Some Results on the Least Squares Formula

Praveen C
CFD Center, Department of Aerospace Engg.,
Indian Institute of Science, Bangalore

Abstract

In this report, we establish some interesting invariance properties of the first order least squares formula. It is shown that it is consistent with the transformation properties of the exact derivatives, both in the case of a scalar and a vector field variable. This leads to the result that when the full stencil is used (i.e. no upwinding), there are no optimum coordinate directions which will minimize some relevant norm of the error vector. A numerical example with upwinding shows the existence of optimum coordinate directions. Finally, some length scales are derived based on the eigenvalues of the least squares matrix which help in characterizing the connectivity for LSKUM.

1 Least Squares Formula

We will first derive the first order least squares formula. Let x^a, $a = 1, 2, .., d$ be the coordinates with d being the number of dimensions. Let $\phi(x^a)$ be a sufficiently smooth function that is given to us. Assume that the values of ϕ are known at n distinct points, with $n \geq d + 1$, which are close to the point $\vec{x}_o$, together with that at $\vec{x}_o$, i.e.,

$$\phi_o, \phi_1, \phi_2, ..., \phi_n$$

are the values taken by ϕ at the points,

$$\vec{x}_o, \vec{x}_1, \vec{x}_2, ..., \vec{x}_n$$

We say that the above points are in the *connectivity* of $\vec{x}_o$. Without loss of generality, assume that $\vec{x}_o = \vec{0}$ and $\phi_o = \phi(\vec{x}_o) = 0$. Since ϕ is smooth, we can expand it using Taylor's formula, about $\vec{x}_o$, i.e.,

$$\phi(\vec{x}) = \phi(\vec{x}_o) + (\vec{x} - \vec{x}_o) \cdot \nabla\phi_o + \mathcal{O}(|\vec{x} - \vec{x}_o|^2)$$

Hence we have

$$\phi(\vec{x}_i) = \phi_i = \vec{x}_i \cdot \nabla\phi_o + \mathcal{O}(|\vec{x} - \vec{x}_o|^2), \ i = 1, ..., n$$

Neglecting the second and higher order terms, let us define the error,

$$E_i = \phi(\vec{x}_i) - \vec{x}_i \cdot \nabla\phi_o, \ i = 1, ..., n$$

We could try to solve these equations by setting $E_i = 0$. But since there are d number of unknowns (derivatives at $\vec{x}_o$), while there are $n > d$ equations, we are in the situation of an over-determined problem. The natural way to solve this problem is to define the norm,

$$\|E\| = \sum_{i=1}^{n} w_i E_i^2 \tag{1.1}$$

and determine the derivatives in such a way that this norm is minimized, i.e., by setting,

$$\frac{\partial}{\partial(\partial_a \phi_o)} \|E\| = 0, \ a = 1, ..., d \tag{1.2}$$

This gives the following system of d linear equations in the d unknowns,

$$\sum_{i=1}^{n} w_i(\phi_i - \vec{x}_i \cdot \nabla\phi_o) x_i^a = 0, \ a = 1, ..., d \tag{1.3}$$

In equation (1.1), we have assumed $w_i = w(|\vec{x}_i - \vec{x}_o|)$ to be a positive weighting function, which gives relative importance to the data ϕ_i based on the proximity of $\vec{x}_i$ to $\vec{x}_o$. Define,

$$A_{ab} = \sum_{i=1}^{n} w_i x_i^a x_i^b \tag{1.4}$$

Then the matrix $A = [A_{ab}]$ is a $d \times d$ symmetric matrix, and we can write (1.3) as,

$$A\nabla\phi_o = f \tag{1.5}$$

where $\nabla\phi_o$ is to interpreted as a matrix-vector, and f is a matrix-vector given by,

$$f^a = \sum_{i=1}^{n} w_i \phi_i x_i^a, \ a = 1, ..., d \tag{1.6}$$

We will call the matrix A defined by (1.4) as the *least squares matrix.*

Lemma 1 *The least squares matrix is symmetric positive definite.*

Proof: The proof follows easily from the fact that we can write A as,

$$A = XX^t \tag{1.7}$$

where X is given by,

$$X = \begin{bmatrix} w_1 x_1^1 & w_2 x_2^1 & ..... & w_n x_n^1 \\ : & & & : \\ : & & & : \\ : & & & : \\ w_1 x_1^d & w_2 x_2^d & ..... & w_n x_n^d \end{bmatrix}$$

