Feature Maps and Kernels

Machine Learning
CSx824/ECEx242
Bert Huang
Virginia Tech

Last Time

- SVM primal problem has a dual optimization
- Dual has box constraints on dual variables
- Dual only considers inner products of data vectors
- Kernel trick: replace inner products with kernel functions
    - Inner products in mapped feature space

Kernel SVM

$$\min_{\alpha} \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_i \alpha_i$$

$$\text{s.t.} \quad \sum_i \alpha_i y_i = 0, \quad \alpha_i \in [0, C]$$

$$f(x) = w^\top x + b = \sum_i \alpha_i y_i K(x_i, x) + b$$

$$b = y_i - \sum_j \alpha_j y_j K(x_i, x_j)$$

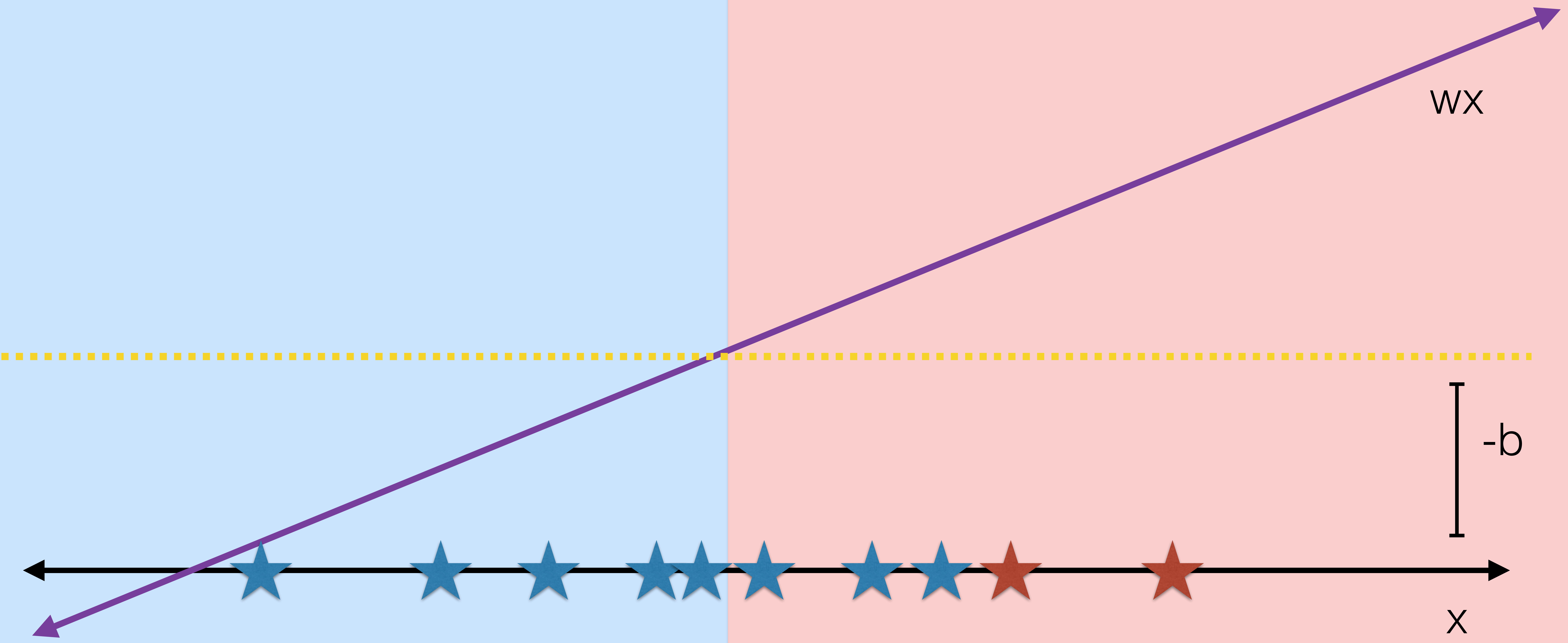
for examples *i* where $0 < \alpha_i < C$

