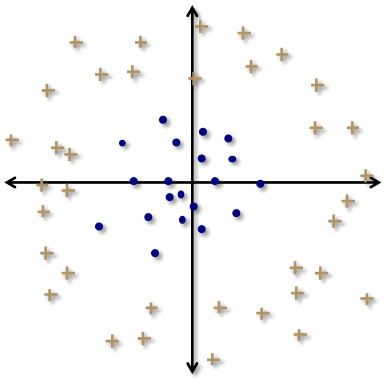
Kernels

We have learned several ways to find linear classifiers (LDA, logistic regression, perceptron, maximum margin) for feature vectors $\boldsymbol{x} \in \mathbb{R}^d$ and binary class labels $y \in \{-1, 1\}$. In this section, we will see how we can create **nonlinear** classifiers by first transforming the data with a nonlinear *feature map*,

$$\boldsymbol{\Phi} : \mathbb{R}^d \to \mathbb{R}^p,$$

and then training a linear classifier on $\boldsymbol{\Phi}(\boldsymbol{x}_1), \boldsymbol{\Phi}(\boldsymbol{x}_2), \ldots, \boldsymbol{\Phi}(\boldsymbol{x}_n)$.

A quick example shows how this works. Consider the following data with $\boldsymbol{x} \in \mathbb{R}^2$:



In this case, every linear classifier is pretty awful. However, we can see that the simple rule

$$x(1)^2 + x(2)^2 - 1 \lesseqgtr 0, \tag{1}$$

will separate the data. This rule is nonlinear in $\mathbb{R}^2$, but we can find a $\boldsymbol{\Phi}(\cdot)$ that makes it linear in a higher dimensional space. In particular,

consider the following map in to $\mathbb{R}^6$:

$$\boldsymbol{\Phi}(\boldsymbol{x}) = \begin{bmatrix} 1 \\ x(1) \\ x(2) \\ x(1)^2 \\ x(2)^2 \\ \sqrt{2}x(1)x(2) \end{bmatrix}.$$

The the linear rule on the mapped data in $\mathbb{R}^6$,

$$\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\Phi}(\boldsymbol{x}) \lessgtr 0, \quad \boldsymbol{\theta} = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

corresponds to the nonlinear rule (1) in $\mathbb{R}^2$.

The introduction of the mapping $\boldsymbol{\Phi}(\cdot)$ greatly enriches the set of choices we have for a classifier. This enrichment comes at a cost, however, as now the algorithms to fit the classifier are working in $\mathbb{R}^p$ instead of $\mathbb{R}^d$, and typically $p \gg n$. For example, the quadratic mapping above had $d = 2$ and $p = 6$; for general d, we will have $p = (d^2 + 3d + 2)/2$, and so p is proportional to d^2. When d is on the order of 10^3, this can be a problem.

Fortunately, there is a trick that allows us to fit linear classifiers in this high dimensional space while still working in $\mathbb{R}^d$, and in fact, we will not even have to compute the $\boldsymbol{\Phi}(\boldsymbol{x}_i)$ on the training data. Suppose that we have a classification algorithm that works only with distances and inner products between data points. Then all we need

is a function $k(\cdot, \cdot)$ that computes[1]:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{\Phi}(\boldsymbol{x}), \boldsymbol{\Phi}(\boldsymbol{x}') \rangle = \boldsymbol{\Phi}(\boldsymbol{x}')^{\mathrm{T}} \boldsymbol{\Phi}(\boldsymbol{x}).$$

Note that this would also give us a way to compute distances, as:

$$k(\boldsymbol{x} - \boldsymbol{x}', \boldsymbol{x} - \boldsymbol{x}') = \|\boldsymbol{\Phi}(\boldsymbol{x}) - \boldsymbol{\Phi}(\boldsymbol{x}')\|_2^2.$$

An example of a classification algorithm that only needs distances between data points is k-nearest neighbor. We will also see how the optimization program for finding a maximum margin hyperplane can be recast to only depend on inner products between data points; this will eventually lead to what is known as a *support vector machine* (SVM)[2]

We will start with a couple of examples of $k(\cdot, \cdot)$ that are easy to compute and correspond to inner products between feature maps in high dimensions.