Lemma 2 *The least squares matrix has positive eigenvalues and a complete set of linearly independent, mutually orthogonal eigenvectors, provided* $\det A \neq 0$.

Proof: The positivity of the eigenvalues follows from the fact that A is positive definite and $\det A \neq 0$. Since A is symmetric, it has a full set of mutually orthogonal eigenvectors.

Example 1 *(One dimensional case)*

The least squares formula in one dimensions is given by,

$$\frac{\mathrm{d}\phi}{\mathrm{d}x} = \frac{\sum_{i=1}^{n} w_i \phi_i x_i}{\sum_{i=1}^{n} w_i x_i^2} \tag{1.8}$$

Example 2 *(Two dimensional case)*

The derivatives are given by,

$$\frac{\partial \phi}{\partial x} = \frac{\sum w_i y_i^2 \sum w_i \phi_i x_i - \sum w_i x_i y_i \sum w_i \phi_i y_i}{\sum w_i x_i^2 \sum w_i y_i^2 - (\sum w_i x_i y_i)^2} \tag{1.9}$$

$$\frac{\partial \phi}{\partial y} = \frac{\sum w_i x_i^2 \sum w_i \phi_i y_i - \sum w_i x_i y_i \sum w_i \phi_i x_i}{\sum w_i x_i^2 \sum w_i y_i^2 - (\sum w_i x_i y_i)^2} \tag{1.10}$$

where all summations extend from $i = 1, ..., n$. The denominator in the above two formula is the determinant of the least squares matrix, and by Cauchy-Schwarz inequality,

$$D = \sum w_i x_i^2 \sum w_i y_i^2 - \left(\sum w_i x_i y_i \right)^2 \geq 0 \tag{1.11}$$

Equality holds if and only if there exists a constant m such that,

$$y_i = m x_i, \;\; i = 1, ..., n$$

i.e., when all the points $\vec{x}_i$ lie on a straight line passing through $\vec{x}_o$.

We next give explicit expressions for the eigenvalues and eigenvectors for two two-dimensional case. Setting $r_i = |\vec{x}_i - \vec{x}_o|$, we have

$$\lambda_1 = \frac{1}{2} \left\{ \sum w_i r_i^2 + \sqrt{\left(\sum w_i r_i^2 \right)^2 - 4D} \right\} \geq 0$$

$$\lambda_2 = \frac{1}{2} \left\{ \sum w_i r_i^2 - \sqrt{\left(\sum w_i r_i^2 \right)^2 - 4D} \right\} \geq 0$$

The eigenvalues are real because,

$$\left(\sum w_i r_i^2 \right)^2 - 4D = \left(\sum w_i (x_i^2 - y_i^2) \right)^2 + 4 \left(\sum w_i x_i y_i \right)^2 \geq 0$$

and obviously positive. Moreover $\lambda_1 \geq \lambda_2 \geq 0$ and $\lambda_1 = \lambda_2$ if and only if,

$$\sum w_i x_i^2 = \sum w_i y_i^2, \quad \sum w_i x_i y_i = 0$$

We also see that $\lambda_2 = 0$ if and only if $D = 0$.

The eigenvectors are given by ($D \neq 0$),

$$e_1 = \begin{bmatrix} 1 \\ \\ \dfrac{\lambda_1 - \sum w_i x_i^2}{\sum w_i x_i y_i} \end{bmatrix}$$

$$e_2 = \begin{bmatrix} 1 \\ \\ \dfrac{\lambda_2 - \sum w_i x_i^2}{\sum w_i x_i y_i} \end{bmatrix}$$

It can be easily shown that e_1 and e_2 are orthogonal.

