# *Introduction*

At its heart, approximation theory is concerned with understanding and quantifying the extent to which one function can be approximated by elements of another *given* collection of functions. As one would expect, it plays a central role in many fields, *inter alia* numerical analysis, differential equations, statistics, and data science. It is intimately related to the geometry of function spaces and compressibility. Indeed, a motif throughout these notes is the role that smoothness plays in the effective dimension of a function space, understood in a suitable sense.

It has a long history, dating at least back to Euler's 1777 work on minimizing errors in distances on maps of Russia; continuing through Laplace's work in 1843 on determining the best ellipsoidal approximation to the Earth; and Bernstein's work on constructive function theory in the 20th century, to today, where it underpins modern approaches to everything from PDE solvers, fast linear algebra, signal processing, and neural networks. In all of this, a key figure is Pafnuty Chebyshev, who in 1853 considered the following problem which arose in the study of locomotives:

*Given an interval* $[a, b]$ *a function* $f : [a, b] \to \mathbb{R}$, *and a natural number* $n$, *solve*

$$\min_{p \in \mathcal{P}_n} \max_{a \leq x \leq b} |f(x) - p(x)|,$$

*where* $\mathcal{P}_n$ *denotes the set of polynomials of degree* at most $n$.

Understanding this innocuous-looking optimization problem will be a central focus of these notes and, as we shall see, has far ranging implications, extensions, and applications. When considering an approximation problem, there are three fundamental mathematical questions to consider: what functions are we approximating? With

what? In what sense? Additionally, from a computational perspective, when assessing a method for approximation, there are three fundamental computational traits that are desirable for a method to possess: is it accurate? Is it numerically stable? Can it be carried out in a reasonable amount of time?

## *A motivating example*

As a first example that will highlight many of the themes of these notes, let us first consider the example of numerical integration. That is to say, we want to obtain information ($\int_0^1 f(x)\,\mathrm{d}x$) from samples. This leads to a problem of *quadrature*.

One of the simplest approaches to approximate such an integral would be through either the left-hand or right-hand Riemann rules:

$$L_n(f) := \frac{1}{n}\sum_{k=0}^{n-1} f(k/n), \quad R_n(f) := \frac{1}{n}\sum_{k=1}^{n} f(k/n),$$

respectively. Note that both of these approximations take the general form

$$\sum_{k=0}^{n} f(x_k)w_k,$$

where $x_k = k/n$ and $w_k = 1/n$ except at the right endpoint (for $L_n$) or the left endpoint (for $R_n$). The $x_k$ are called *quadrature points* and the $w_k$ are referred to as the corresponding *quadrature weights.* The error is clearly dependent on the smoothness of $f$. Indeed, it is easy to construct a function for which the integral, in a Lebesgue sense, exists and is equal to zero, but the left-hand and right-hand Riemann rules with equispaced points take on any desired value for all $n$.

Thus, we must buy convergence at the price of imposing more assumptions on our function $f$. One way to think about this is that the space of all functions, even all bounded functions, is too large to allow for numerical integration. Instead, we might consider restricting our attention to a subset of functions which are almost low dimensional. Indeed, approximation theory is fundamentally about function "compression" - the extent to which a given set of functions can be well-approximated by a finite dimensional subspace. One mechanism for a function space to be compressible is smoothness. An old game in numerical analysis (or analysis in general) is to buy decay at the cost of derivatives.

Returning to our example, we see that from Taylor's theorem, if $f$ is smooth enough, then

$$|R_n(f) - I(f)| \leq C\|f'\|_\infty n^{-1}.$$

Here $\|\cdot\|_\infty$ denotes the $L^\infty$ norm, $C$ is a constant independent of $f$, and $I(f)$ is the true integral of $f$

$$I(f) := \int_0^1 f(x)\,\mathrm{d}x.$$

A similar result holds for the left-hand sum. These rules have a classical interpretation: if the integral denotes the (signed) area under the curve $y = f(x)$, then $R_n(f)$ is the area under the piecewise constant approximation to $f$ which is equal to $f(k/n)$ for all $x \in ((k-1)/n, k/n]$, $k = 1, \ldots, n$.