K = kernel function

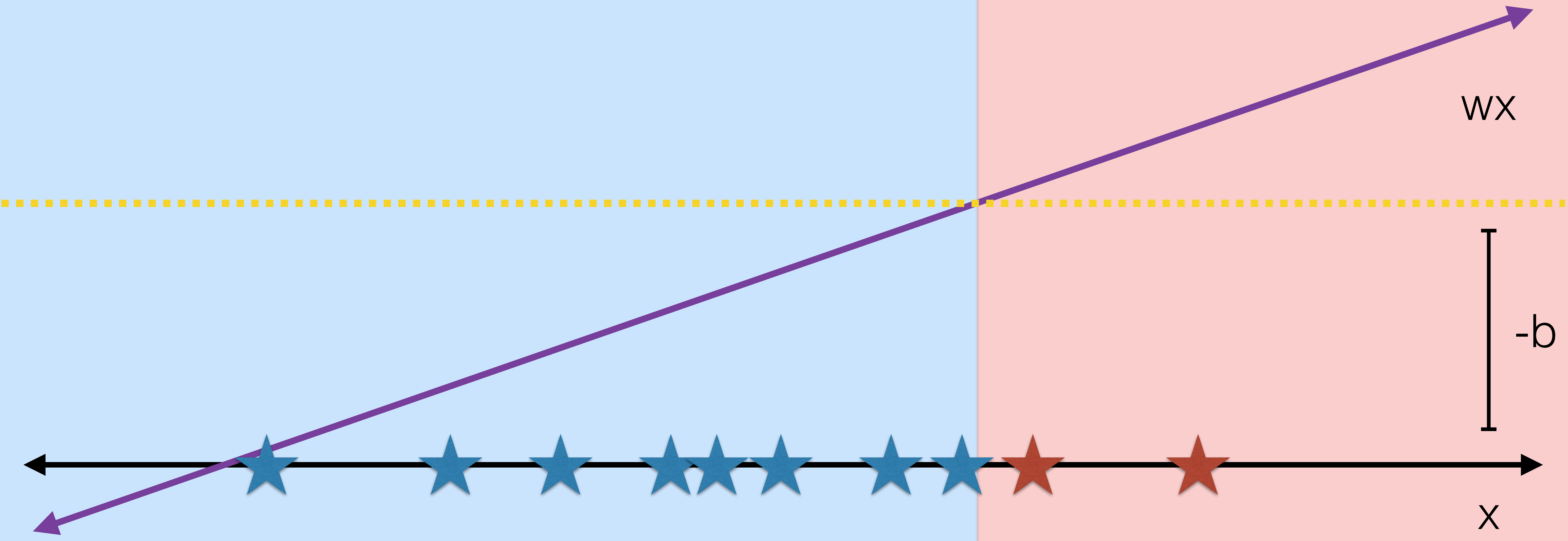
Outline

- Feature maps and nonlinearity
- Efficient kernel functions
  - Polynomial kernel
  - Gaussian radial basis function kernel (correction from last video)