Lemma 3 *The least squares matrix is invariant under translation.*

Proof: This follows easily because equation (1.4) is actually,

$$A_{ab} = \sum w(|\vec{x}_i - \vec{x}_o|)(x_i^a - x_o^a)(x_i^b - x_o^b)$$

This corresponds to the translation invariance of the least squares derivatives and is a property shared by the exact derivatives.

Lemma 4 *The least squares matrix undergoes a similarity transformation under a coordinate rotation.*

Proof: Let the coordinates x^a be transformed to ξ^b, which can be achieved by an orthogonal matrix $R = [l_b^a]$, i.e,

$$\xi = Rx$$

or in terms of the coordinates,

$$\xi^a = \sum_{b=1}^{d} l_b^a x^b = l_b^a x^b$$

The least squares matrix $\bar{A}$ in the new coordinates is given by,

$$\begin{aligned} \bar{A}_{ab} &= \sum_i w_i \xi_i^a \xi_i^b \\ &= \sum_i w_i (l_p^a x_i^p)(l_q^b x_i^q) \\ &= l_p^a A_{ab} l_q^b \end{aligned}$$

which implies that $\bar{A} = RAR^t$. Since R is an orthogonal matrix, $R^t = R^{-1}$, so that,

$$\bar{A} = RAR^{-1}$$

Corollary 1 *The determinant and the eigenvalues of the least squares matrix are invariant under coordinate translation and rotation.*

Proof: The invariance under translation follows because the least squares matrix is itself invariant. The invariance under rotation follows because of the similarity transformation[1].

Corollary 2 *The least squares matrix becomes diagonal in the eigenvector-frame, with the eigenvalues along the diagonal.*

Proof: Since A is symmetric, it is diagonalizable, i.e.,

$$A = \mathcal{R}\Lambda\mathcal{R}^{-1}$$

where $\mathcal{R}$ is the matrix of eigenvectors of A and Λ is the diagonal matrix with the eigenvalues along the diagonal. In the eigenvector frame,

$$\bar{A} = \mathcal{R}A\mathcal{R}^{-1} = \Lambda$$

Example 3

The least squares formula takes a particularly simple form in the eigenvector-frame. Let ξ be one of the coordinates in the eigenvector-frame and ϕ be a scalar function. Then

$$\frac{\partial\phi}{\partial\xi} = \frac{\sum_i w_i\xi_i\phi_i}{\sum_i w_i\xi_i^2}$$

Lemma 5 *If $\phi(\vec{x})$ is a scalar then the least squares gradient $\nabla\phi$ transforms like the exact gradient.*

Proof: First notice that,

$$\begin{aligned}\bar{f}^a &= \sum_i w_i\phi_i\xi_i^a \\ &= \sum_i w_i\phi_i l_b^a x_i^b \\ &= l_b^a f^b\end{aligned}$$

This implies that,

$$\bar{f} = Rf$$

Hence,

$$\begin{aligned}\nabla_\xi\phi &= \bar{A}^{-1}\bar{f} \\ &= (RA^{-1}R^{-1})(Rf) \\ &= R(A^{-1}f) \\ &= R\nabla_x\phi\end{aligned}$$

[1]See appendix (A) for a different proof in the case of 2-D

Lemma 6 *If $\vec{u}(\vec{x})$ is a vector, then $\nabla\vec{u}$ is a second order tensor. The least squares estimate of $\nabla\vec{u}$ transforms like the exact tensor.*

Proof: The least squares estimate of $\nabla\vec{u}$ can be written as,

$$A\nabla\vec{u} = F$$

where $F = [F^{ab}]$ and,

$$F^{ab} = \sum_i w_i u_i^a x_i^b$$

Under a coordinate rotation, F^{ab} transforms as,

$$\begin{aligned}
\bar{F}^{ab} &= \sum_i w_i \bar{u}_i^a \xi_i^b \\
&= \sum_i w_i (l_p^a u_i^p)(l_q^b x_i^q) \\
&= l_p^a F^{pq} l_q^b
\end{aligned}$$

