PROPOSITION 1

The L_2 Boost estimate in iteration m can be represented as:

$$\hat{F}_m = \sum_{j=0}^{m} S(I-S)^j y = \underbrace{\left(I-(I-S)^{m+1}\right)}_{B_m \text{ boosting operator}} y$$

such that $\hat{F}_m = B_m y$

Proof

L_2 Boost = gradient boosting with L_2 loss; and linear learners:

$$L = L_2 = \frac{1}{2}(y-f)^2$$

$\Rightarrow$ negative gradient = residuals:

$$-\left.\frac{\partial L}{\partial f}\right|_{f=\hat{f}_{m-1}} = -\left.\frac{\partial}{\partial f}\left[\frac{1}{2}(y-f)^2\right]\right|_{f=\hat{f}_{m-1}} =$$

$$= -\left[-2\cdot\frac{1}{2}\left(y-\hat{f}_{m-1}\right)\right] = y-\hat{f}_{m-1}$$

$\Rightarrow$ WE ARE ITERATIVELY FITTING OUR LEARNER ON THE RESIDUALS OF THE MODEL AT THE PREVIOUS STEP.

At the beginning of each step j we have the residuals from the previous step minus the update we just added:

$$v_j = v_{j-1} - \hat{f}_{j-1} = v_{j-1} - S v_{j-1} = (I-S)v_{j-1} \quad \text{for } j=1,\dots,$$

Similarly, $u_{j-1} = u_{j-2} - S u_{j-2}$

$$\Rightarrow u_j = (I-S)(u_{j-2} - S u_{j-2}) =$$

$$= (I-S)\left[(I-S)u_{j-2}\right] =$$

$$= (I-S)^2 u_{j-2} = (I-S)^2\left[(I-S)u_{j-3}\right] =$$

$$= (I-S)^3 u_{j-3} = \dots = (I-S)^j u_0 = (I-S)^j y$$

($u_0 = y$)

Therefore, the update at step j is:

$$\hat{f}_j = S u_j = S(I-S)^j y .$$

~~The~~ ~~first~~ ~~estimate~~

The estimate at step m is the sum of the contributions from each step:

$$\hat{f}_m = \sum_{j=0}^{m} \hat{f}_j = \sum_{j=0}^{m} S(I-S)^j y$$

for $m=0$ (FIRST STEP): $\hat{f}_0 = Sy$ i.e. we just apply the smoother to the data

$m=1$ $\hat{f}_1 = \sum_{j=0}^{1} S(I-S)^j y =$

$= \underbrace{Sy}_{\hat{f}_0} + S\underbrace{(I-S)y}_{u_1}$ u_1: residuals after step 1.

TELESCOPIC SUM (subsequent terms cancel each other, leaving only initial and final terms)

$$\left[I - (I-S)^{m+1}\right] y = B_m y .$$

PROPOSITION 3

Linear, symmetric weak learner S with eigenvalues $\{\lambda_k ; k = 1, \dots, n\}$

- $\text{bias}^2(m, S; f) = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{E}[\hat{F}_m(x_i)] - f(x_i) \right)^2 =$

$$= \frac{1}{n} f^T U \, \text{diag}\left[(1 - \lambda_k)^{2m+2} \right] U^T f$$

- $\text{Var}(m, S; \sigma^2) = \frac{1}{n} \sum_{i=1}^{n} \text{Var}(\hat{F}_m(x_i)) =$

$$= \frac{1}{n} \cdot \sigma^2 \sum_{i=1}^{n} \left(1 - (1 - \lambda_k)^{m+1} \right)^2$$

and $\text{MSE}(m, S; f, \sigma^2) = \text{bias}^2 + \text{Var}$

Proof

Note:

- Eigenvalues of B_m: $\{1 - (1 - \lambda_k)^{m_{stop}+1} ; k = 1, \dots, N\}$
- Since S is symmetric (i.e. $S = S^T$), B_m can be diagonalized with an orthonormal transformation:

$$B_m = U D_m U^T \quad \text{with} \quad D_m = \text{diag}\left[1 - (1 - \lambda_k)^{m_{stop}+1} \right]$$

$$\text{and} \quad U U^T = U^T U = I.$$

- $\hat{F}_m = B_m Y \quad \text{with} \quad B_m = I - (I - S)^{m+1}$

BIAS TERM

$$\text{bias}^2 = \frac{1}{n}\left[\mathbb{E}[\hat{f}_m] - f\right]^T\left[\mathbb{E}[\hat{f}_m] - f\right] =$$

$$= \frac{1}{n}\left[\mathbb{E}[B_m y] - f\right]^T\left[\mathbb{E}[B_m y] - f\right] =$$

$$= \frac{1}{n}\left[B_m \mathbb{E}[y] - f\right]^T\left[B_m \mathbb{E}[y] - f\right] =$$

$$= \frac{1}{n}\left[B_m f - f\right]^T\left[B_m f - f\right] =$$

$$= \frac{1}{n}\left[(B_m - I) f\right]^T\left[(B_m - I) f\right]$$

where $B_m - I = U D_m U^T - U U^T =$

$$= U(D_m - I)U^T =$$

$$= U\left[\text{diag}\left(1 - (1-\lambda_k)^{m+1}\right) - I\right]U^T =$$

$$= U\left[\text{diag}\left(1 - (1-\lambda_k)^{m+1} - I\right)\right]U^T =$$

$$= U\ \text{diag}\left(-(1-\lambda_k)^{m+1}\right)U^T$$

giving

$$\text{bias}^2 = \frac{1}{n}\left[U\,\text{diag}\left(-(1-\lambda_k)^{m+1}\right)U^T f\right]^T\left[U\,\text{diag}\left(-(1-\lambda_k)^{m+1}\right)U^T f\right] =$$

$$= \frac{1}{n} f^T U\,\text{diag}\left(-(1-\lambda_k)^{m+1}\right)\underbrace{U^T U}_{=I}\,\text{diag}\left(-(1-\lambda_k)^{m+1}\right)U^T f =$$

$$= \frac{1}{n} f^T U\,\text{diag}\left((1-\lambda_k)^{2m+2}\right)U^T f$$

VARIANCE

$$Var = \frac{1}{n}\sum_{i=1}^{n} Var\left(\hat{f}_m(x_i)\right) = \frac{1}{n}\sum_{i=1}^{n} Var(B_m y_i) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} B_m^2 \underbrace{Var(y_i)}_{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} \left(1-(1-\lambda_k)^{m+1}\right)^2 \sigma^2 =$$

$$= \frac{\sigma^2}{n}\sum_{i=1}^{n} \left(1-(1-\lambda_k)^{m+1}\right)^2 .$$