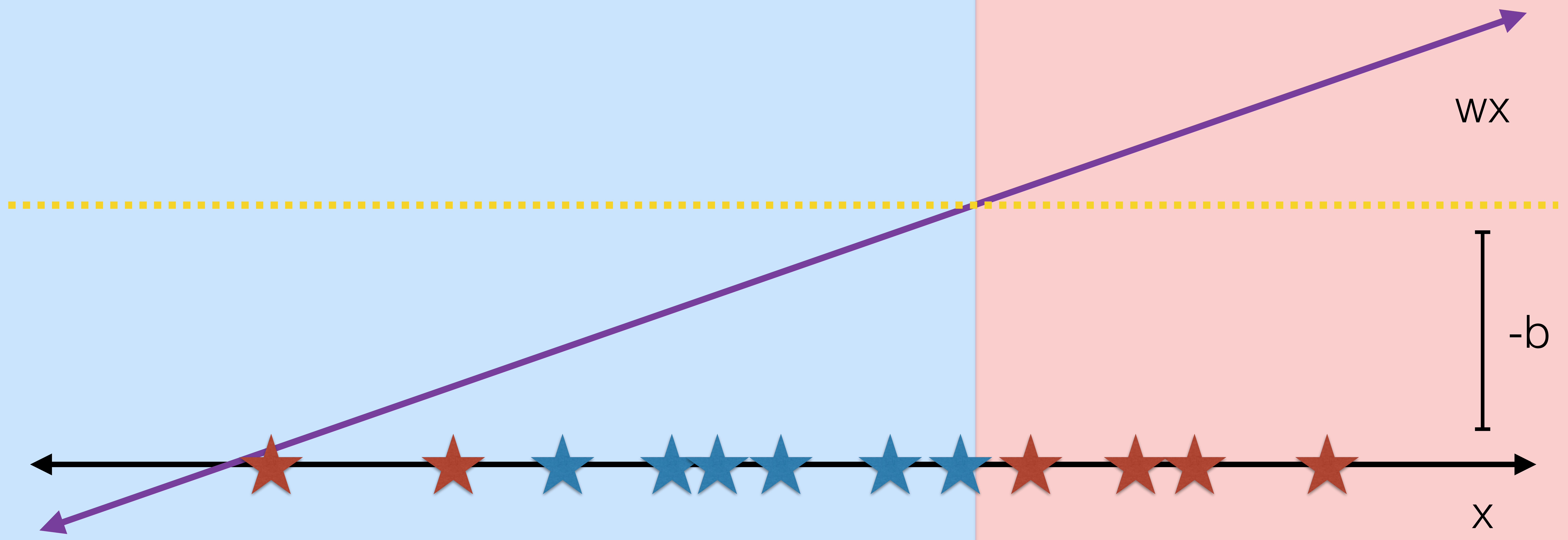
Nonlinear Decision Boundary



Nonlinear Decision Boundary



Nonlinear Decision Boundary



Nonlinear Decision Boundary

Polynomial Feature Map

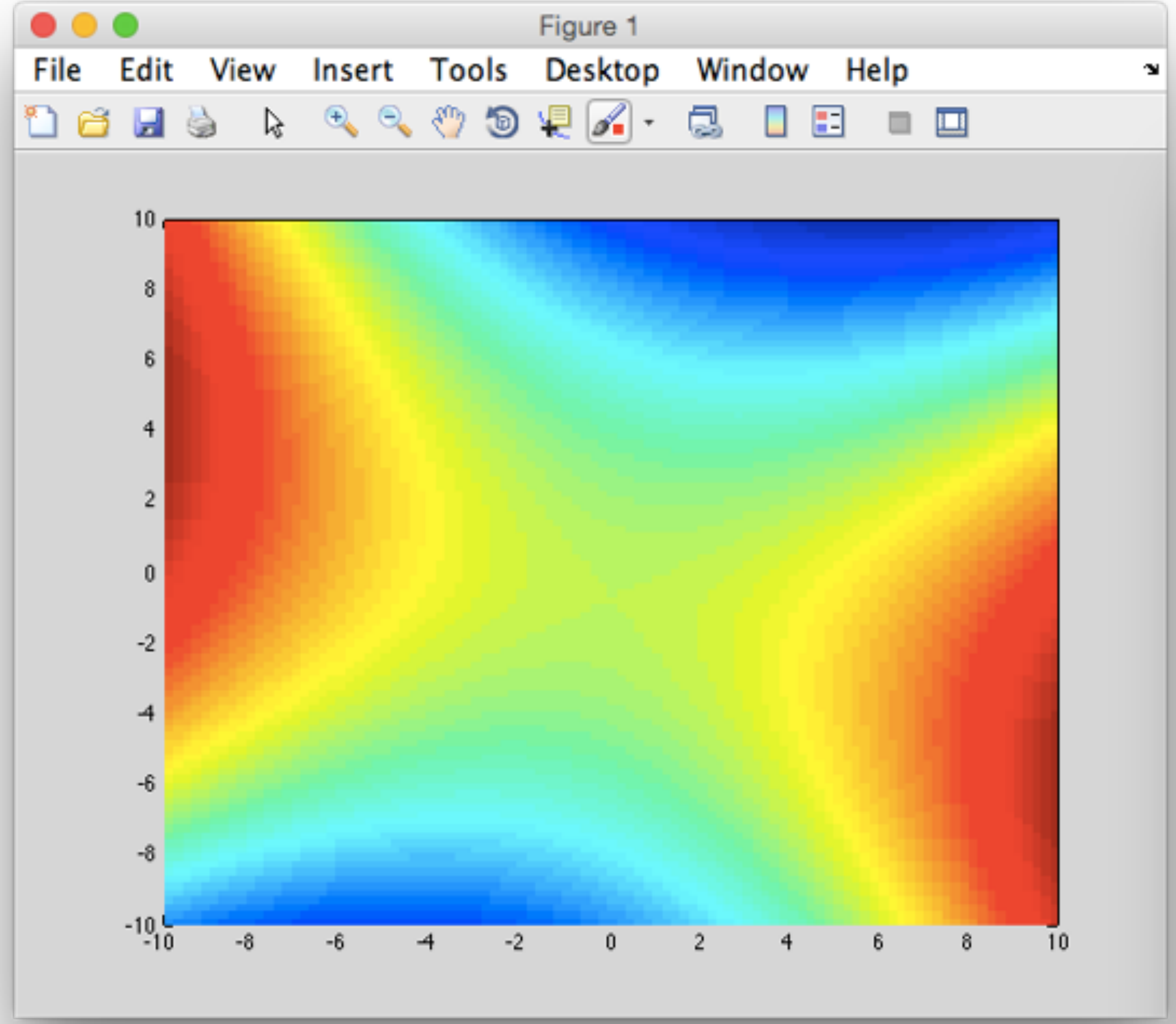
$$\Phi(x) = [x^1, \ldots, x^d, x^1 x^1, \ldots, x^1 x^d, \ldots, x^d x^1, \ldots, x^d x^d, \ldots]^\top$$