This implies that,

$$\bar{F} = RFR^t = RFR^{-1}$$

Hence,

$$\begin{aligned}
\nabla_\xi \vec{u} &= \bar{A}^{-1}\bar{F} \\
&= (RA^{-1}R^{-1})(RFR^{-1}) \\
&= R(A^{-1}F)R^{-1} \\
&= R(\nabla_x \vec{u})R^t
\end{aligned}$$

2 Transformation of Tensor-divergence

Let $\vec{u}$ be the momentum vector and σ be the momentum flux tensor. Then, the momentum equation is,

$$\frac{\partial \vec{u}}{\partial t} + \text{div } \sigma = 0$$

The update equation for the momentum is,

$$\vec{u}^{n+1} = \vec{u}^n - \Delta t \text{div } \sigma^n$$

Define the quantity,

$$v^a = \frac{\partial \sigma^{ab}}{\partial x^b} \tag{2.1}$$

which is obviously a vector. The next lemma shows that the least squares estimate of this quantity transforms like a vector. If the divergence is evaluated using the least

squares formula with full stencil, then this result shows that the update is invariant under coordinate rotations. This property is a prerequisite for any genuinely multi-dimensional upwind method.

Lemma 7 *The least squares estimate of the quantity v^a as defined in equation (2.1) transforms like a vector.*

Proof: Let $B = A^{-1} = [b_{ab}]$. Then $\bar{B} = \bar{A}^{-1} = RA^{-1}R^{-1} = RBR^t$. Then

$$\begin{aligned} \bar{v}_c^{ab} &= \frac{\partial \bar{\sigma}^{ab}}{\partial \bar{x}^c} \\ &= \bar{b}_{ce} \sum_i w_i \bar{x}_i^e \bar{\sigma}_i^{ab} \\ &= (l_c^p b_{pq} l_e^q) \sum_i w_i (l_r^e x_i^r)(l_s^a \sigma_i^{st} l_t^b) \end{aligned}$$

Omitting the summation sign, we get,

$$\begin{aligned} \bar{v}^a &= \bar{v}_b^{ab} \\ &= \frac{\partial \bar{\sigma}^{ab}}{\partial \bar{x}^b} \\ &= (l_b^p b_{pq} l_e^q) w_i (l_r^e x_i^r)(l_s^a \sigma_i^{st} l_t^b) \\ &= (l_b^p l_t^b)(l_e^q l_r^e) b_{pq} w_i x_i^r l_s^a \sigma_i^{st} \\ &= \delta_t^p \delta_r^q b_{pq} w_i x_i^r l_s^a \sigma_i^{st} \\ &= b_{pq} w_i x_i^q l_s^a \sigma_i^{sp} \\ &= l_s^a b_{pq} \sum_i w_i x_i^q \sigma_i^{sp} \\ &= l_s^a \frac{\partial \sigma^{sp}}{\partial x^p} \\ &= l_s^a v^s \end{aligned}$$

This proves the lemma.

3 Optimal Coordinate System

The least squares formula can be used to determine the gradients in any direction and does not rely on the existence of coordinate lines in the grid. Hence it is interesting to ask whether there is a coordinate direction in which the error in the gradients is minimized. Let ϕ be scalar function and $\vec{u}$ a vector function. Then we define the errors in the gradients,

$$E_v = E_v(\theta) = \nabla_\xi \phi - (\nabla_\xi \phi)^e \tag{3.1}$$

$$E_t = E_t(\theta) = \nabla_\xi \vec{u} - (\nabla_\xi \vec{u})^e \tag{3.2}$$

where θ gives the orientation of the axes with reference to some fixed axes. We have the following simple consequence of lemma (5)-(6).

Corollary 3 *The errors defined in equation (3.1) and (3.2) transform like a vector and a second order tensor respectively.*

We thus obtain the following important result.

