Exponential Families and Kernels

Lecture 3

Alexander J. Smola
Alex.Smola@nicta.com.au

Machine Learning Program
National ICT Australia
RSISE, The Australian National University

Outline

Exponential Families

- Maximum likelihood and Fisher information
- Priors (conjugate and normal)

Conditioning and Feature Spaces

- Conditional distributions and inner products
- Clifford Hammersley Decomposition

Applications

- Classification and novelty detection
- Regression

Applications

- Conditional random fields
- Intractable models and semidefinite approximations

Lecture 3

Novelty Detection

- Density estimation
- Thresholding and likelihood ratio

Classification

- Log partition function
- Optimization problem
- Examples
- Clustering and transduction

Regression

- Conditional normal distribution
- Estimating the covariance
- Heteroscedastic estimators

Density Estimation

Maximum a Posteriori

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{m} g(\theta) - \langle \phi(x_i), \theta \rangle + \frac{1}{2\sigma^2} \|\theta\|^2$$

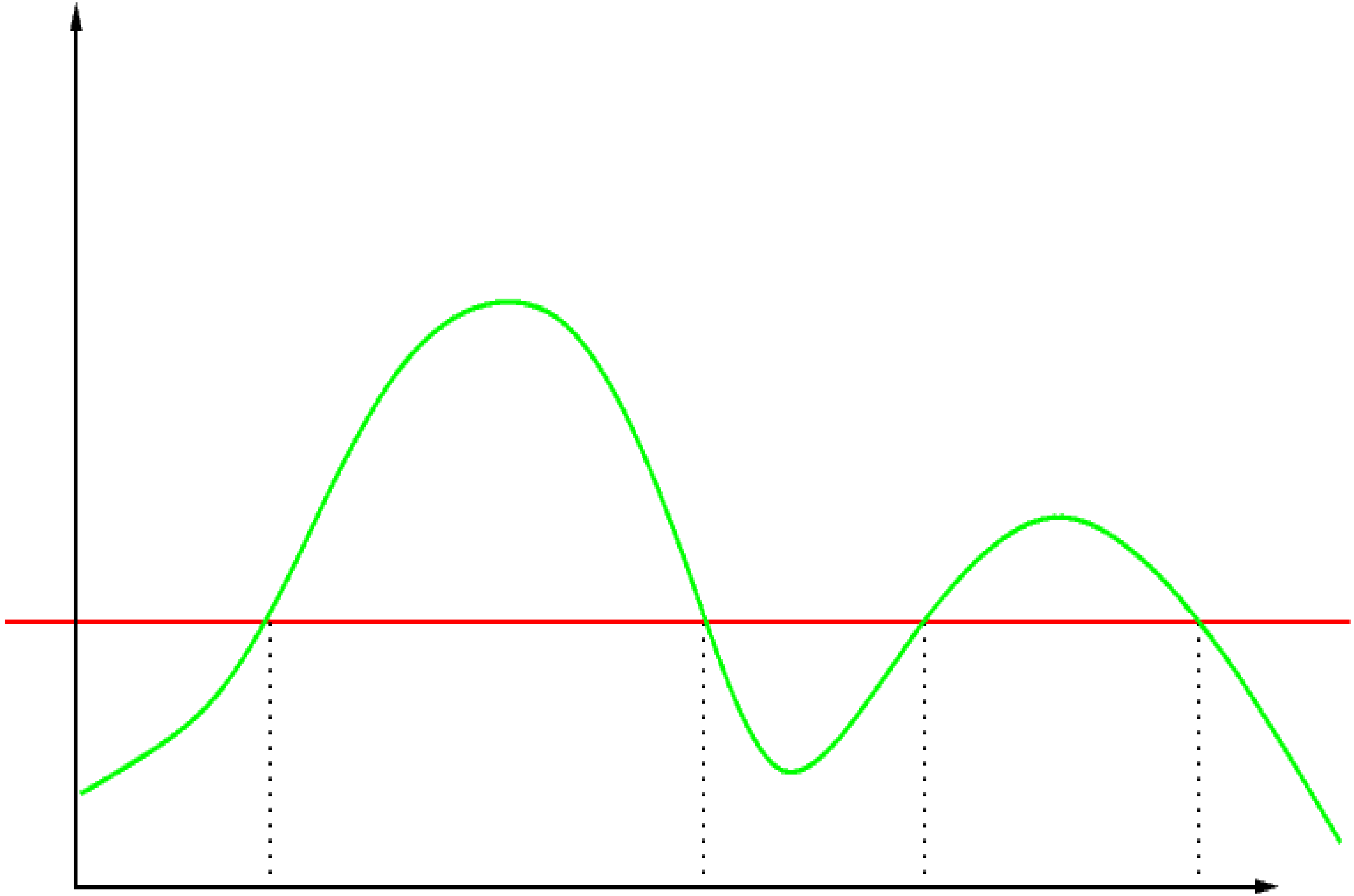
Advantages

- Convex optimization problem
- Concentration of measure

Problems

- Normalization $g(\theta)$ may be painful to compute
- For density estimation we need no normalized $p(x|\theta)$
- No need to perform particularly well in high density regions

Novelty Detection



Novelty Detection

Optimization Problem

MAP $$\sum_{i=1}^{m} -\log p(x_i|\theta) + \frac{1}{2\sigma^2}\|\theta\|^2$$

Novelty $$\sum_{i=1}^{m} \max\left(-\log \frac{p(x_i|\theta)}{\exp(\rho - g(\theta))}, 0\right) + \frac{1}{2}\|\theta\|^2$$

