Plan

Orthogonalization (Gram-Schmidt Procedure)

\[ \downarrow \]

Orthogonal Polynomials: approximation of nonlinear functions \( f(x) \)

\[ \downarrow \]

Orthogonal Multinomials: approximation of nonlinear functions \( f(x, x_2, \ldots, x_k) \)

\[ \downarrow \]

Orthogonal Functionals: approximation of \( f[X(t)] \)

\[ \Rightarrow \] Discrete time approach \[ \Rightarrow \] "White Noise" \[ \Rightarrow m\text{-sequences} \]

\[ \Rightarrow \] Frequency domain approach \[ \Rightarrow \] sinusoids

We will see that

\[ \Rightarrow \text{describing a nonlinear system can be viewed as a regression problem} \]

\[ \Rightarrow \text{in contrast to the analysis of linear systems, there is no universal choice of "regressors" (in linear systems, the sinusoids)} \]

\[ \Rightarrow \text{the description of a nonlinear system depends on the choice of regressors ("context")} \]

\[ \Rightarrow \text{some choices are better than others} \]

- efficiency of analysis
- usefulness of description
Orthogonalization, approximating Gram-Schmidt procedure.

Working in a vector space $V$ (over $\mathbb{R}$) with an inner product $\langle \cdot, \cdot \rangle$.

Here, $V, \mathbf{v}, f \in V$, abstract - but we have in mind vectors that represent functions of single variable $f(x)$, functions of multiple discrete variables $f(x_1, \ldots, x_k)$, "functions" of a continuum of variables $f(x)$.

Say we want to approximate $f = \sum_{j=1}^{n} \alpha_j \mathbf{v}_j$, for some given library $\mathbf{v}_1, \ldots, \mathbf{v}_n$.

Then, we wish to find the $\alpha_j$ that minimize $R = \| f - \sum_{j=1}^{n} \alpha_j \mathbf{v}_j \|^2$.

where $\| u \|^2 = \langle u, u \rangle$.

We can write $f = f_{\text{app}} + f_{\text{err}}$, where $f_{\text{app}} = \sum_{j=1}^{n} \alpha_j \mathbf{v}_j$.

$R = \| f_{\text{err}} \|^2$.

We'd like to show that when $\| f_{\text{err}} \|^2$ is minimized, $f_{\text{err}}$ is orthogonal to $f_{\text{app}}$, i.e.,

$\langle f_{\text{app}}, f_{\text{err}} \rangle = 0$

i.e., we are projecting $f$ into the subspace spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_n$. 

\[ \mathbf{f} \]

\[ \mathbf{f}_{\text{app}} \]

\[ \mathbf{f}_{\text{err}} \]
Say, at the minimum, \( f_{\text{err}} = \sum_{j=1}^{n} \beta_j v_j + \varepsilon \),

where \( \varepsilon \perp v_j \), i.e., \( \langle \varepsilon, v_j \rangle = \langle v_j, \varepsilon \rangle = 0 \).

We have \( f = \sum_{j=1}^{n} (v_j + \beta_j) v_j + \varepsilon \)

\[ \underbrace{\sum_{j=1}^{n} \beta_j v_j}_{f_{\text{app}}} + \underbrace{\varepsilon}_{f_{\text{err}}} \]

But this can be reorganized into

\[ f = \sum_{j=1}^{n} (v_j + \beta_j) v_j + \varepsilon \]

\[ \underbrace{\sum_{j=1}^{n} \beta_j v_j}_{f_{\text{new-app}}} + \underbrace{\varepsilon}_{f_{\text{new-err}}} \]

We know that \( |f_{\text{new-err}}|^2 > |f_{\text{err}}|^2 \),

because we hypothesized that \( f_{\text{err}} \) was minimum.

\[ |f_{\text{err}}|^2 = \left| \sum_{j=1}^{n} \beta_j v_j + \varepsilon \right|^2 = \langle \sum_{j=1}^{n} \beta_j v_j + \varepsilon, \sum_{k=1}^{n} \beta_k v_k + \varepsilon \rangle \]

\[ = \sum_{j=1}^{n} \sum_{k=1}^{n} \beta_j \beta_k \langle v_j, v_k \rangle + \sum_{j=1}^{n} \sum_{k=1}^{n} \beta_j \varepsilon_k \langle v_j, \varepsilon \rangle + \sum_{k=1}^{n} \sum_{j=1}^{n} \varepsilon \beta_k \langle \varepsilon, v_k \rangle \]

\[ = \sum_{j=1}^{n} \beta_j \langle v_j, v_j \rangle + |\varepsilon|^2 = \sum_{j=1}^{n} \beta_j \langle v_j, v_j \rangle + |f_{\text{new-err}}|^2 \]

So \( \sum_{j=1}^{n} \beta_j v_j = 0 \implies f_{\text{err}} = f_{\text{new-err}} = \varepsilon. \)
In words, $f_{\text{ppr}}$ has no component in the subspace spanned by $v_1, \ldots, v_r$, since if it did, we could improve the approximation by adding this component into $f_{\text{ppr}}$.