Theorem 1 *The errors defined by (3.1) and (3.2) satisfy the property,*

$$\|E_v(\theta)\|_2 = \|E_v(0)\|_2 \,, \ \forall\theta \tag{3.3}$$

$$\text{trace}\, E_t(\theta) = \text{trace}\, E_t(0), \ \forall\theta \tag{3.4}$$

Proof: $E_v(\theta)$ transforms like a vector and hence it's l^2-norm is invariant. $E_t(\theta)$ transforms like a second order tensor and hence it's trace is invariant.

Remark: The above theorem effectively says that when we use the least squares formula together with the full stencil, then there is no local coordinate system in which some relevant norm of the error is minimized. For a scalar function, the l^2-norm of the error in it's gradient is invariant. In the case of a vector function, we usually require only its trace in the numerical scheme. For example, the continuity equation is,

$$\frac{\partial \rho}{\partial t} + \frac{\partial \rho u}{\partial x} + \frac{\partial \rho v}{\partial y} = 0$$

Here only the trace of $\nabla \rho \vec{u}$ is required.
Conclusion: When the full stencil is used, there is no optimal coordinate system.

4 Numerical Example

The above results have been checked by taking some distribution of points which is shown in figure (1) and computing the gradients using the least squares formula for given scalar and vector functions. The functions chosen are,

$$\phi = xy, \qquad \vec{u} = (xy, x^2)$$

The results of computations are shown in appendix (B) which clearly show the invariance properties.

5 Numerical Example with Upwind Biasing

Let us choose a vector function,

$$\vec{u} = \left(xy, \frac{1}{2}(x^2 + y^2) \right)$$

and a KFVS-like splitting,

$$F_i^{\pm} = u_i A_i^{\pm} \pm B_i$$

where

$$A_i^{\pm} = \frac{1}{2}(1 + \mathrm{erf}\,(u_i)), \qquad B_i = \frac{1}{\sqrt{4\pi}} \exp(-u_i^2)$$

which corresponds to taking $\rho = \beta = 1$ in the mass flux. The trace of $\nabla \vec{u}$ is written as,

$$\nabla \cdot \vec{u} = \sum_{i=1}^{2} \left(\frac{\partial F_i^+}{\partial x_i} + \frac{\partial F_i^-}{\partial x_i} \right)$$

The exact value of the trace is,

$$(\nabla \cdot \vec{u})^e = 2y$$

Figure (2) shows the variation of the trace calculated using upwinding. The two figures correspond to two different *streamline* directions. *It is seen that there are certain rotated coordinate axes in which the error is a minimum.* Note that these angles are independent of the weights used. The dependence of these minimum-error angles must be determined in terms of the flux-splitting and the least-squares formula.

6 Length Scales and the Bounding Ellipse

Assuming that $\vec{x}_o = 0$, define the average coordinates,

$$\bar{x}_o = \frac{1}{n} \sum_{j=1}^{n} x_j, \quad \bar{y}_o = \frac{1}{n} \sum_{j=1}^{n} y_j$$

and the new coordinates with respect to $(\bar{x}_o, \bar{y}_o)$,

$$X_i = x_i - \bar{x}_o, \quad Y_i = y_i - \bar{y}_o \tag{6.1}$$

Let $\underline{\mathrm{A}}$ be the least squares matrix constructed from (X_i, Y_i). Let $\underline{\lambda}_1$, $\underline{\lambda}_2$ be the eigenvalues and $\underline{e}_1$, $\underline{e}_2$ be the corresponding eigenvectors, with $\underline{\lambda}_1 \geq \underline{\lambda}_2$. Define the lengths,

$$L_\lambda = \sqrt{\underline{\lambda}_1} \tag{6.2}$$

$$l_\lambda = \sqrt{\underline{\lambda}_2} \tag{6.3}$$

so that

$$l_\lambda \leq L_\lambda \tag{6.4}$$

We define *the bounding ellipse to be the ellipse with centre* $(\bar{x}_o, \bar{y}_o)$*, semi-major and semi-minor axes of lengths* L_λ*,* l_λ *oriented along* $\underline{e}_1$*,* $\underline{e}_2$ *respectively.*

Figure (3) illustrate the bounding ellipse for a given connectivity. We can draw the following conclusions.