Using this intuition, generalizing is rather straightforward. Rather than piecewise constant, we could instead use a piecewise linear approximant, giving us the trapezoid rule:

$$T_n(f) := \frac{f(0)+f(1)}{2n} + \frac{1}{n}\sum_{k=1}^{n-1} f(k/n).$$

For smooth enough functions this looks much better. And it is! Indeed, one can easily show that

$$|T_n(f) - I(f)| \leq C\|f''\|_\infty n^{-2}.$$

**Exercise 1.** *Prove this. Note that the increase in accuracy relies on smoothness.*

**Exercise 2.** *Can you find a counterexample when the function is differentiable, but not twice differentiable?*

Clearly, one could continue on in this way. But, there is a more subtle point that the geometric picture misses. Looking at the formulae for the three rules, we get

$$T_n(f) = \frac{L_n(f) + R_n(f)}{2},$$

and so the trapezoid rule is the average of the left and right. So far, so reasonable. Generically speaking, one "overshoots" the other "undershoots" and the two errors almost cancel. But, returning to the formula for $T_n$,

$$T_n(f) = \frac{f(0/n)}{2n} + \frac{1}{n}\sum_{k=1}^{n-1} f(k/n) + \frac{f(n/n)}{2n}.$$

In particular, only the endpoints are affected. From this, we conclude that almost all of the error comes only from the endpoints. How should we interpret this?

*A first approach: Fourier series*

In the following, we will assume that $f$ is smooth, and it, along with all of its derivatives, is periodic (i.e. $f(0) = f(1)$, $f'(0) = f'(1)$, $\cdots$). Then it can be shown that $f$ can be expressed as a Fourier series

$$f(x) = \sum_{m\in\mathbb{Z}} e^{2\pi imx} f_m,$$

where

$$f_m = \int_0^1 e^{-2\pi imx} f(x)\,\mathrm{d}x.$$

In general, one needs much less regularity for the Fourier representation to exist. Later in these notes we will revisit decompositions of this type, exploring under what conditions they are defined and in what sense they converge to $f$. For now, we will not worry too much about these issues.

After substituting our expression for $f$ into the formula for $T_n(f)$, we find that

$$T_n(f) = \sum_{j=0}^{n-1} \left( \sum_{m\in\mathbb{Z}} e^{2\pi imx_j} f_m \right) \frac{1}{n} = \sum_{m\in\mathbb{Z}} \frac{f_m}{n} \sum_{j=0}^{n-1} \left(e^{2\pi im/n}\right)^j.$$

Here $x_j = j/n$, and we have used the periodicity of $f$ to equate the contributions from $j = 0$ and $j = n$. In particular, as a first sign that there is something interesting happening, for equispaced points the left-hand Riemann rule, the right-hand Riemann rule, and the trapezoid rule all give the same result.

The inner sum on the right-hand side of the previous expression is a truncated geometric series, and can be explicitly summed, which gives

$$\sum_{j=0}^{n-1} \left(e^{2\pi im/n}\right)^j = \begin{cases} n, & e^{2\pi im/n} = 1, \\ \frac{e^{2\pi im}-1}{e^{2\pi im/n}-1}, & e^{2\pi im/n} \neq 1. \end{cases}$$

So,

$$T_n(f) = \sum_{m\in n\mathbb{Z}} f_m.$$

Now, $I(f) = f_0$ and hence

$$E_n(f) := |T_n(f) - I(f)| = \left| \sum_{m\in n\mathbb{Z},\, m\neq 0} f_m \right|.$$

Thus, the speed of the convergence is directly related to the rate of decay of the Fourier coefficients. How does this relate to smoothness?

We begin by observing that if we take the expression for $f_m$ and integrate by parts, for $m \neq 0$,

$$f_m = -\frac{1}{2\pi im} \int_0^1 \frac{\mathrm{d}}{\mathrm{d}x} \left(e^{-2\pi imx}\right) f(x)\,\mathrm{d}x = \frac{1}{2\pi im} \int_0^1 e^{-2\pi imx} f'(x)\,\mathrm{d}x.$$

The boundary conditions arising in the integration by parts disappear because of the periodicity of $f$ and its derivatives. Indeed, now we see an explanation for the heuristic observation that the error is dominated by the endpoints. If $f$ is not periodic then we must keep the boundary terms, which would in turn contribute to slower decay in the Fourier coefficients.