$$\sum_{i=1}^{m} \max(\rho - \langle \phi(x_i), \theta \rangle, 0) + \frac{1}{2}\|\theta\|^2$$

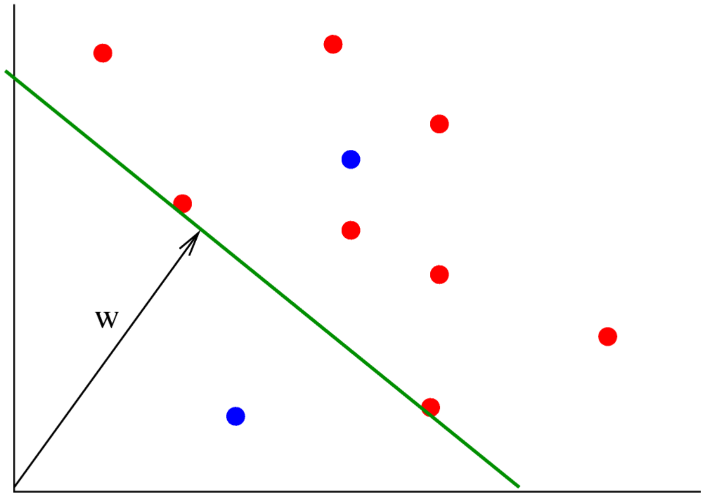
Advantages

- No normalization $g(\theta)$ needed
- No need to perform particularly well in high density regions (estimator focuses on low-density regions)
- Quadratic program

Geometric Interpretation

Idea

Find hyperplane that has **maximum distance from origin**, yet is still closer to the origin than the observations.



Hard Margin

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|\theta\|^2 \\ \text{subject to} \quad & \langle \theta, x_i \rangle \geq 1 \end{aligned}$$

Soft Margin

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|\theta\|^2 + C \sum_{i=1}^{m} \xi_i \\ \text{subject to} \quad & \langle \theta, x_i \rangle \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned}$$

Dual Optimization Problem

Primal Problem

minimize $\frac{1}{2}\|\theta\|^2 + C\sum_{i=1}^{m}\xi_i$

subject to $\langle\theta, x_i\rangle - 1 + \xi_i \geq 0$ and $\xi_i \geq 0$

Lagrange Function

We construct a **Lagrange Function** L by subtracting the constraints, multiplied by **Lagrange multipliers** (α_i and η_i), from the **Primal Objective Function**.

L has a **saddlepoint** at the optimal solution.

$$L = \frac{1}{2}\|\theta\|^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i\left(\langle\theta, x_i\rangle - 1 + \xi_i\right) - \sum_{i=1}^{m}\eta_i\xi_i$$

where $\alpha_i, \eta_i \geq 0$. For instance, if $\xi_i < 0$ we could increase L without bound via η_i.