1. The bounding ellipse completely encloses the connectivity.
2. The lengths L_λ, l_λ indicate the length scales in the connectivity and the aspect ratio of the ellipse L_λ/l_λ indicates the aspect ratio of the connectivity.
3. The optimum way to reduce the aspect ratio is by adding extra points along $\underline{e}_2$.
4. The orientation of the ellipse indicates the major directional orientation of the connectivity.
5. For a connectivity in which all the $\vec{x}_i$, $i = 1, ..., n$ are on a straight line, the bounding ellipse degenerates to a straight line, i.e., $l_\lambda = 0$.

The above properties can be used to assess the quality of a given connectivity and to improve it if required. Since the eigenvectors are orthogonal, we can apply the least squares formula in the eigenvector frame with upwinding. Anandhanarayan has found improvement in the convergence characteristics when this is done for q-LSKUM. Investigation of these aspects is being planned.

A Invariance of the determinant in 2-D

Let (x, y) be some fixed axes, and $(\bar{x}, \bar{y})$ be some axes obtained by rotating (x, y) through an angle θ. The determinant is given by,

$$D(\theta) = \sum \bar{x}_i^2 \sum \bar{y}_i^2 - \left(\sum \bar{x}_i \bar{y}_i\right)^2$$

Since,

$$\bar{x} = x \cos\theta + y \sin\theta$$

$$\bar{y} = -x \sin\theta + y \cos\theta$$

we have

$$\frac{\mathrm{d}\bar{x}}{\mathrm{d}\theta} = \bar{y}, \quad \frac{\mathrm{d}\bar{y}}{\mathrm{d}\theta} = -\bar{x}$$

Hence,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\theta} D(\theta) &= 2 \sum \bar{x}_i \frac{\mathrm{d}\bar{x}_i}{\mathrm{d}\theta} \sum \bar{y}_i^2 + 2 \sum \bar{x}_i^2 \sum \bar{y}_i \frac{\mathrm{d}\bar{y}_i}{\mathrm{d}\theta} \\
&\quad -2 \left(\sum \bar{x}_i \bar{y}_i\right) \left(\sum \bar{x}_i \frac{\mathrm{d}\bar{y}_i}{\mathrm{d}\theta} + \sum \frac{\mathrm{d}\bar{x}_i}{\mathrm{d}\theta} \bar{y}_i\right) \\
&= 2 \sum \bar{x}_i \bar{y}_i \sum \bar{y}_i^2 - 2 \sum \bar{x}_i^2 \sum \bar{y}_i \bar{x}_i \\
&\quad -2 \left(\sum \bar{x}_i \bar{y}_i\right) \left(-\sum \bar{x}_i^2 + \sum \bar{y}_i^2\right) \\
&= 0
\end{aligned}$$

which implies that,

$$D(\theta) = \text{constant}$$

We have not indicated the weights in the above derivation, but the same result holds with weights since they are independent of θ.

