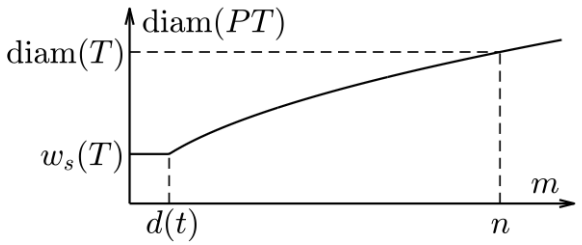


Figure 7.5 The diameter of a random m-dimensional projection of a set T as a function of m.

As we decrease the dimension m of the random projection, initially it shrinks t by roughly $\sqrt{m/n}$ as stated in the Johnson-Lindenstrauss lemma. But once m dips below the effective dimension $d(T)$, the shrinking stops and the diameter stays near the spherical width $w_s(T)$. This is because $\text{conv}(PT)$ looks like a ball of radius $w_s(T)$, as we will see in Section 9.7.2.