Naturally, if $f$ is smooth enough, then we can iterate, to obtain

$$f_m = \frac{1}{(2\pi i m)^k} \int_0^1 e^{-2\pi i m x} f^{(k)}(x)\, \mathrm{d}x.$$

This formula yields the following bound for $f_m$

$$|f_m| \leq \frac{1}{2\pi |m|^k} \|f^{(k)}\|_1,$$

and hence

$$E_n \leq \frac{\|f^{(k)}\|_1}{(2\pi)^k} \sum_{j \in \mathbb{Z}, j \neq 0} \frac{1}{|jn|^k} \leq \frac{2\|f^{(k)}\|_1}{(2\pi n)^k} \sum_{j=1}^{\infty} \frac{1}{j^k}.$$

The last sum is independent of $f$ and $n$, and converges for all $j > 1$. A crude bound can be obtained by comparing it to the integral of $1/x^k$ on $[1, \infty)$, which gives the upper bound $k/(k-1)$. So, for all $n \geq 2$,

$$E_n \leq C \frac{\|f^{(k)}\|_1}{(2\pi n)^k}.$$

Here we see the trade-off between smoothness and integration error very clearly.

*A second approach: the Euler-Maclaurin formula*

For our next approach, we will use the Euler-Maclaurin formula to bound the error in our quadrature rules. To do so, we need a little bit of notation. Given integers $m, n$ and an integrable function $g : [m, n] \to \mathbb{R}$, we set

$$I_{m,n}(g) = \int_m^n g(x)\, \mathrm{d}x$$

and

$$S_{m,n}(g) = g(m+1) + \cdots + g(n).$$

**Theorem 1.** *If $k \in \mathbb{N}$ and $g \in C^k([m, n])$ then*

$$S_{m,n} - I_{m,n} = \sum_{k=1}^{k} \frac{B_j}{j!} \left[ f^{(j-1)}(n) - f^{(j-1)}(m) \right] + R_k.$$

*Here, $B_k$ are the Bernoulli numbers (see Table 1). The remainder term, $R_k$, is bounded by*

$$|R_k| \leq \frac{2\zeta(k)}{(2\pi)^k} \int_m^n |f^{(k)}(x)|\, \mathrm{d}x$$

*if $k$ is even. For odd $k$ the Riemann zeta function, $\zeta(k)$, can be omitted.*

| $n$ | $B_n$ |
|---|---|
| 0 | 1 |
| 1 | 1/2 |
| 2 | 1/6 |
| 4 | -1/30 |
| 6 | 1/42 |
| 8 | -1/30 |
| 10 | 5/66 |
| 12 | -691/2730 |
| 14 | 7/6 |
| 16 | -3617/510 |
| 18 | 43867/798 |
| 20 | -174611/330 |
| ⋮ | ⋮ |

Table 1. *Bernoulli numbers. All odd Bernoulli numbers greater than one are zero.*

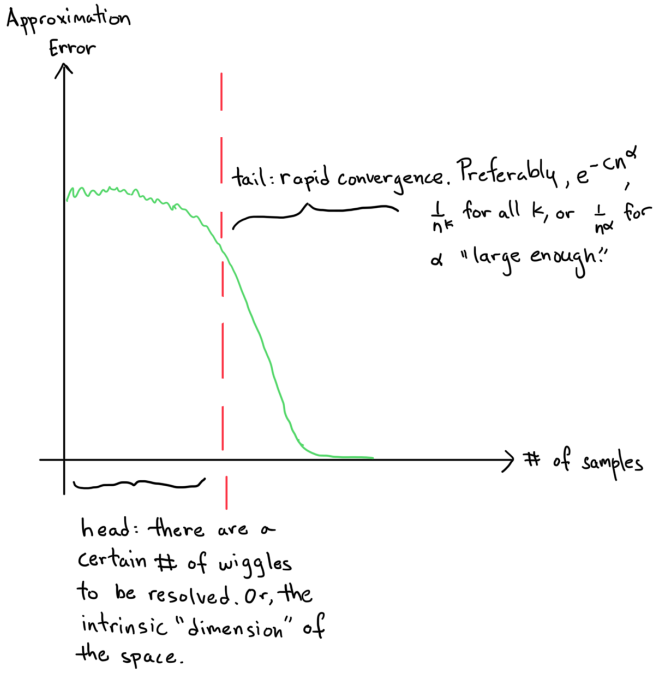
Rescaling to $[0,1]$ and setting $k=2$, $m=0$, we see that