B Numerical Example of Invariance

```
---------------------------------------------------------------
                      Scalar Function
---------------------------------------------------------------
ANGLE = 0.000000, Determinant = 2.010650, norm = 0.066661
             Computed          Exact             Error
phix         0.981231          1.000000          -0.018769
phiy         0.936036          1.000000          -0.063964
---------------------------------------------------------------
ANGLE = 45.000000, Determinant = 2.010650, norm = 0.066661
             Computed          Exact             Error
phix         1.355712          1.414214          -0.058501
phiy        -0.031958          0.000000          -0.031958
---------------------------------------------------------------
ANGLE = 72.000000, Determinant = 2.010650, norm = 0.066661
             Computed          Exact             Error
phix         1.193440          1.260074          -0.066634
phiy        -0.643955         -0.642040          -0.001916
---------------------------------------------------------------
                      Vector Function
---------------------------------------------------------------
ANGLE = 0.000000, trace = -0.030755
    Computed                 Exact                Error
0.981231   0.936036   1.000000   1.000000  -0.018769   -0.063964
1.937545  -0.011986   2.000000   0.000000  -0.062455   -0.011986
---------------------------------------------------------------
ANGLE = 45.000000, trace = -0.030755
    Computed                 Exact                Error
1.921413  -0.997363   2.000000  -1.000000  -0.078587   0.002637
0.004146  -0.952168   0.000000  -1.000000   0.004146   0.047832
---------------------------------------------------------------
ANGLE = 72.000000, trace = -0.030755
    Computed                 Exact                Error
 0.927382  -1.955042   0.977169  -2.007418  -0.049788   0.052376
-0.953532   0.041863  -1.007418   0.022831   0.053886   0.019033
---------------------------------------------------------------
```

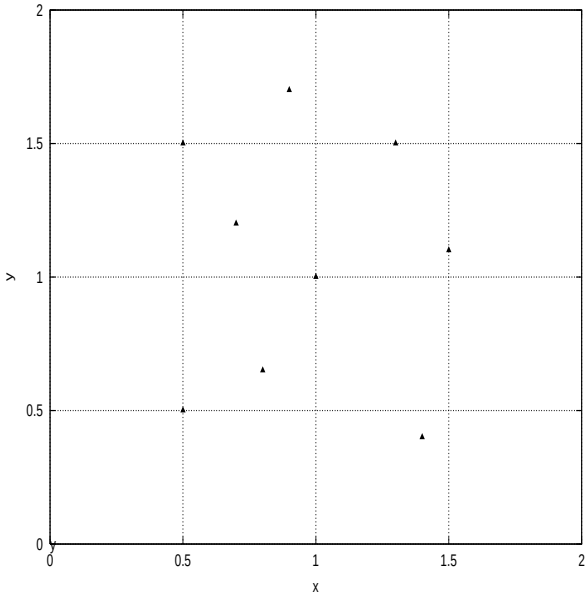
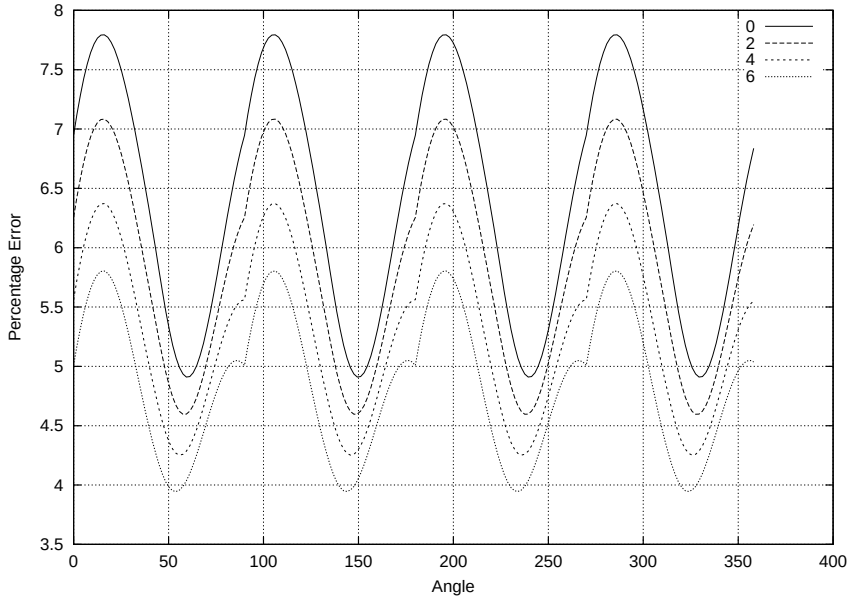