Quadratic Kernels

Consider the simple act of taking an inner product and squaring it:

$$k(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u}^{\mathrm{T}} \boldsymbol{v})^2.$$

We might ask: does $k(\boldsymbol{u}, \boldsymbol{v}) = \langle \boldsymbol{\Phi}(\boldsymbol{u}), \boldsymbol{\Phi}(\boldsymbol{v}) \rangle$ for some nonlinear map $\boldsymbol{\Phi}(\cdot)$? If so, what is it and what is the corresponding p?

[1] Here and below, we will use $\langle \cdot, \cdot \rangle$ to denote an inner product

[2] SVM = maximum margin + a kernel.

Expanding the expression above provides the answer:

$$
\begin{aligned}
(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v})^2 = \left(\sum_{i=1}^{d} u(i)v(i)\right)^2 &= \left(\sum_{i=1}^{d} u(i)v(i)\right)\left(\sum_{j=1}^{d} u(j)v(j)\right) \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d} u(i)v(i)u(j)v(j) \\
&= \sum_{i=1}^{d} u(i)^2 v(i)^2 + \sum_{i\neq j} u(i)u(j)\cdot v(i)v(j) \\
&= \langle \boldsymbol{\Phi}(\boldsymbol{u}), \boldsymbol{\Phi}(\boldsymbol{v}) \rangle
\end{aligned}
$$

where

$$
\boldsymbol{\Phi}(\boldsymbol{u}) = \begin{bmatrix}
u(1)^2 \\
\vdots \\
u(d)^2 \\
\sqrt{2}u(1)u(2) \\
\vdots \\
\sqrt{2}u(1)u(d) \\
\sqrt{2}u(2)u(3) \\
\vdots \\
\vdots \\
\sqrt{2}u(d-1)u(d)
\end{bmatrix}
$$

In this case $p = d + d(d-1)/2$.

A related mapping is the *nonhomogenous quadratic kernel*

$$k(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v} + 1)^2.$$

In this case, calculations similar to the one above show that the corresponding $\boldsymbol{\Phi}(\boldsymbol{u})$ will replicate $\boldsymbol{u}$ and then include all its quadratic

terms:

$$k(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v} + 1)^2 = (\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v})^2 + 2\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v} + 1 = \langle \boldsymbol{\Phi}(\boldsymbol{u}), \boldsymbol{\Phi}(\boldsymbol{v}) \rangle,$$

where

$$\boldsymbol{\Phi}(\boldsymbol{u}) = \begin{bmatrix} 1 \\ \sqrt{2}u(1) \\ \vdots \\ \sqrt{2}u(d) \\ u(1)^2 \\ \vdots \\ u(d)^2 \\ \sqrt{2}u(1)u(2) \\ \vdots \\ \vdots \\ \sqrt{2}u(d-1)u(d). \end{bmatrix}$$

In this case, $p = 1 + d + d + d(d-1)/2 = (d^2 + 3d + 2)/2$.

Polynomial Kernels

We can just as well consider arbitrary orders:

$$k(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v})^m = \left(\sum_{i=1}^{d} u(i)v(i) \right)^m.$$

We can expand the term on the right above using the *multinomial theorem*[3]. For $x_1, \ldots, x_d \in \mathbb{R}$, we have for any integer m,

$$(x_1 + x_2 + \cdots + x_d)^m = \sum_{(j_1, \ldots, j_d) \in \mathcal{P}} \binom{m}{j_1, \ldots, j_d} x_1^{j_1} x_2^{j_2} \cdots x_d^{j_d},$$

[3]The Wikipedia article on this is decent, https://en.wikipedia.org/wiki/Multinomial_theorem.

where $\mathcal{P}$ is the set of multi-indexes that sum to m,

$$\mathcal{P} = \left\{ j_1, \ldots, j_d \mid k_i \geq 0, \ \sum_{i=1}^{d} j_i = m \right\},$$

and

$$\binom{m}{j_1, \ldots, j_d} = \frac{m!}{j_1! j_2! \cdots j_d!}.$$

Thus

$$k(\boldsymbol{u}, \boldsymbol{v}) = \sum_{(j_1, \ldots, j_d) \in \mathcal{P}} \binom{m}{j_1, \ldots, j_d} u(1)^{j_1} v(1)^{j_1} \cdots u(d)^{j_d} v(d)^{j_d},$$

and we can take

$$\boldsymbol{\Phi}(\boldsymbol{u}) = \begin{bmatrix} \vdots \\ \sqrt{\binom{m}{j_1, \ldots, j_d}} u(1)^{j_1} \cdots u(d)^{j_d} \\ \vdots \end{bmatrix}$$

The dimension of $\boldsymbol{\Phi}(\boldsymbol{u})$ is now much, much larger than d. It is something like $\sim d^m$.

