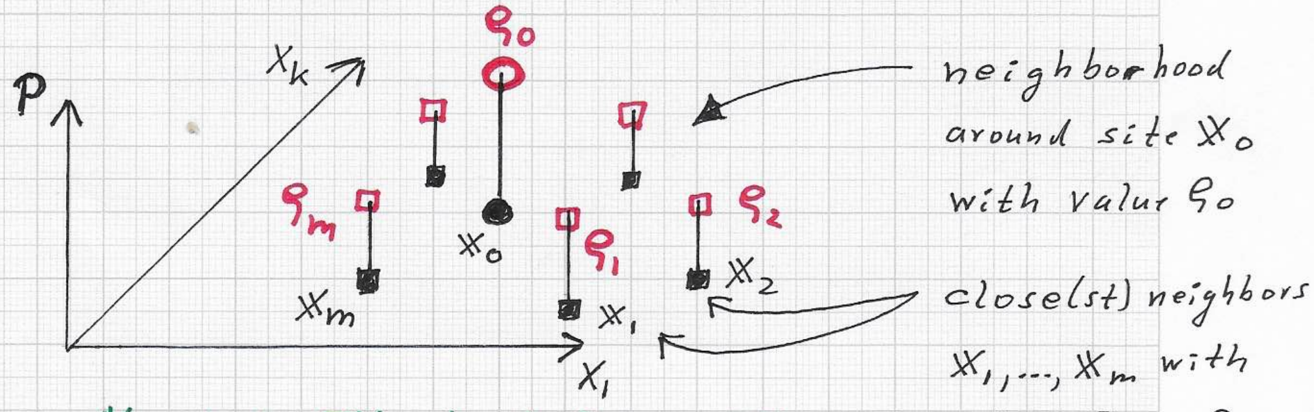
DISTRIBUTION APPROXIMATION – Cont'd.

→ Considerations...

13) Gradients must be estimated for all sites $\mathbb{x}_i$ with density values ϱ_i. Gradient estimates are denoted as ∇_i. Local linear Taylor approximations of probability density function $p(\mathbb{x})$ can be defined per site as

$$p_i(\mathbb{x}) = \varrho_i + \nabla_i \cdot (\mathbb{x} - \mathbb{x}_i) \text{ over } k\text{-dim. feature space.}$$



Needed: ∇_0 to define $p_0(\mathbb{x})$

- Use least squares method to compute linear or quadratic polynomial using the data $\{(\mathbb{x}_i, \varrho_i)\}_{i=0}^{m}$

(i) Compute linear polynomial $L_0(\mathbb{x}) = c_0 + \sum_{j=1}^{k} c_j x_j$:

$$\Downarrow$$

$$\nabla_0 = (c_1, \ldots, c_k)$$

$$c_0 + c_1 x_1^0 + \ldots c_k x_k^0 = \varrho_0$$
$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$
$$c_0 + c_1 x_1^m + \ldots c_k x_k^m = \varrho_m$$

$$\begin{bmatrix} 1 & x_1^0 & \ldots & x_k^0 \\ \vdots & & & \vdots \\ 1 & x_1^m & \ldots & x_k^m \end{bmatrix} \begin{bmatrix} c_0 \\ \vdots \\ c_k \end{bmatrix} = \begin{bmatrix} \varrho_0 \\ \vdots \\ \varrho_m \end{bmatrix} \Leftrightarrow \mathbb{X}\mathbb{c} = \varrho$$

$$\mathbb{c} = (X^T X)^{-1} X^T \varrho$$

DISTRIBUTION APPROXIMATION – Cont'd.

→ Considerations...

13) ... Gradient estimation ...

(ii) Compute quadratic polynomial

$$Q_0(\mathbf{x}) = \sum_{\substack{i_1,\ldots,i_k \geq 0 \\ i_1+\ldots+i_k \leq 2}} c_{i_1,\ldots,i_k}\, x_1^{i_1} \cdots x_k^{i_k}$$

$$= \sum_{\|\mathbf{i}\| \leq 2} c_{\mathbf{i}}\, \mathbf{x}^{\mathbf{i}}$$ (multi-index notation)

$\Rightarrow$ Conditions: $Q_0(\mathbf{x}_j) = \varphi_j$, $j = 0 \ldots m$

$$Q_0(\mathbf{x}_0) = \varphi_0 \quad \ldots \quad Q_0(\mathbf{x}_m) = \varphi_m \iff \boxed{\begin{array}{c} \sum_{\|\mathbf{i}\| \leq 2} c_{\mathbf{i}}\, \mathbf{x}_0^{\mathbf{i}} = \varphi_0 \\ \vdots \\ \sum_{\|\mathbf{i}\| \leq 2} c_{\mathbf{i}}\, \mathbf{x}_m^{\mathbf{i}} = \varphi_m \end{array}}$$

$$\Rightarrow X\mathbf{c} = \boldsymbol{\varphi} \Rightarrow \mathbf{c} = (X^T X)^{-1} X^T \boldsymbol{\varphi}$$

$$\Rightarrow \nabla\varphi_0 = \nabla Q_0 \Big|_{\mathbf{x}=\mathbf{x}_0} = \left(\frac{\partial}{\partial x_1} Q_0, \ldots, \frac{\partial}{\partial x_k} Q_0 \right)_{\mathbf{x}=\mathbf{x}_0}$$

- **NOTES:**
  - It is simpler and more efficient to compute the LINEAR polynomial $L_0(\mathbf{x})$.
  - Depending on feature space dimension k and the degree of the local Taylor approximation, the value of m must be large enough to avoid an under-determined equation system.