$$T_n(f) - I(f) = \frac{n^{-2}}{12}|f'(1) - f'(0)| + R_2 \in O(n^{-2}).$$

Again, higher-order error terms come from the failure of derivatives of $f$ to be periodic.

### *Summary*

For periodic, smooth functions, the trapezoid, left-hand, and right-hand rules are the gold standard. They converge faster than any power of $n$ (though one has to be a little bit careful with interpreting this). Schematically, the approximation error of an ideal scheme looks like the plot below.




For the trapezoid rule, we use differentiability to get compressibility. Finally, the error is not spread uniformly throughout the interval.

## *Additional Exercises*

**Exercise 3.** *Assuming that $f \in C^4(0,1)$, construct an improved quadrature rule which computes the integral of $f$ on $(0,1)$ with an error which is $O(h^4)$. Your rule should only involve a constant number of modifications to the standard left-hand or right-hand Riemann sums as $n \to \infty$. Implement your quadrature rule and plot the convergence for: i) $f(x) = e^x$*

*and ii) $f(x) = \sqrt{x}$ (of course this isn't differentiable up to the boundary!). Compare your numerical results with those obtained from Simpson's rule – using a cubic approximation on each subinterval, where the cubic is chosen to agree with $f$ at the two endpoints and the midpoint of each subinterval.*

# *The Basics of Bases*

In this chapter, our goal will be to formulate means of describing, classifying, and quantifying spaces of functions and notions of distance therein. The properties of the functions we wish to approximate and those with which we approximate them, determine the rate of convergence, the algorithm, and the manner of the proofs. Generic theorems (or algorithms) work for broad classes of functions but then do not typically exploit the unique structure of a specific problem. The same considerations apply to measures of distance, our definition of "success".

More importantly, in applications there is typically a natural "budget" for accuracy of certain quantities, which induce natural objective functions for approximation problems. Do we care about outliers (i.e. catastrophic failure due to rare events, machining parts, etc.)? Or, do we care about the average, or a weighted average, being small (like antenna design)?

One of the most natural and convenient ways of specifying collections of functions is through *bases*. Every finite dimensional linear vector space, $E$, has a basis - a maximal linearly independent collection of vectors in $E$. This generalizes (rather unsatisfyingly) to infinite dimensions in the following way.

**Definition 1.** *A subset $\mathcal{X}$ of a vector space $E$ is a Hamel basis if every vector $v \in E$ can be expressed uniquely as a finite linear combination of elements in $\mathcal{X}$. That is to say, for every $v \in E$, there exist $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, $v_1, \ldots, v_n \in \mathcal{X}$ such that*

$$v = \sum_{i=1}^{n} \alpha_i v_i.$$

Without more, we are limited to finite sums. In particular, we need a notion of convergence to use infinite sums. Even for relatively pedestrian spaces, $\mathcal{X}$ is typically uncountable.

**Proposition 1.** *Every linear vector space has a Hamel basis.*

*Proof.* A somewhat sterile application of Zorn's lemma which we leave to the reader. Note that linear independence involves only finite linear combinations. □

Still, this gives us our first approximation result. We now impose additional structure on our vector spaces, with the hope of obtaining more useful results. In particular, we now focus on the case in which $E$ is a Banach space.