Inner Product Kernels

We are now ready to ask a general question:

What kinds of functions $k(\boldsymbol{u}, \boldsymbol{v})$ correspond to inner products between nonlinear mappings of $\boldsymbol{u}$ and $\boldsymbol{v}$?

We can answer this question in a very concrete mathematical way. But first, we should carefully define what we are looking for. We want to determine whether or not a particular $k(\boldsymbol{u}, \boldsymbol{v})$ corresponds to an inner product in a higher dimensional space. This space might be infinite dimensional, and hence more abstract that our standard vector space $\mathbb{R}^p$.

A *Hilbert space* is collection of objects (e.g. functions of continuous variables, or infinite sequences of numbers) that has some notion of inner product (and hence norm and distance) associated with it. These spaces are very similar to $\mathbb{R}^p$ in that we can do thing like project onto subspaces (i.e. solve least-squares problems), diagonalize linear operators using something like the SVD, and define linear functionals using inner products. The main difference is that they can be infinite dimensional, and so there are always technical convergence concerns that have to be handled in a principled manner.

We say that $\mathcal{H}$ is a (real-valued) **Hilbert space** if

1. $\mathcal{H}$ is a linear vector space[4]. This essentially means that $\mathcal{H}$ is closed under linear combinations: if $\boldsymbol{f}, \boldsymbol{g} \in \mathcal{H}$, then $\alpha \boldsymbol{f} + \beta \boldsymbol{g} \in \mathcal{H}$ for all $\alpha, \beta \in \mathbb{R}$.
2. There is an associated inner product $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ that

[4] See https://en.wikipedia.org/wiki/Vector_space for a complete definition.

obeys the three rules:

a) Symmetry: $\langle \boldsymbol{f}, \boldsymbol{g} \rangle = \langle \boldsymbol{g}, \boldsymbol{f} \rangle$ for all $\boldsymbol{f}, \boldsymbol{g} \in \mathcal{H}$.

b) Linearity: $\langle \alpha_1 \boldsymbol{f}_1 + \alpha_2 \boldsymbol{f}_2, \boldsymbol{g} \rangle = \alpha_1 \langle \boldsymbol{f}_1, \boldsymbol{g} \rangle + \alpha_2 \langle \boldsymbol{f}_2, \boldsymbol{g} \rangle$ for all $\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{g} \in \mathcal{H}$ and $\alpha_1, \alpha_2 \in \mathbb{R}$.

c) $\langle \boldsymbol{f}, \boldsymbol{f} \rangle \geq 0$ for all $\boldsymbol{f} \in \mathcal{H}$, and $\langle \boldsymbol{f}, \boldsymbol{f} \rangle = 0 \Leftrightarrow \boldsymbol{f} = \boldsymbol{0}$.

Having a valid inner product immediately imbues $\mathcal{H}$ with two important types of structure:

- a **distance** defined through the induced norm:

$$\|\boldsymbol{f} - \boldsymbol{g}\| = \sqrt{\langle \boldsymbol{f} - \boldsymbol{g}, \boldsymbol{f} - \boldsymbol{g} \rangle},$$

- the **angle** θ between two elements of $\mathcal{H}$:

$$\cos\theta = \frac{\langle \boldsymbol{f}, \boldsymbol{g} \rangle}{\|\boldsymbol{f}\| \, \|\boldsymbol{g}\|}$$

These definitions are exactly the same as they are in Euclidean space, and that is the point: if you have an inner product, all your favorite geometrical results generalize. Here are three more examples of things that are very familiar in $\mathbb{R}^p$, but are also true in general Hilbert spaces:

Pythagorean theorem If $\langle \boldsymbol{f}, \boldsymbol{g} \rangle = 0$, then

$$\|\boldsymbol{f} + \boldsymbol{g}\|^2 = \|\boldsymbol{f}\|^2 + \|\boldsymbol{g}\|^2$$

Cauchy-Schwarz For any $\boldsymbol{f}, \boldsymbol{g} \in \mathcal{H}$,

$$|\langle \boldsymbol{f}, \boldsymbol{g} \rangle| \leq \|\boldsymbol{f}\| \cdot \|\boldsymbol{g}\|$$

with equality holding if and only if $\boldsymbol{f}$ and $\boldsymbol{g}$ are colinear: $\boldsymbol{f} = \alpha \boldsymbol{g}$ for some $\alpha \in \mathbb{R}$.