Key ingredient: $u < \gamma$.

Geometrically, $f_{\text{ppr}}$ is the projection of $f$ onto the subspace spanned by the $v_j$. Projection is linear.

Formal solution: Let $A = \text{matrix of columns } v_1, \ldots, v_r$.

The projection is $P = A(ATA)^{-1}A^T$.

Verify $P^2 = A(ATA)^{-1}A^TA(ATA)^{-1}A^T$

$= A(ATA)^{-1}(ATA)(ATA)^{-1}A^T$

$= A(ATA)^{-1}A^T = P$

and $\text{span of } P$ include each column ($P = A\cdot x$)

Not a "useful" solution, in the sense that we'd need to calculate $(ATA)^{-1}$.

[When does this not exist?]

We could also try to minimize $R(x_1, \ldots, x_r) = \| f - \sum_{i=1}^{r} x_i v_i \|_2^2$

by $\frac{\partial R}{\partial x_i} = 0$, leading to a linear system of equations for the $v_i$'s.

Better solution is to replace $\{ v_1, \ldots, v_r \}$ by an orthonormal set

$\{ \psi_1, \ldots, \psi_r \}$ with the same span.
Then the approximation \( f_{\text{app}} = \sum_{j=1}^{\infty} \alpha_j v_j = \sum_{j=1}^{\infty} \alpha_j \langle \phi_h, v_j \rangle \),

and \( \langle f_{\text{app}}, \psi_h \rangle = \sum_{j=1}^{\infty} \alpha_j \langle \phi_h, v_j \rangle = \alpha_h \langle \phi_h, \psi_h \rangle \)

so \( \alpha_h = \frac{\langle f_{\text{app}}, \psi_h \rangle}{\langle \psi_h, \psi_h \rangle} = \frac{\langle f, \psi_h \rangle}{\langle \psi_h, \psi_h \rangle} \),

\( \text{so} \) \( \text{min}(f - f_{\text{app}})^2 \).

**Advantage 1:** No system of equations to solve.

**Advantage 2:** Straightforward to improve the approximation by adding new terms.

Say we have \( \sum_{j=1}^{\infty} \alpha_j v_j \) and we want to add \( v_{j+1} \).

\( f_{\text{app}} = \sum_{j=1}^{\infty} \alpha_j v_j \), no guarantee that \( \alpha_j = \alpha' \).

In fact, if \( \langle v_{j+1}, v_j \rangle \neq 0 \), then typically \( v_j \neq v' \).

But adding a new \( v_{j+1} \) doesn't resolve proof (1): \( \alpha_h = \frac{\langle f, \psi_h \rangle}{\langle \psi_h, \psi_h \rangle} \).

Think \( f \approx \sum_{j=1}^{\infty} \alpha_j v_j \) as a regression and

\( f \approx \sum_{j=1}^{\infty} \alpha_j \phi_j \) is a way to solve it.
How to create the \( \psi_j \)'s, such that the span of 
\[ \langle \psi_1, \cdots, \psi_n \rangle = \text{span of} \langle \psi_1, \cdots, \psi_n \rangle, \text{ and } \psi_j \text{'s orthogonal?} \]

"Gram Schmidt" procedure.

\[ \psi_1 = \psi_1 \]
\[ \psi_2 = \psi_2 - \text{projection of } \psi_2 \text{ onto space spanned by } \langle \psi_1 \rangle \]
\[ \psi_3 = \psi_3 - \text{projection of } \psi_3 \text{ onto space spanned by } \langle \psi_1, \psi_2 \rangle \]
\[ \psi_4 = \psi_4 - \text{projection of } \psi_4 \text{ onto space spanned by } \langle \psi_1, \psi_2, \psi_3 \rangle \]

At each stage, space spanned by \( \langle \psi_1, \cdots, \psi_n \rangle \) = space spanned by \( \langle \psi_1, \cdots, \psi_n \rangle \).