Third-order terms $\{x^1 x^1 x^1, x^1 x^1 x^2, \ldots, x^1 x^d x^d, \ldots, x^d x^d x^d\}$

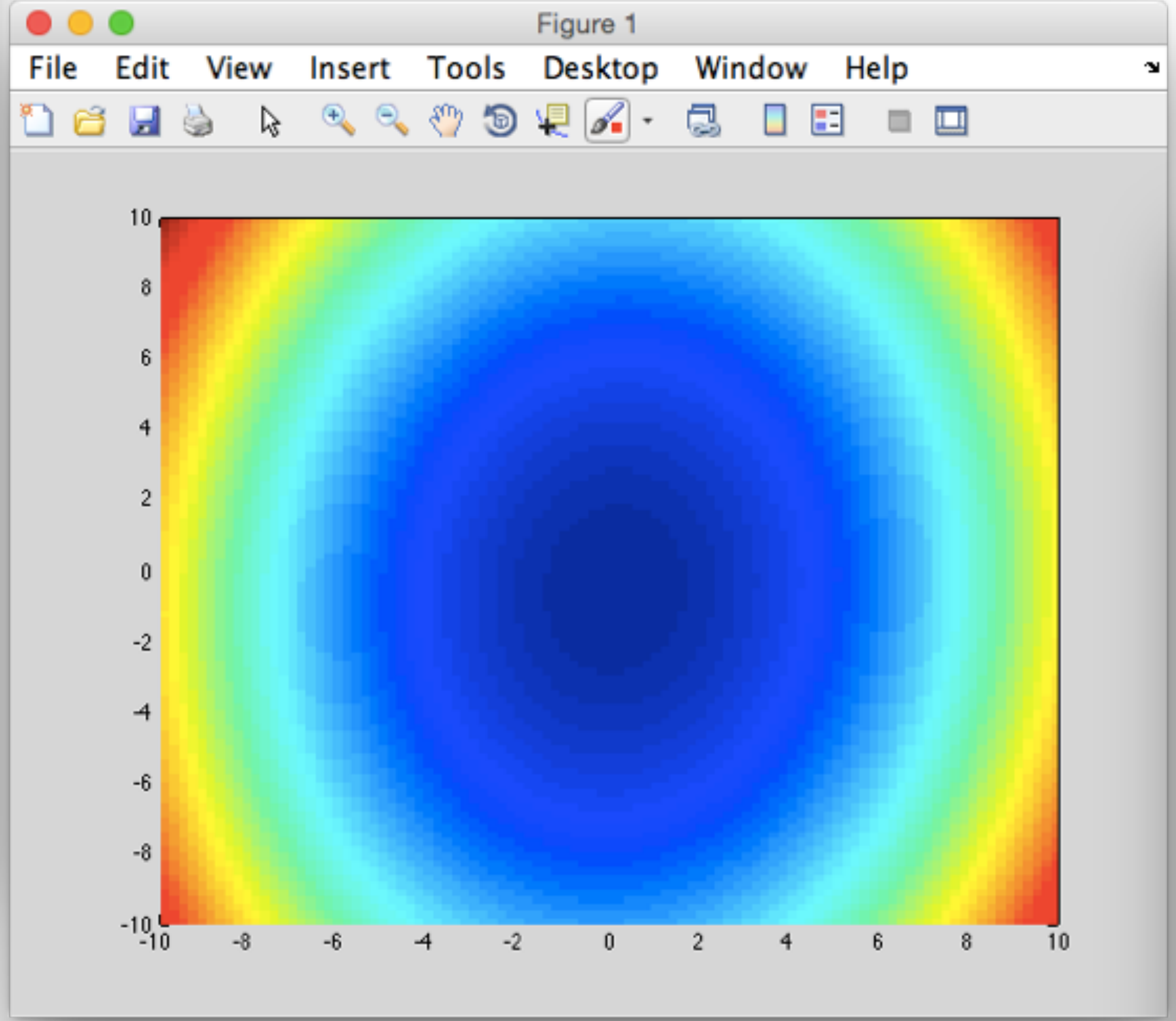
Fourth-order terms $\{x^i x^j x^k x^\ell | i, j, k, \ell \in \{1, \ldots, d\}\}$