Triangle inequality For general $\boldsymbol{f}, \boldsymbol{g} \in \mathcal{H}$,

$$\|\boldsymbol{f} + \boldsymbol{g}\| \leq \|\boldsymbol{f}\| + \|\boldsymbol{g}\|.$$

If $\boldsymbol{f} = \alpha \boldsymbol{g}$ for $\alpha > 0$, then

$$\|\boldsymbol{f} + \boldsymbol{g}\| = \|\boldsymbol{f}\| + \|\boldsymbol{g}\|.$$

Basically what all of this means is that we have all of the mathematical structure in place to make the problem of fitting a separating hyperplane well-posed.

We now return to our central question of what kinds of kernels correspond to mappings into inner product (Hilbert) spaces. The following definition is central to the answer:

Definition: We say that $k(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a **symmetric positive semi-definite kernel** (PSD kernel) if the following hold:

1. $k(\boldsymbol{u}, \boldsymbol{v}) = k(\boldsymbol{v}, \boldsymbol{u})$ for all $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$,
2. for all $n \in \mathbb{N}$ and all $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, the matrix $\boldsymbol{K}$ with entries
$$K(i, j) = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$$
is positive semidefinite, meaning $\boldsymbol{z}^{\mathrm{T}} \boldsymbol{K} \boldsymbol{z} \geq 0$ for all $\boldsymbol{z} \in \mathbb{R}^n$.

Given a kernel function $k(\boldsymbol{u}, \boldsymbol{v})$, there exists a Hilbert space $\mathcal{H}$ with associated inner product $\langle \cdot, \cdot \rangle$ and mapping $\boldsymbol{\Phi} : \mathbb{R}^d \to \mathcal{H}$ with

$$\langle \boldsymbol{\Phi}(\boldsymbol{u}), \boldsymbol{\Phi}(\boldsymbol{v}) \rangle = k(\boldsymbol{u}, \boldsymbol{v}),$$

if and only if k is a PSD kernel.

To see this, first notice that when we fix one of the inputs to the kernel k, it becomes a function on $\mathbb{R}^d$. By convention, we will fix the second argument, writing $f_{\boldsymbol{v}}(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{v})$. As we vary $\boldsymbol{v}$, this function changes.

We will take $\mathcal{H}$ as the space of *functions* with domain $\mathbb{R}^d$ that can be written as a finite linear combination of "columns" of the kernel $k(\boldsymbol{u}, \boldsymbol{v})$. More precisely, all $\boldsymbol{f} \in \mathcal{H}$ are functions $f : \mathbb{R}^d \to \mathbb{R}$ that can be written as

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i k(\boldsymbol{x}, \boldsymbol{v}_i) \tag{2}$$

for some natural number N, some sequence of vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N \in \mathbb{R}^d$, and some sequence of scalars $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$. A simpler way to say this is that

$$\mathcal{H} = \text{Span}\left\{k(\cdot, \boldsymbol{v}),\ \boldsymbol{v} \in \mathbb{R}^d\right\}.$$

Now for two such functions $\boldsymbol{f}, \boldsymbol{g} \in \mathcal{H}$,

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i k(\boldsymbol{x}, \boldsymbol{v}_i), \quad g(\boldsymbol{x}) = \sum_{j=1}^{M} \beta_j k(\boldsymbol{x}, \boldsymbol{w}_j),$$

we consider the following candidate for the inner product:

$$\langle \boldsymbol{f}, \boldsymbol{g} \rangle = \sum_{i=1}^{N} \sum_{j=1}^{M} \alpha_i \beta_j\, k(\boldsymbol{v}_i, \boldsymbol{w}_j). \tag{3}$$

If we take $\boldsymbol{\Phi} : \mathbb{R}^d \to \mathcal{H}$ as

$$\boldsymbol{\Phi}(\boldsymbol{u}) = k(\cdot, \boldsymbol{u}),$$

then

$$\langle \boldsymbol{\Phi}(\boldsymbol{u}), \boldsymbol{\Phi}(\boldsymbol{v}) \rangle = k(\boldsymbol{u}, \boldsymbol{v}),$$

since for $k(\cdot, \boldsymbol{u})$ we can take $N = 1, \alpha_1 = 1$ in the representation (2), and similarly for $k(\cdot, \boldsymbol{v})$. It still remains to show that this is indeed a valid inner product on $\mathcal{H}$.