DISTRIBUTION APPROXIMATION – Cont'd.

→ Considerations...

13) ... Gradient estimation ...

• **NOTES**: – The numbers of coefficients of the Local Taylor approximation are:

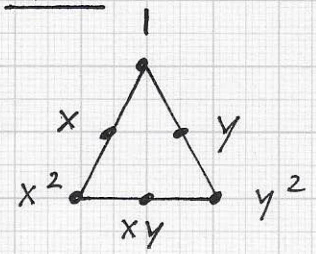
(i) Linear polynomial: $\mathbf{1+k}$

(ii) quadratic polynomial:

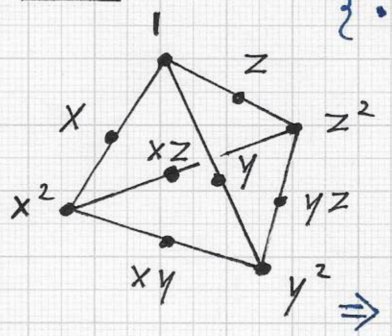
$$1+k+k+\binom{k}{2} = 1+2k+\frac{k!}{2(k-2)!} = 1+2k+\frac{1}{2}((k-1)k)$$

$$= 1+2k+\frac{1}{2}(k^2-k) = \mathbf{1+\frac{3}{2}k+\frac{1}{2}k^2}$$

k=2:



k=3:



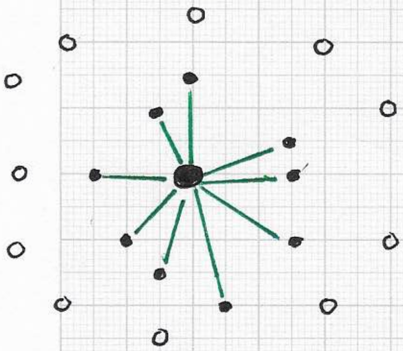
{ • Ex.: $k=3$

(i) linear poly: $1, x, y, z \Rightarrow 4$ coeffs.

(ii) quadratic poly: $1, x, y, z, x^2, y^2, z^2, xy, xz, yz \Rightarrow 10$ coeffs. }

⇒ These numbers define the numbers of sites $\mathbf{x}_i$ needed to avoid under-determined systems.

m = 9:



– The least squares computations of Local linear or quadratic Taylor approximations must be done efficiently. Thus, a k-D tree, for example, is a desirable data structure to use for all sites $\mathbf{x}_i$ to determine quickly the m nearest neighbors of a specific site.

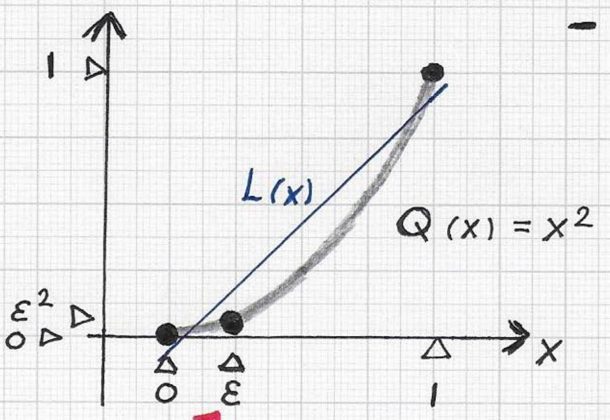
● specific site

• m closest sites

DISTRIBUTION APPROXIMATION - Cont'd.

→ Considerations ...

14) Linear vs. quadratic local Taylor approx. using least squares - Effect on gradient estimates. The effect is shown for the 1D case:



grad. estimate needed for $x = \varepsilon$

- Ex.: 3 data points in local neighborhood: $(0,0), (\varepsilon, \varepsilon^2), (1,1)$

⇒ Quadratic least squares polynomial Q(x) interpolates the 3 points with

⇒ $Q'(x) = 2x$

⇒ $Q'(\varepsilon) = 2\varepsilon$

⇒ least squares system for L(x):

- $L(x) = A + Bx$
- $\begin{pmatrix} 1 & 0 \\ 1 & \varepsilon \\ 1 & 1 \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} 0 \\ \varepsilon^2 \\ 1 \end{pmatrix}$

  $M \quad \mathbb{a} = \mathbb{f}$
- solve: $\mathbb{a} = (M^T M)^{-1} M^T \mathbb{f}$

  $\mathbb{a} = \ldots$

⇒ $L'(x) = B$ ⇒ $L'(\varepsilon) = B$

Consider $\varepsilon \to 0$:

$Q'(\varepsilon) \to 0$

$L'(\varepsilon) \to 1$

⇒ When data sites in k-dim. feature space are "highly non-uniformly distributed" and/or some sites are "very close to each other," linear and quadratic Taylor approx. can generate "very different grad. estimates."

BH