$$|\Phi(x)| = \sum_{a=1}^{M} d^a = O(d^M)$$

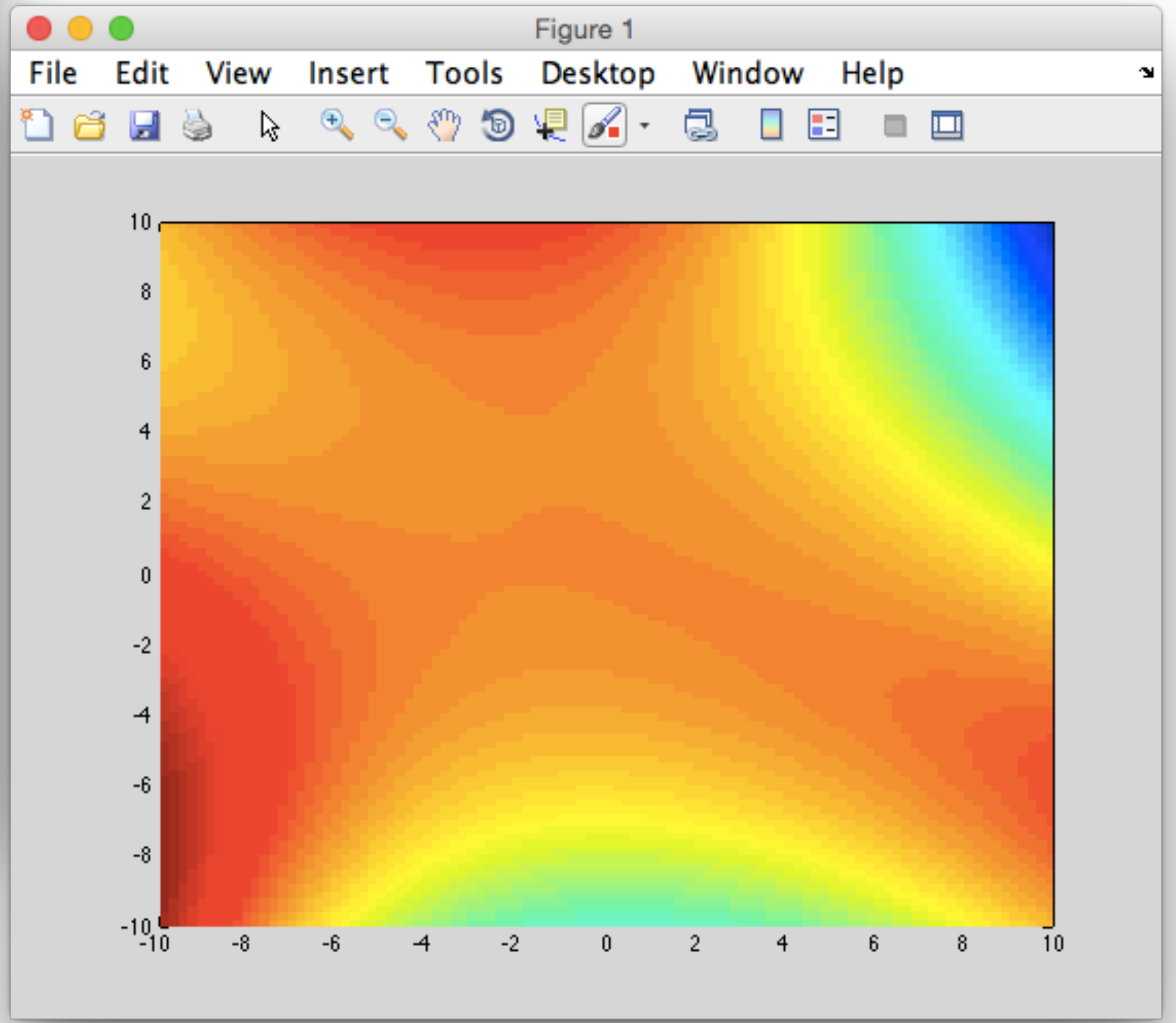
Polynomial Decision Scores



quadratic



quadratic



third order

Efficient Kernel Computation

$\Phi(x)$ $\quad$ $K(x_i, x_j) = \Phi(x_i)^\top \Phi(x_j)$

$$\Phi(x_i) = [x_i^1, \ldots, x_i^d, x_i^1 x_i^1, \ldots, x_i^1 