Your first concern might be that the $N, \{\boldsymbol{v}_i\}, \{\alpha_i\}, M, \{\boldsymbol{w}_i\}, \{\beta_i\}$ are not unique, and indeed there could in general be many different choices that lead to the same $\boldsymbol{f}$ and $\boldsymbol{g}$. But over all such choices, the inner product above will evaluate to the same thing, as

$$\langle \boldsymbol{f}, \boldsymbol{g} \rangle = \sum_{i=1}^{N} \sum_{j=1}^{M} \alpha_i \beta_j \, k(\boldsymbol{v}_i, \boldsymbol{w}_j) = \sum_{j=1}^{M} \beta_j f(\boldsymbol{w}_j) = \sum_{i=1}^{N} \alpha_i g(\boldsymbol{v}_i). \quad (4)$$

Now we verify that (3) meets the three criteria for an inner product when $k(\boldsymbol{u}, \boldsymbol{v})$ is symmetric positive semi-definite kernel.

Symmetry follows immediately from the symmetry of $k(\boldsymbol{u}, \boldsymbol{v})$. The linearity property is also straightforward to verify given the form of (3). For the third property, consider that for a fixed $\boldsymbol{f} \in \mathcal{H}$,

$$\langle \boldsymbol{f}, \boldsymbol{f} \rangle = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j k(\boldsymbol{v}_i, \boldsymbol{v}_j) = \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{K} \boldsymbol{\alpha},$$

where $\boldsymbol{\alpha} \in \mathbb{R}^N$ contains the α_i, and $\boldsymbol{K}$ is the $N \times N$ matrix with entries $K(i,j) = k(\boldsymbol{v}_i, \boldsymbol{v}_j)$. Since $k(\cdot, \cdot)$ is a symmetric positive semidefinite kernel, we know that $\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{K} \boldsymbol{\alpha} \geq 0$. It is also clear that $\boldsymbol{f} = \boldsymbol{0} \Rightarrow \langle \boldsymbol{f}, \boldsymbol{f} \rangle = 0$.

It remains to show that $\langle \boldsymbol{f}, \boldsymbol{f} \rangle = 0 \Rightarrow \boldsymbol{f} = \boldsymbol{0}$. Suppose that indeed $\langle \boldsymbol{f}, \boldsymbol{f} \rangle = 0$, then consider $\boldsymbol{f}$ evaluated at an arbitrary $\boldsymbol{x} \in \mathbb{R}^d$:

$$\begin{aligned} |f(\boldsymbol{x})| &= \left| \sum_{i=1}^{N} \alpha_i k(\boldsymbol{x}, \boldsymbol{v}_i) \right| \\ &= |\langle \boldsymbol{f}, k(\cdot, \boldsymbol{x}) \rangle| \, , \end{aligned}$$

where the second equality follows from the fact that $k(\cdot, \boldsymbol{x}) \in \mathcal{H}$ and (4). Using properties of $\langle \cdot, \cdot \rangle$ we have established so far and the variation of the Cauchy-Schwarz inequality in Lemma 1 in the Technical Details section below, we have

$$|f(\boldsymbol{x})|^2 \leq \langle \boldsymbol{f}, \boldsymbol{f} \rangle \langle k(\cdot, \boldsymbol{x}), k(\cdot, \boldsymbol{x}) \rangle = 0.$$

Thus $f(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$, and so $\boldsymbol{f} = \boldsymbol{0}$.

Other important examples of kernels

So far, we have only explicitly talked about polynomial kernels, where the mapping $\boldsymbol{\Phi}(\cdot)$ leads us to another finite (albeit high) dimensional space. There are plenty of examples of kernels used in practice that lead to infinite dimensional Hilbert spaces.

Perhaps the most prominent example is the Gaussian kernel:

$$k(\boldsymbol{u}, \boldsymbol{v}) = \exp\left(-\frac{\|\boldsymbol{u} - \boldsymbol{v}\|_2^2}{2\sigma^2}\right),$$

where σ^2 is a width parameter that we choose. The Hilbert space associated with this kernel is the space of all functionals on $\mathbb{R}^d$ that can be approximated arbitrarily well by a finite number of multidimensional "Gaussian bump" functionals at different locations. This Hilbert space is infinite dimensional ... it can be shown[5] that for any set of N distinct points $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N$, the matrix

$$K(i, j) = \exp\left(-\frac{\|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2^2}{2\sigma^2}\right),$$

[5] See C. A. Micchelli, *Algebraic aspects of interpolation*, 1986.

has a rank of N — this means that the functions $k(\cdot, \boldsymbol{v}_1), \ldots, k(\cdot, \boldsymbol{v}_N)$ are linearly independent. Since N can be made arbitrarily large, there are a set of linearly independent vectors in $\mathcal{H}$ with unbounded size, meaning the dimension of $\mathcal{H}$ is infinite.