## *Moving to Banach spaces*

**Definition 2.** *A sequence $\{v_j\}_1^\infty$ in a Banach space $X$ is called a Schauder basis if for all $v \in X$ there exist unique coefficients $\alpha_1, \alpha_2, \dots$ such that*

$$v = \sum_{j=1}^{\infty} \alpha_j v_j,$$

*where implicitly we require the sum on the right-hand side in the last expression to converge.*

It is easy to show that $X$ must be separable to have a Schauder basis. Unfortunately, this is not sufficient (though finding counterexamples is non-trivial - see Enflo 1973).

The next result allows us to control the size of the partial sums in terms of the norm of $v$.

**Theorem 2.** *Suppose $\{v_j\}_1^\infty$ is a Schauder basis of the Banach space $X$. Then there exists a constant $M$ such that for all $v \in X$,*

$$\left\| \sum_{j=1}^{N} \alpha_j v_j \right\| \leq M \|v\|$$

*for all $N = 1, 2, \cdots$. Here, the $\alpha_j$'s are the coefficients such that*

$$v = \sum_{j=1}^{\infty} \alpha_j v_j.$$

*The minimal $M$ for which the above inequality holds for all $M$ and $N$ is called the* basis constant.

*Proof.* We define a new space $E$ and show that it is isomorphic to $X$. In particular, set

$$E := \left\{ \{\alpha_j\}_1^\infty \mid \sum_{j=1}^{\infty} \alpha_j v_j \text{ converges in } X \right\}.$$

Moreover, for $\alpha \in E$, set

$$\|\alpha\|_E := \sup_{N \in \mathbb{N}} \left\| \sum_{j=1}^{N} \alpha_j v_j \right\|.$$

It is easily verified that, equipped with this norm, $E$ is a Banach space. Moreover, clearly there is a bounded bijection from $E$ to $X$. Thus, by the inverse mapping theorem, this is an isomorphism. □

Now, for each positive integer $N$ we can define the operator $S_N : X \to \mathbb{R}$ by

$$S_N(v) := \sum_{j=1}^{N} \alpha_j v_j,$$

where again $\{v_j\}_1^\infty$ is a Schauder basis and the $\{\alpha_j\}_1^\infty$ are the corresponding coefficients of $v$. Note that the $\alpha_j$ are independent of $N$. Clearly $S_N$ is a projection onto the span of $\{v_1, \ldots, v_N\}$ and, by the previous theorem, all the $S_N$'s are uniformly bounded. It follows that that the coefficients are linear functionals on $X$, called *biorthogonal functionals*, and are denoted by $v_j^*$. Hence,

$$v = \sum_{j=1}^{\infty} v_j v_j^*(v).$$

Clearly,

$$\|v_j\||v_j^*(v)| = \|v_j v_j^*(v)\| = \|\|S_j(v) - S_{j-1}(v)\| \leq 2M\|v\|$$

and so $\|v_j\|\,\|v_j^*\| \leq 2M$ for all $j$.

This gives a slick characterization of compactness.

**Theorem 3.** *A subset $A \subset X$ is compact if and only if $A$ is bounded and the basis expansions converge uniformly, i.e.*

$$\|v - S_N(v)\| \leq \epsilon_N \to 0,$$

*for all $v \in A$.*

Note here that the $\epsilon_N$ do not depend on $v$.

*Proof. Necessity:* from above, we know that $S_N$ converges to the identity operator pointwise. Using a standard $\epsilon$-net argument, since convergence is uniform on compact sets, it follows that $S_N$ converges uniformly to the identity on $A$.
*Sufficiency:* For any $\epsilon$, there exists an $N$ such that $\|v - S_N(v)\| \leq \epsilon$ for all $v \in A$. The image of $S_N$ is a finite dimensional subspace and $A \subset \mathcal{N}_\epsilon(S_N)$ ($A$ is contained within an $\epsilon$-neighborhood of the image of $S_N$). □

We now impose even more mathematical structure to obtain stronger results. For Banach spaces we added a notion of length. Going to Hilbert spaces, we now also add a notion of angle.