x_i^d, \ldots, x_i^d x_i^1, \ldots, x_i^d x_i^d, \ldots]^\top$$

$$| \quad | \quad | \quad | \quad | \quad | \quad | \quad | \quad | \quad | \quad |$$

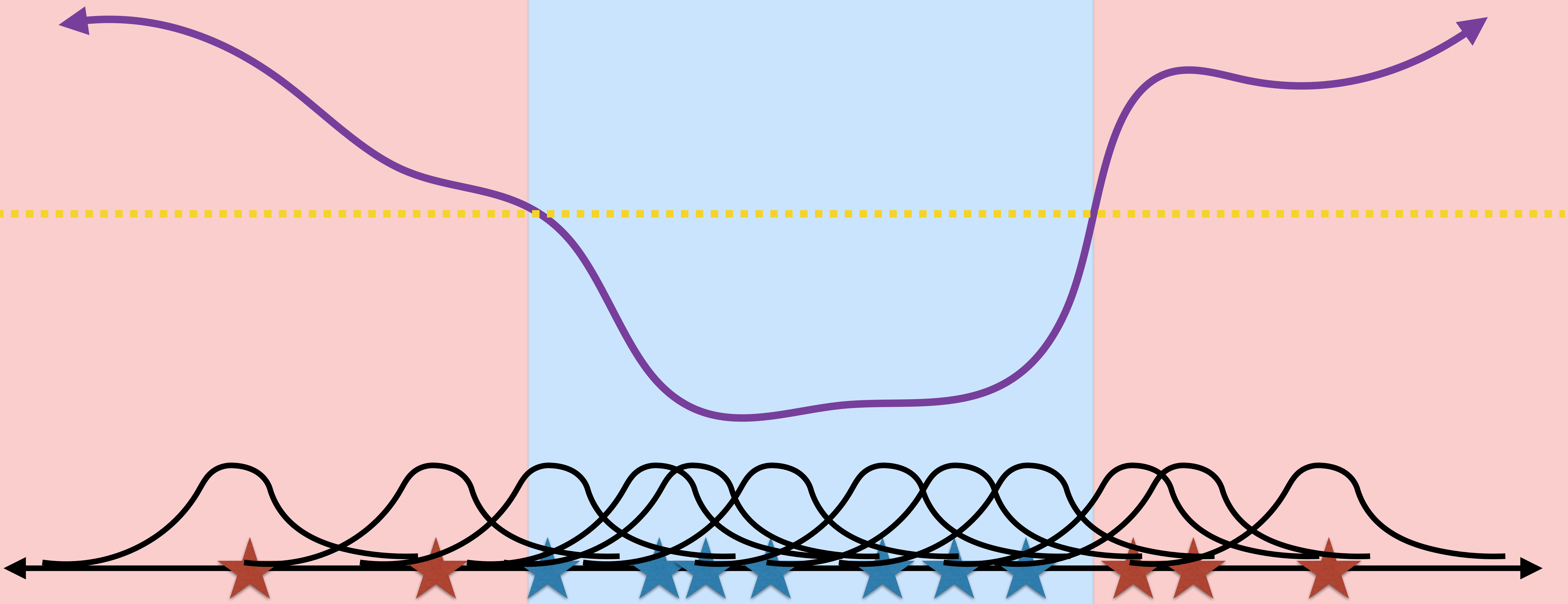
$$\Phi(x_j) = [x_j^1, \ldots, x_j^d, x_j^1 x_j^1, \ldots, x_j^1 x_j^d, \ldots, x_j^d x_j^1, \ldots, x_j^d x_j^d, \ldots]^\top$$