The Gaussian kernel is one example of a **radial basis function**; it is a function of two d-dimensional vectors that depends only on the distance between them. We might consider similar kernels

$$k(\boldsymbol{u}, \boldsymbol{v}) = r(\|\boldsymbol{u} - \boldsymbol{v}\|),$$

where we could use different functions $r : \mathbb{R} \to \mathbb{R}$ and notions of distance $\|\cdot\|$. Other popular choices for $r(\cdot)$ include B-splines or Laplacian functions. The dimension of the resulting spaces depends of course on properties of the distance measure and r.

Technical Details

This is a variation on the Cauchy-Schwarz inequality:

Lemma 1 *Suppose that $Q(\cdot,\cdot)$ is a symmetric bilinear function on a real-valued vector space $\mathcal{H}$ obeying*

1. *$Q(\boldsymbol{f},\boldsymbol{g}) = Q(\boldsymbol{g},\boldsymbol{f})$ for all $\boldsymbol{f},\boldsymbol{g} \in \mathcal{H}$,*
2. *$Q(\alpha_1\boldsymbol{f} + \alpha_2\boldsymbol{h},\boldsymbol{g}) = \alpha_1 Q(\boldsymbol{f},\boldsymbol{g}) + \alpha_2 Q(\boldsymbol{h},\boldsymbol{g})$, and*
3. *$Q(\boldsymbol{f},\boldsymbol{f}) \geq 0$,*

for all $\boldsymbol{f},\boldsymbol{g},\boldsymbol{h} \in \mathcal{H}$ and $\alpha_1,\alpha_2 \in \mathbb{R}$. Then

$$|Q(\boldsymbol{f},\boldsymbol{g})|^2 \;\leq\; Q(\boldsymbol{f},\boldsymbol{f})\,Q(\boldsymbol{g},\boldsymbol{g}).$$

Proof For any $\boldsymbol{f},\boldsymbol{g} \in \mathcal{H}$, the properties of $Q(\cdot,\cdot)$ tell us that $Q(\boldsymbol{f}+\boldsymbol{g},\boldsymbol{f}+\boldsymbol{g}) \geq 0$, and

$$Q(\boldsymbol{f}+\boldsymbol{g},\boldsymbol{f}+\boldsymbol{g}) = Q(\boldsymbol{f},\boldsymbol{f}) + 2\,Q(\boldsymbol{f},\boldsymbol{g}) + Q(\boldsymbol{g},\boldsymbol{g}).$$

Since the expression above holds for *all* $\boldsymbol{f},\boldsymbol{g} \in \mathcal{H}$, and $\mathcal{H}$ is a linear vector space, it must hold for $\alpha\boldsymbol{f}$ and $\beta\boldsymbol{g}$ for all $\alpha,\beta \in \mathbb{R}$, meaning

$$\alpha^2 Q(\boldsymbol{f},\boldsymbol{f}) + 2\alpha\beta\,Q(\boldsymbol{f},\boldsymbol{g}) + \beta^2\,Q(\boldsymbol{g},\boldsymbol{g}) \geq 0, \quad \text{for all } \alpha,\beta \in \mathbb{R}.$$

This is the same as saying that

$$\begin{bmatrix}\alpha & \beta\end{bmatrix}\begin{bmatrix}Q(\boldsymbol{f},\boldsymbol{f}) & Q(\boldsymbol{f},\boldsymbol{g}) \\ Q(\boldsymbol{f},\boldsymbol{g}) & Q(\boldsymbol{g},\boldsymbol{g})\end{bmatrix}\begin{bmatrix}\alpha \\ \beta\end{bmatrix} \geq 0,$$

which means that the 2×2 matrix above is symmetric positive semi-definite for all $\boldsymbol{f},\boldsymbol{g}$. This means that its determinant is non-negative, and so

$$Q(\boldsymbol{f},\boldsymbol{f})Q(\boldsymbol{g},\boldsymbol{g}) - |Q(\boldsymbol{f},\boldsymbol{g})|^2 \geq 0.$$

■