## *Hilbert spaces*

**Definition 3.** *A set $\{v_j\}_1^\infty$ in $H$ (a Hilbert space) is an orthogonal system if $\langle v_j\,,\, v_k\rangle = 0$ for all $j \neq k$. It is orthonormal if additionally $\|v_j\| = 1$ for all $j$.*

**Theorem 4.** *Let $\{v_j\}_1^\infty$ be an orthogonal system. The following are equivalent:*

*i* $\sum_j \alpha_j v_j$ *converges in* $H$

*ii* $\sum_j |\alpha_j|^2 \|v_j\|^2 < \infty$

*iii* $\sum_j \alpha_j v_j$ *converges unconditionally*

*If there is convergence then*

$$\|\sum_j \alpha_j v_j\|^2 = \sum_j |\alpha_j|^2 \|v_j\|^2.$$

*Proof.* We leave the proof as an exercise. □

**Definition 4.** *Let $\{v_j\}_1^\infty$ be an orthonormal system in $H$. The* Fourier series *of $v$ with respect to $\{v_j\}_1^\infty$ is*

$$\sum_{j=1}^{\infty} \langle v_j, v\rangle v_j.$$

*The scalar quantities $\langle v_j, v\rangle$ are called the* Fourier coefficients.

We end this section with several useful results about orthonormal systems on Hilbert spaces.

**Lemma 1.** *If $\{v_j\}_1^\infty$ is an orthonormal system then for each $N$, $S_N$ is an orthogonal projection.*

**Lemma 2** (Bessel's inequality)**.** *If $\{v_j\}_1^\infty$ is again an orthonormal system in a Hilbert space $H$,*

$$\sum_{j=1}^{\infty} |\langle v_j, v\rangle|^2 \leq \|v\|^2.$$

We also have the following optimality result.

**Lemma 3.** *Among all convergent sequences $s = \sum_j \alpha_j v_j$, $\|v - s\|$ is minimized by the Fourier series of $v$. The same is true of partial sums.*

Thus, finding a best approximation from a subspace is easy in a Hilbert space provided we have an orthonormal system. Here we should add that "best" is being measured with respect to the norm of the Hilbert space. A natural follow-up question is how to find orthonormal systems. If $H$ is separable, it's easy, at least in practice. Just apply Gram-Schmidt to the countably dense set.

## *Projections*

Projection operators arise naturally in approximation theory. Here we briefly review some general theory of projections which will be useful throughout these notes.

**Definition 5.** *Given a Banach space $E$, the linear operator $P : E \to E$ is called a* projection *if it is bounded and* idempotent, *i.e. $P^2 = P$.*

The following theorem gives a handy characterization of projections.

**Theorem 5.** *Suppose $P : E \to E$ is a projection and $\mathcal{R}(P) = V$. Then $V$ is a closed subspace of $E$, and $V = \ker(I - P)$. Here $I$ is the identity operator on $E$.*

*Proof.* To prove the first assertion, note that $I - P$ is continuous, so its kernel is closed. Thus, it suffices to show that $V = \ker(I - P)$. To see this, observe that if $v \in V$, then there exists a $u \in E$ such that $Pu = v$. Then $Pv = P^2u = Pu = v$. Thus $v \in \ker(I - P)$ and hence $V \subseteq \ker(I - P)$. Similarly, if $w \in \ker(I - P)$ then $Pw = w$ and so $w \in V$, implying the reverse inclusion. □

Given a projection $P$, there are a few other natural projections we can construct.

**Proposition 2.** *Let $P : E \to E$ be a projection. Then $(I - P) : E \to E$ is also a projection. So is the adjoint of $P$, $P^* : E^* \to E^*$.*

*Proof.* Both are clearly linear and bounded. For idempotency, for $I - P$, we observe that $(I - P)^2 = I - 2P + P^2 = I - P$. For the adjoint, recall that it is defined by: $(P^*\ell)(v) = \ell(Pv)$, for all $v \in E$. Hence, $((P^*)^2\ell)(v) = (P^*\ell)(Pv) = \ell(P^2v) = \ell(Pv) = (P^*\ell)(v)$. □

Given two projections, $P, Q : E \to E$ it is natural to ask whether it is possible to combine them. Unfortunately, in general, neither $PQ$ nor $QP$ are projections. The following result gives some conditions on which certain new projections can be created from two existing ones.