$(x_i^\top x_j + 1)^M$ $\quad$ $(x_i^\top x_j)(x_i^\top x_j) + 2x_i^\top x_j + 1$

$X \in \mathbb{R}^{d \times n}$ $\quad$ $K = (X^\top X + 1)^M$

elementwise exponentiation

Radial Basis Functions



RBF Kernel

$$K(x_i, x_j) = \exp\left(-\frac{1}{2\sigma^2}||x_i - x_j||^2\right)$$

(mistake from last video)

$$\Phi(x) = \left[\exp\left(\frac{1}{\sigma}||x - x_1||^2\right), \exp\left(\frac{1}{\sigma}||x - x_2||^2\right), \ldots, \exp\left(\frac{1}{\sigma}||x - x_n||^2\right)\right]$$

What is $\Phi(x)$?

Taylor Expansion of RBF Kernel

$$K(x_i, x_j) = \exp\left(-\frac{1}{2\sigma^2}||x_i - x_j||^2\right) \qquad \sigma = 1/\sqrt{2}$$

$$= \exp\left(-||x_i - x_j||^2\right)$$

$$= \exp\left(-x_i^\top x_i\right) \exp\left(-x_j^\top x_j\right) \exp(2x_i^\top x_j) \qquad \exp(z) = \sum_{n=0}^{\infty} \frac{z^n}{n!}$$

$$= \exp\left(-x_i^\top x_i\right) \exp\left(-x_j^\top x_j\right) \sum_{n=0}^{\infty} \frac{2^n (x_i^\top x_j)^n}{n!}$$

← order-n polynomial kernel* $\Phi^n(x)$

$$\Phi^{\text{rbf}} = \exp(-x^\top x)\left[\Phi^1(x)^\top, \Phi^2(x)^\top, \ldots, \Phi^\infty(x)^\top\right]^\top$$

Kernel Formulas

Linear

$$K(x_i, x_j) = x_i^\top x_j \qquad X_i \in \mathbb{R}^{d \times m} \qquad X_j \in \mathbb{R}^{d \times n}$$

$$K = X_i^\top X_j$$

Polynomial

$$K(x_i, x_j) = (x_i^\top x_j + 1)^M \qquad K = (X_i^\top X_j + 1)^M$$

RBF

$$K(x_i, x_j) = \exp\left(-\frac{1}{2\sigma^2}||x_i - x_j||^2\right)$$

$$K = \exp\left(-\frac{1}{2\sigma^2}\left(\text{diag}(X_i^\top X_i)\vec{1}^\top + \vec{1}\text{diag}(X_j^\top X_j)^\top - 2X_i^\top X_j\right)\right)$$

Kernels

- Map input data to new feature space (usually higher dimensional)
- Efficient method for computing inner product in mapped space
- Methods using inner products can directly use kernel
  - E.g., dual SVM

Next Time

- How to alleviate the computational cost of SVM training?
  - QP: roughly $O(n^3)$ for n constraints or variables