Figure 1: Point distribution

$\alpha = 45^o$



$\alpha = 51^o$

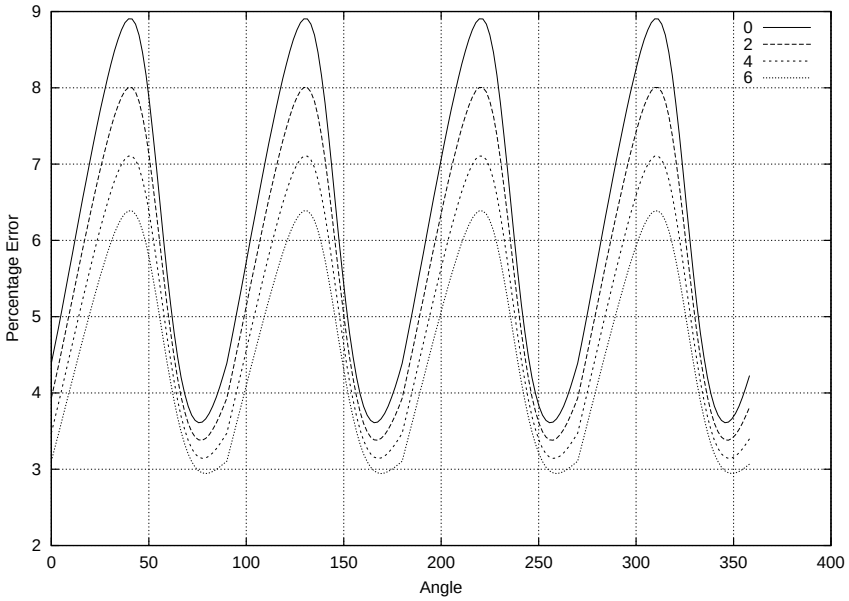


Figure 2: Percentage error in the trace with upwind biasing. The numbers indicate the weights.

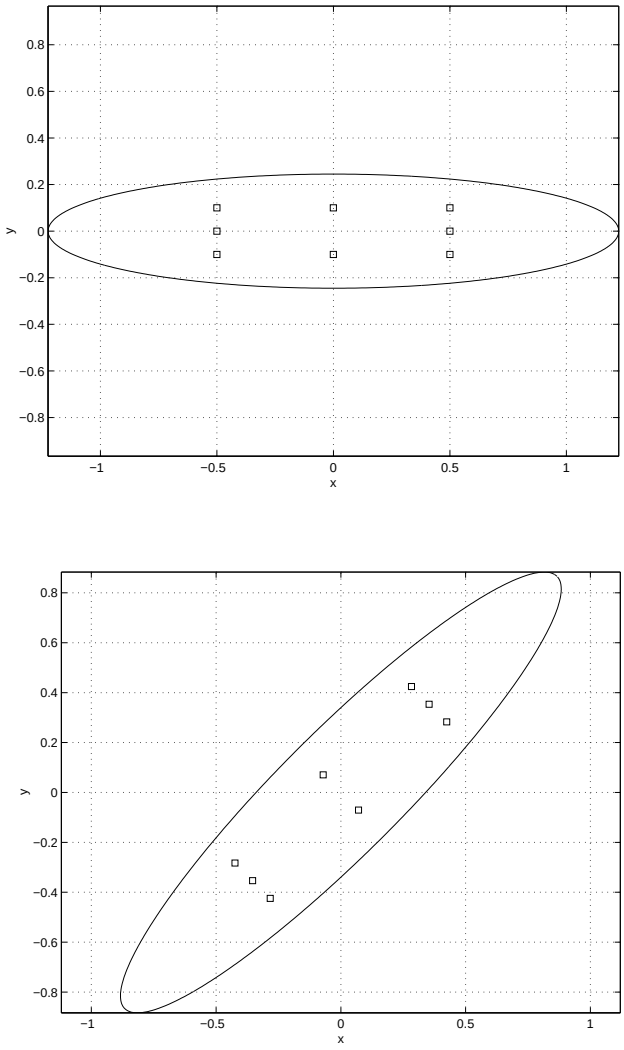


Figure 3: Illustration of the bounding ellipse for 2-D connectivity.