**Proposition 3.** *Let $P, Q : E \to E$ be two projections. Define the operator $P \oplus Q = P + Q - PQ$. Suppose $PQP = QP$. Then*

1. *$P \otimes Q$ is a projection onto $\mathcal{R}(P) + \mathcal{R}(Q)$.*
2. *$QP$ is a projection onto $\mathcal{R}(P) \cap \mathcal{R}(Q)$.*

*Proof.* The proof follows by a direct calculation.

$$\begin{aligned}(P \oplus Q)^2 &= (P + Q - PQ)^2 = (P^2 + PQ - P^2Q) + (QP + Q^2 - QPQ) - (PQP + PQQ - PQPQ) \\ &= P + QP + Q - QPQ - PQP - PQ + (PQP)Q = P + Q - QPQ - PQ + QPQ \\ &= P + Q - PQ.\end{aligned}$$

Moreover, $(P \oplus Q)P = P$, and so $(P \oplus Q)u = u$ for all $u \in \mathcal{R}(P)$. Similarly, $(P \oplus Q)Q = Q$ and hence $(P \oplus Q)v = v$ for all $v \in \mathcal{R}(Q)$. If $w \in \mathcal{R}(P) + \mathcal{R}(Q)$ then $w = u + v$ where $u \in \mathcal{R}(P)$ and $v \in \mathcal{R}(Q)$. From above, both $u$ and $v$ are fixed by $P \oplus Q$. □

**Remark 1.** *The operator $P \otimes Q$ is called the* Boolean sum *of the linear operators.*

We finish our discussion with the following result which shows that if we are willing to consider only a finite-dimensional subspace $U$, a projection onto $U$ always exists, and the norm is not too big.

**Theorem 6.** *Let $U$ be an $n$-dimensional subspace of a Banach space $E$. There exists a projection $P : E \to U$ with norm at most $n$.*

*Proof.* Before proving the main result we need the following tool from functional analysis (it is a simplified version of Auerbach's theorem):

*If $U$ is an $n$-dimensional normed space, then there exist vectors $u_1, \cdots, u_n$ and functionals $\ell_1, \cdots, \ell_n$ such that $\|u_i\| = 1$, $\|\ell_i\|_1, 1, \cdots, n$, and $\ell_j(u_i) = \delta_{i,j}$.*

We sketch the proof briefly. Let $v_1, \cdots, v_n$ be a basis of $U$. Define $M : (U^*)^{\otimes n} \to \mathbb{R}$, by $M(\nu_1, \cdots, \nu_n) = |\det(\nu_j(v_i))|$. Since $U^*$ is finite-dimensional, $M$ attains its maximum $m_*$ in the unit ball $B_{U^*} \times \cdots B_{U^*}$ at some point $(\ell_1, \cdots, \ell_n)$. Note that $m_* > 0$ since otherwise $A_{i,j} = \nu_j(v_i)$ would have a left nullvector for any choice of $\nu$'s. This in turn would imply that there is a $v$ such that $\ell(v)$ for all $\ell \in U^*$ (just choose $\nu_1, \cdots, \nu_n$ to be a basis of $U^*$ for example).

Now set $A_{j,i} = \ell_i(v_j)$ and let $C = A^{-1}$. We set $u_j = \sum_{k=1}^n C_{j,k} v_k$. It is easy to see that $\ell_i(u_j) = \delta_{i,j}$, $1 \le i, j \le n$. To show the bound on $\|u_j\|$, note that for any $\psi \in E^*$,

$$\psi(u_i) = \sum_{j=1}^n C_{i,j} \psi(v_j).$$

Next, we observe that $C_{i,j}\psi(v_j)$ (as a vector) is a solution to the equation $Ax = \psi(\vec{v})$. By Cramer's rule,

$$\sum_{j=1}^n C_{i,j} \psi(u_j) = \frac{\det A_i}{\det(A)},$$

where $A_i$ is the matrix formed by replacing the $i$th column with $\psi(u_j)$. In particular,

$$|\psi(u_i)| = \left| \frac{M(\ell_1, \cdots, \ell_{i-1}, \psi, \ell_{i+1}, \cdots, \ell_n)}{M(\ell_1, \cdots, \ell_n)} \right| \le \|\psi\|.$$