So calculation of the projection is easy, if you proceed iteratively

\[ \psi_1 = \psi_1 \]
\[ \psi_2 = \psi_2 - \frac{\langle \psi_2, \psi_1 \rangle \psi_1}{\langle \psi_1, \psi_1 \rangle} \]
\[ \psi_3 = \psi_3 - \frac{\langle \psi_3, \psi_2 \rangle \psi_2 - \langle \psi_3, \psi_1 \rangle \psi_1}{\langle \psi_2, \psi_2 \rangle} \]

etc.

Will fail if some \( \psi_n = 0 \), which will happen if \( \psi_n \) is an span of \( \psi_1, \cdots, \psi_{n-1} \).
An example: vectors are functions of a single variable, f(x).

Inner product:  \[ \langle f, g \rangle = \int f(x)g(x)W(x)dx \]

for some \( W(x) \geq 0 \) \( V \) is a set of functions for which

\[ \int f(x)^2 W(x)dx < \infty. \]

By minimizing \( f(x) \), we minimize

\[ \int |f - f_{\text{app}}|^2 W(x)dx \]

i.e., \( f_{\text{app}} \) is a good approximation where \( W \) is large.

Say \( V_0 = x^0, V_1 = x^1, V_2 = x^2, \) etc.

Say \( \int x^k W(x)dx = M_k \), \( M_0 = 1 \) \( W(x) \) is a probability distribution.

\[ \varphi_0 = x^0. \]

\[ \varphi_1 = x^1 - \frac{\langle x^1, \varphi_0 \rangle}{\langle \varphi_0, \varphi_0 \rangle} \quad \varphi_0 = x^1 - \frac{M_1}{M_0} \quad x^0 = x^1 - M_1. \]

\[ \langle x^1, \varphi_0 \rangle = \int x^1 \cdot x^0 W(x)dx = \int x^1 W(x)dx = M_1. \]
\[ \psi_2 = x^2 - \frac{\langle x^2, \psi_1 \rangle}{\langle \psi_1, \psi_1 \rangle} \psi_1 - \frac{\langle x^2, \psi_0 \rangle}{\langle \psi_0, \psi_0 \rangle} \psi_0 \]

\[ \langle x^2, \psi_1 \rangle = \int x^2 (x' - M_1) W(x') dx' = \int x^2 (x - M_1) W(x) dx = M_3 - M_1 M_2 \]

\[ \langle x^2, \psi_0 \rangle = \int x^2 W(x) dx = M_2. \]

\[ \langle \psi_1, \psi_1 \rangle = \int (x' - M_1)^2 W(x') dx' = \int (x^2 - 2M_1 x' + M_1^2) W(x) dx = M_2 - M_1^2. \]

\[ \psi_2 = x^2 - \frac{M_3 - M_1 M_2}{M_2 - M_1^2} (x' - M_1) - M_2. \]

\[ = x^2 - \left( \frac{M_3 - M_1 M_2}{M_2 - M_1^2} \right) x' + \frac{M_1 M_3 - M_2^2}{M_2 - M_1^2} \]

It can be done, but it's messy. Important special case: symmetric \( W \).

Say \( W(x) = W(-x) \). Then \( M_1, M_3, M_5, \ldots \) are 0.

\[ \psi_0 = x^0 \]

\[ \psi_1 = x' \]

\[ \psi_2 = x^2 - \frac{M_2}{M_2} x^0 \]

\[ \psi_3 = x^3 - \frac{M_4}{M_2} x^1 \]
"Classical" special cases:

\[ W(x) = e^{-x^2}, \quad x > 0 \quad \rightarrow \quad \text{Hermite polynomials} \]

\[ W(x) = e^{x^2}, \quad x < 0 \quad \rightarrow \quad \text{Laguerre polynomials} \]

uniform approx. en\'train

\[ W(x) = \begin{cases} \frac{1}{\sqrt{2}}, & |x| < 1 \\ 0, & |x| > 1 \end{cases} \quad \rightarrow \quad \text{Legendre polynomials} \]

\[ W(x) = \begin{cases} \frac{i}{\sqrt{2\pi}}, & |x| < 1 \\ 0, & |x| > 1 \end{cases} \quad \rightarrow \quad \text{Chebyshev} p-n's \]

