Lecturer: Prof. Homrighausen Scribe: Chuck Vollmer

1 Parseval's Theorem

If a function has a Fourier series given by

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos(nx) + \sum_{n=1}^{\infty} b_n \sin(nx) \tag{1}$$

then Bessel's inequality becomes an equality know as *Parseval's theorem*, from (1):

$$\begin{aligned}[f(x)]^2 = \frac{1}{4}a_0^2 + a_0 \sum_{n=1}^{\infty} a_n \cos(nx) + b_n \sin(nx) \\ + \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} a_n a_m \cos(nx)\cos(mx) + a_n b_m \cos(nx)\sin(mx) + a_m b_n \sin(nx)\cos(mx) + b_n b_m \sin(nx)\sin(mx)\end{aligned} \tag{2}$$

Integrating

$$\begin{aligned}\int_{-\pi}^{\pi}[f(x)]^2 dx &= \frac{1}{4}a_0^2\int_{-\pi}^{\pi} dx + a_0\int_{-\pi}^{\pi}\sum_{n=1}^{\infty} a_n\cos(nx) + b_n\sin(nx)dx \\ &+ \int_{-\pi}^{\pi}\sum_{n=1}^{\infty}\sum_{m=1}^{\infty} a_n a_m \cos(nx)\cos(mx) + a_n b_m \cos(nx)\sin(mx) + a_m b_n \sin(nx)\cos(mx) + b_n b_m \sin(nx)\sin(mx)dx \\ &= \frac{1}{4}a_0^2(2\pi) + 0 + \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} a_n a_m \pi\delta_{nm} + 0 + 0 + b_n b_m \pi\delta_{nm}\end{aligned} \tag{3}$$

So

$$\frac{1}{\pi}\int_{-\pi}^{\pi}[f(x)]^2 dx = \frac{1}{2}a_0^2 + \sum_{n=1}^{\infty}(a_n^2 + b_n^2) \tag{4}$$

Example 1.1. *Thus, under our conditions the problem is well-understood if the error of the representation is measured using the Euclidean or ℓ_2 norm. Since the ℓ_2 norm is preserved under rotations, by Parseval's theorem, we have*

$$\|f - \hat{f}\|_2^2 = \sum_i (f_i - \sum_k \beta_k \phi_k)^2 = \sum_i (< f, \phi_k > -\beta_k)^2 \tag{5}$$

and it is clear that the solution under this error measure is to retain the largest K inner products $< f, \phi_k >$, which are also the coefficients of the basis expansion of f.

2 Vandermonde Matrix

A *Vandermonde Matrix* - of order n- is of the form

$$\begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{pmatrix} \tag{6}$$

A Vandermonde Matrix is a type of matrix that arises in the polynomial least squares fitting, Lagrange interpolating polynomials (Hoffman and Kunze p. 114), and the reconstruction of a statistical distribution from the distribution's moments (von Mises 1964; Press et al. 1992, p. 83).

Notes:

- Some authors define the transpose of this matrix as the Vandermonde matrix.
- The solution of an $n \times n$ Vandermonde matrix equation requires $O(n^2)$ operations.
- The determinants of Vandermonde matrices have a particularly simple form.

3 Spectral (operator) Norm

The *operator norm* of a linear operator $\mathbf{T} : \mathbf{V} \to \mathbf{W}$ is the largest value by which $\mathbf{T}$ stretches an element of $\mathbf{V}$,

$$\|T\| = \sup_{\|v\|=1} \|T(v)\| \tag{7}$$

It is necessary for V and W to be normed vector spaces. The operator norm of a composition is controlled by the norms of the operators,

$$\|TS\| \leq \|T\|\|S\| \tag{8}$$

When $\mathbf{T}$ is given by a matrix, say $\mathbf{T}(v) = \mathbf{A}v$, then $||T||$ is the square root of the largest eigenvalue of the symmetric matrix $A^T A$, all of whose eigenvalues are nonnegative. So, if we have the following linear operator

Example 3.1. *So, if we have the following linear operator*

$$A = \begin{pmatrix} 2 & 0 & 0 \\ 3 & 0 & 2 \end{pmatrix} \tag{9}$$

Then, we find the following

$$A^T A = \begin{pmatrix} 13 & 0 & 6 \\ 0 & 0 & 0 \\ 6 & 0 & 4 \end{pmatrix} \tag{10}$$

which has eigenvalues $\{0, 1, 16\}$*, so* $\|A\| = 4$.