Thus, since $\psi$ was arbitrary, $\|u_i\| \le 1$.
We now turn to the proof of the theorem. From Auerbach's theorem, we can find $u_1, \cdots, u_n \in U$ and $\ell_1, \cdots, \ell_n \in U^*$. By Hahn-Banach, the latter can be extended to act on all of $E$, while not increasing the

norm. Then, set $P(v) = \sum_i u_i \ell_i(v)$. Clearly, $\mathcal{R}(P) = U$ and $P^2 = P$. Finally,

$$\|Pv\| = \|\sum_i u_i \ell_i(v)\| \leq \sum_i \|u_i\|\|\ell_i\|\|v\| \leq n\|v\|.$$

□

## *Further reading*

There are many good functional analysis textbooks which cover this. See the notes from R. Vershynin, for example. For projections, see *A Comprehensive Course in Analysis, Volume 5* by B. Simon. For a more applied approach, see *A Course in Approximation Theory* by W. Cheney and W. Light.

## *Additional Exercises*

**Exercise 4.** *Let $\{v_j\}_{j=1}^{\infty}$ be an orthonormal basis of a Hilbert space $H$. Let $A$ be a bounded operator from $H$ to itself with bounded inverse. Consider the sequence of vectors $w_j := Av_j$. Is $\{w_j\}_1^{\infty}$ an orthonormal basis? A Schauder basis? If so, can you bound its basis constant?*

**Exercise 5.** *Let $\psi : \mathbb{R} \to \mathbb{R}$ be the function defined by*

$$\psi(x) = \begin{cases} 0, & x \notin (0,1), \\ 2x, & x \in [0, 1/2], \\ 2(1-x), & x \in [1/2, 1]. \end{cases}$$

*For $i = 0, \cdots, \infty$ and $k = 0, \cdots, 2^i - 1$ define $\psi_{i,k}$ by*

$$\psi_{i,k}(x) = \psi(2^i x + k).$$

*Consider the collection of functions $\mathcal{X} = \{1, x, \psi_{0,0}, \psi_{1,0}, \psi_{1,1}, \psi_{2,0}, \cdots\}$. Sketch the first five functions of $\mathcal{X}$ on $[0,1]$. Show that $\mathcal{X}$ is a Schauder basis of $C([0,1])$.*

**Exercise 6.** *Let $H$ be a Hilbert space and let $P, Q : H \to H$ be two projections. In the following, set $C = i[P,Q] = i(PQ - QP)$. We also assume that $P$ and $Q$ are* orthogonal, *i.e. $P^* = P$ and $Q^* = Q$.*

a) *Find explicit examples of $P$ and $Q$ which do not commute. Make sure to give clear definitions of both them and $H$. Sketch a geometric characterization of when this happens.*

b) *Show that $\langle v, Pv\rangle \geq 0$ and $\langle v, Pv\rangle \leq 1$ for all $v \in H$.*

c) *For any $v \in H$, set $\bar{P}(v) = \langle v, Pv\rangle$, $\bar{Q}(v) = \langle v, Qv\rangle$, $\sigma_P^2(v) = \langle v, (P - \bar{P}\, I)^2 v\rangle$ and $\sigma_Q^2(v) = \langle v, (Q - \bar{Q}\, I)^2 v\rangle$. Show that $\sigma_P, \sigma_Q \leq \|v\|^2/2$.*

*d) If $C$ is the commutator of $P$ and $Q$, show that $\frac{1}{4}|\langle v, Cv\rangle|^2 \leq \sigma_P^2(v)\sigma_Q^2(v)$. Argue that $\|C\| \leq 1/2$. Hint: use Cauchy-Schwarz. For the last part, you can use without proof that $\|C\| = \sup_{\|v\|\leq 1} |\langle v, Cv\rangle|$ (this follows from the fact that $C$ is Hermitian).*