\( x = \sin \theta \) makes connection with frequency-domain approach

\[ W(x) = \frac{1}{\sqrt{2\pi \rho}} e^{-x^2/2\rho} \]

\[ M_1, M_3, M_5, \ldots = 0 \]

\[ M_0 = 1, \quad M_2 = \frac{1}{2}, \quad M_4 = \frac{3}{8}, \quad M_6 = \frac{15}{16} \]

\[ M_{2n} = \frac{(2n)!}{2^n \cdot n!} \rho^n = \frac{(2n)(2n-1)(2n-2)(2n-3)\cdots 1}{2^n \cdot n!} \rho^n 
= \frac{2n-1)(2n-3)(2n-5)\cdots 1}{\rho^n} \]
\[ \begin{align*}
    h_0 &= 1 \\
    h_1 &= x \\
    h_2 &= x^2 - 3p \\
    h_3 &= x^3 - 3px \\
    h_4 &= x^4 - 6px^2 + 3p^2 \\
    h_5 &= x^5 - 10px^3 + 15p^2x \\
\end{align*} \]

The Hermite have many other properties, usually with parallels in the other classical families.

\[ \frac{dh_n}{dx} = nh_{n-1} \]

\[ \frac{dh_n}{dp} = \frac{n(n-1)}{2} h_{n-2} \quad (\text{non-} \text{normal}) \]

\[ h_{n+1} = x h_n - nh_{n-1} \quad \text{"recursion"} \]

\[ nh_n^2 = n! P^n \quad \text{"normalized"} \]

\[ h_n \text{ has } n \text{ roots, all real.} \]

\[ \sum_{n=0}^{\infty} \frac{h_n(x) e^x}{n!} = e^{-p^2/2} \quad \text{"Gaußian Factor"} \]
Generally, factoring in the specific case deducing all the other properties,

consider in \( E^{tx-Pt^{2}/2} \), how \( x^n, \ x^{n-2}, \ x^{n-4}, \ldots \), since

\[
e^u = \sum_{k=0}^{\infty} \frac{u^k}{k!} \quad (u = tx - \frac{Pt^2}{2})
\]

For example, to see that the \( h_k \)'s are orthogonal:

Let \( c_{jk} = \langle h_j, h_k \rangle = \int h_j(x) h_k(x) W(x) \, dx \).

Let \( C = \sum_{j=0}^{\infty} \frac{c_{jk}}{k!} s_j x^j \), \( \sum_{j=0}^{\infty} \int \frac{s_j}{k!} h_j(x) h_k(x) W(x) \, dx \).

\[
C = \int_{-\infty}^{\infty} \sum_{j=0}^{\infty} \frac{s_j}{j!} \frac{x^j}{k!} W(x) \, dx
\]

\[
= \int e^{tx-Pt^2/2} e^{-x^2/2P} dx = \frac{1}{\sqrt{4\pi P}}
\]

\[
= \frac{1}{\sqrt{2\pi P}} e^{-Ps/2 - Pt^2/2} \int e^{-x^2/2P + sx + tx} \, dx
\]
Completing the square:

$$\frac{x^2}{2p} + sx + tx = \frac{x^2}{2p} + sx + tx - \frac{p}{2} (s+t)^2 + \frac{p}{2} (s+t)^2$$

So

$$-\frac{x^2}{2p} + sx + tx = -\frac{1}{2p} (x - p(s+t))^2 + \frac{p}{2} (s+t)^2$$

$$C = \frac{1}{\sqrt{2\pi p}} \int e^{-\frac{p}{2} (x - p(s+t))^2 + \frac{p}{2} (s+t)^2} dx$$

A de-central Gaussian of variance $p$

$$C = e^{-\frac{p_s^2}{2} - \frac{p_t^2}{2} + \frac{p}{2} (s+t)^2} = C_{st}$$

$$C_{ab} = \Sigma \frac{C_{jk} s^j t^k}{j!k!}$$

So

$$\Sigma \frac{C_{jk} s^j t^k}{j!k!} = C_{st} = \Sigma \left(\frac{st}{n}\right)^n \frac{p^n}{n!}$$

So $C_{jk} = 0$ unless $j = k = n$

$$\frac{c_{nm}}{n!} = \frac{p^n}{n!}, \text{ so } c_n = n!$$
From orthogonal polynomials to orthogonal multionomials

Above, \( S_k = x^k, x^2, x^3, \ldots \) \( (x \) replaces \( n_0 x^n \) while we're approximating \( f(x) \).

What if we wanted to approximate \( f(xy) \)?