4 Kullback-Leibler Divergence

It is interesting to note a few things about this pseudo-metric: the Kullback-Leibler Discrepancy (KLD).

1. the KLD can be seen as a measure of statistical independence
2. a Pythagorean Theorem exists in which the KLD plays a role similar to that of squared distance in Euclidean Geometry

4.1 a measure of independence

This can easily be seen by measuring the KLD from the distribution of $P(Y)$ of Y to the product of its marginal distributions:

$$I(Y) \stackrel{\text{def}}{=} K[P(Y), \Pi_i P(Y_i)] \tag{11}$$

where the divergence K[Q,R] is defined as in the class notes:

$$K[Q, R] \stackrel{\text{def}}{=} \int_Y Q(y) \log \frac{Q(y)}{R(y)} dy \tag{12}$$

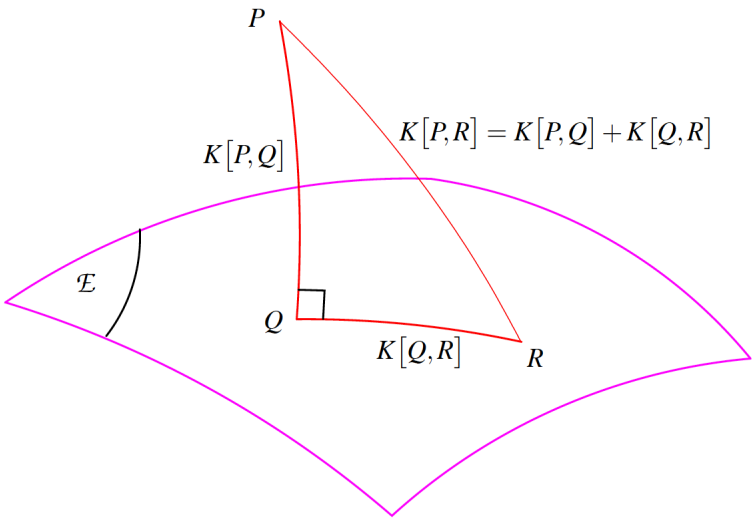
4.2 KLD Pythagorean Theorem

Now, another interesting feature is how the KLD plays a role similar to the squared distance in Euclidean Geometry.

Theorem 4.1. *For an exponential family, E, of distributions and for any P (not necessarily in E), there is a unique distribution $\mathcal{Q}$ of E such that the divergence from P to any distribution R of E decomposes as*

$$K[P,R] = K[P,Q] + K[Q,R]. \tag{13}$$

This Theorem is illustrated below:



This decomposition shows that Q is the closest distribution to P in E in the KLD sense since K[Q,R] reaches its minimum value of 0 for Q = R. Thus, Q can be called the "orthogonal projection" of P onto E and decomposition (13) should be read as a Pythagorean theorem in distribution space with the KL divergence being the analogue of a squared Euclidean distance.

And remember, since the divergence is not a true distance the Pythagorean theorem is actually a point-to-point relation.

5 KLD and the AIC

Akaike (1973) showed firstly that the maximized log-likelihood is biased upward as an estimator of the model selection criterion. Second, he showed that under certain conditions (these conditions are important, but quite technical), that this bias is approximately equal to K, the number of estimable parameters in the approximating model. Thus, an approximately unbiased estimator of the relative, expected K-L information is

$$\mathbb{P}_f[\log(g(Y;\theta))] \approx \log(g(Y;\hat{\theta}_{ML})) - K \tag{14}$$

which thus gives us that

$$-\mathbb{P}[\log(g(Y;\hat{\theta}_{ML}))] \approx -\log(g(Y;\hat{\theta}_{ML})) + K \tag{15}$$

Akaike (1973) then defined "*an* information criterion" (AIC) -in his words- "taking historical reasons into account" by multiplying by 2 to finally get

$$\mathbf{AIC} = -2\log(g(Y;\hat{\theta}_{ML})) + 2|\hat{\beta}_{ML}|$$