\( (x, y \) replace two \( n_0 x^n \)s, \( n_0 \) at two \( x \)s.)

\[ S_k = 1, x, x^2, x^3, \ldots, y, y^2, y^3, \ldots, xy, x^2y, \ldots, xy^2, \ldots \]

The orthogonalization process will depend on the order chosen; can't be approximating.

But we'd like \( f(xy) = A xy + B y \) to be simple to represent if \( A \neq B \) are both linear.

So,

\[ 1, x, y, x^2, xy, y^2, x^3, x^2y, xy^2, y^3, \ldots \]

With multiple variables, we'd like to do the analogous:

\[ 1, x_1, x_2, x_3, \ldots, x_Q, x_1^2, x_1 x_2, \ldots, x_2^2, x_2 x_3, \ldots \]

A notational catastrophe!

Use vectorized subscripts expanded: \( x_i^k = x_1^{k_1} x_2^{k_2} \ldots x_Q^{k_Q} \)

"Order" of \( k = \sum k_i \);

1 is 2nd-order;

\( x_1, \ldots, x_Q \) are 1st-order;

\( x_1^2, \ldots, x_2^2, x_1 x_2, \ldots \) have \( 2^nd \)-order (\( 2 \) kinds)
Orthogonalize as before for any $W(x) > 0$, for
factors $f(x)$ with $\int |f(x)|^2 W(x) dx < \infty$. 

Proceed column by column.

$W(x)$ is the distribution of $x_1, \ldots, x_k$; a multivariate distribution $x_1, \ldots, x_k$.

If $W(x) = W_0(x_1) W_0(x_2) \cdots W_0(x_k)$, then

$$
\begin{aligned}
\langle \mathbf{x}, \mathbf{x} \rangle &= \int x x^T W(x) dx \\
&= \left( \int x_1^T x_1 W(x_1, x_2) dx_1 \right) \cdots \left( \int x_k^T x_k W(x_1, x_2, \ldots, x_k) dx_k \right) \\
&= M_{k, 1} \cdots M_{k, k_2 + k_3} \quad \text{so we immediately have}
\end{aligned}
$$

$$
\phi_\mathbf{x}^T = \phi_{k_1}(x_1) \phi_{k_2}(x_2) \cdots \phi_{k_d}(x_d).
$$

For Gaussian case:

$$
\phi_0 = 1
$$

$\phi_0 \cdots \phi_0(x) = \phi_0(x) = x_k$

$\phi_0 \cdots \phi_0(x) = \phi_0(x_k) = x_k^2 - 1 \quad \Rightarrow \quad |\phi_0|^2 = 2\rho^2$

$\phi_0 \cdots \phi_0(x) = \phi_0(x_k) \phi_0(x_k) = x_k^2 \quad \Rightarrow \quad |\phi_0|^2 = \rho^2$
\[ \psi_0 \cdots 3 \cdots 0 (x) = \psi_3(x_4) = x_4^3 - 3^2 x_4, \quad \|\psi_3\|^2 = 6p^3 \]

\[ \psi_0 \cdots 2 \cdots 1 \cdots 0 (x) = \psi_2(x_4)\psi_1(x_5) = x_4^2x_5 - 2^2x_5, \quad \|\psi_2\|^2 = 2p^3 \]

\[ \psi_0 \cdots 1 \cdots 1 \cdots 0 (x) = \psi_1(x_4)\psi_1(x_5)\psi_0(x_6) = x_4x_5x_6^2, \quad \|\psi_1\|^2 = p^3. \]

Does this make sense? \( Q \geq \infty \)? Yes, infinite number of basis elements of each order.

\[ f(x) = \sum q_k \psi_k = a_0 \]

\[ + \sum \epsilon_k q_{k_0 \cdots 1 \cdots 0} \psi_{k_0 \cdots 1 \cdots 0} (x) \]

\[ + \sum \epsilon_k q_{k_0 \cdots 2 \cdots 0} \psi_{k_0 \cdots 2 \cdots 0} (x) + \epsilon_{k_0 \cdots 1 \cdots 0} \psi_{k_0 \cdots 1 \cdots 0} (x) \cdots \]

\[ = a_0 + \sum q_k x_k + \sum b_k (x_k^2 - p) + \sum c_{kl} x_k x_l + \cdots \]

\[ b_k = a_0 \cdots 2 \cdots 0, \quad c_{kl} = a_0 \cdots 1 \cdots 1 \cdots l, \cdots k. \]

Each \( a, b, \cdots \) obtained by

\[ a_k = \frac{\langle f(x)\psi_k(x) \rangle}{\langle \psi_k(x)\psi_k(x) \rangle}. \]