Dual Problem, Part II

Optimality Conditions

$$\begin{aligned}\partial_\theta L &= \theta - \sum_{i=1}^m \alpha_i x_i = 0 \implies \theta = \sum_{i=1}^m \alpha_i x_i \\ \partial_{\xi_i} L &= C - \alpha_i - \eta_i = 0 \implies \alpha_i \in [0, C]\end{aligned}$$

Now we **substitute** the two optimality conditions **back into** L and eliminate the **primal variables**.

Dual Problem

$$\begin{aligned}\text{minimize} \quad & \frac{1}{2}\sum_{i=1}^m \alpha_i \alpha_j \langle x_i, x_j \rangle - \sum_{i=1}^m \alpha_i \\ \text{subject to} \quad & \alpha_i \in [0, C]\end{aligned}$$

Convexity ensures uniqueness of the optimum.

The ν-Trick

Problem

Depending on how we choose C, the number of points selected as lying on the "wrong" side of the hyperplane $H := \{x | \langle \theta, x \rangle = 1\}$ will vary.

- We would like to **specify a certain fraction** ν beforehand.
- We want to make the setting more adaptive to the data.

Solution

Use adaptive hyperplane that separates data from the origin, i.e. find

$$H := \{x | \langle \theta, x \rangle = \rho\},$$

where the threshold ρ is **adaptive**.

The ν-Trick

Primal Problem

$$\text{minimize} \quad \frac{1}{2}\|\theta\|^2 + \sum_{i=1}^{m} \xi_i - m\nu\rho$$

$$\text{subject to} \quad \langle \theta, x_i \rangle - \rho + \xi_i \geq 0 \text{ and } \xi_i \geq 0$$

Dual Problem

$$\text{minimize} \quad \frac{1}{2}\sum_{i=1}^{m} \alpha_i \alpha_j \langle x_i, x_j \rangle$$

$$\text{subject to} \quad \alpha_i \in [0, 1] \text{ and } \sum_{i=1}^{m} \alpha_i = \nu m.$$

Difference to before

The $\sum_i \alpha_i$ term vanishes from the objective function but we get one more constraint, namely $\sum_i \alpha_i = \nu m$.

The ν-Property

Optimization Problem

$$\text{minimize} \quad \frac{1}{2}\|\theta\|^2 + \sum_{i=1}^{m} \xi_i - m\nu\rho$$

$$\text{subject to} \quad \langle \theta, x_i \rangle - \rho + \xi_i \geq 0 \text{ and } \xi_i \geq 0$$

Theorem

- At most a fraction of ν points will lie on the "wrong" side of the margin, i.e., $y_i f(x_i) < 1$.
- At most a fraction of $1 - \nu$ points will lie on the "right" side of the margin, i.e., $y_i f(x_i) > 1$.
- In the limit, those fractions will become exact.

Proof Idea

At optimum, shift ρ slightly: only the active constraints will have an influence on the objective function.

Classification

Maximum a Posteriori Estimation

$$-\log p(\theta|X,Y) = \sum_{i=1}^{m} -\langle \phi(x_i, y_i), \theta \rangle + g(\theta|x_i) + \frac{1}{2\sigma^2}\|\theta\|^2 + c$$

Domain

- Finite set of observations $\mathcal{Y} = \{1, \ldots, m\}$
- Log-partition function $g(\theta|x)$ easy to compute.
- Optional centering

$$\phi(x, y) \to \phi(x, y) + c$$

leaves $p(y|x, \theta)$ unchanged (offsets both terms).

Gaussian Process Connection

Inner product $t(x, y) = \langle \phi(x, y), \theta \rangle$ is drawn from Gaussian process, so same setting as in literature.

Classification

Sufficient Statistic

We pick $\phi(x, y) = \phi(x) \otimes e_y$, that is

$$k((x, y), (x', y')) = k(x, x')\delta_{yy'} \text{ where } y, y' \in \{1, \ldots, n\}$$

Kernel Expansion

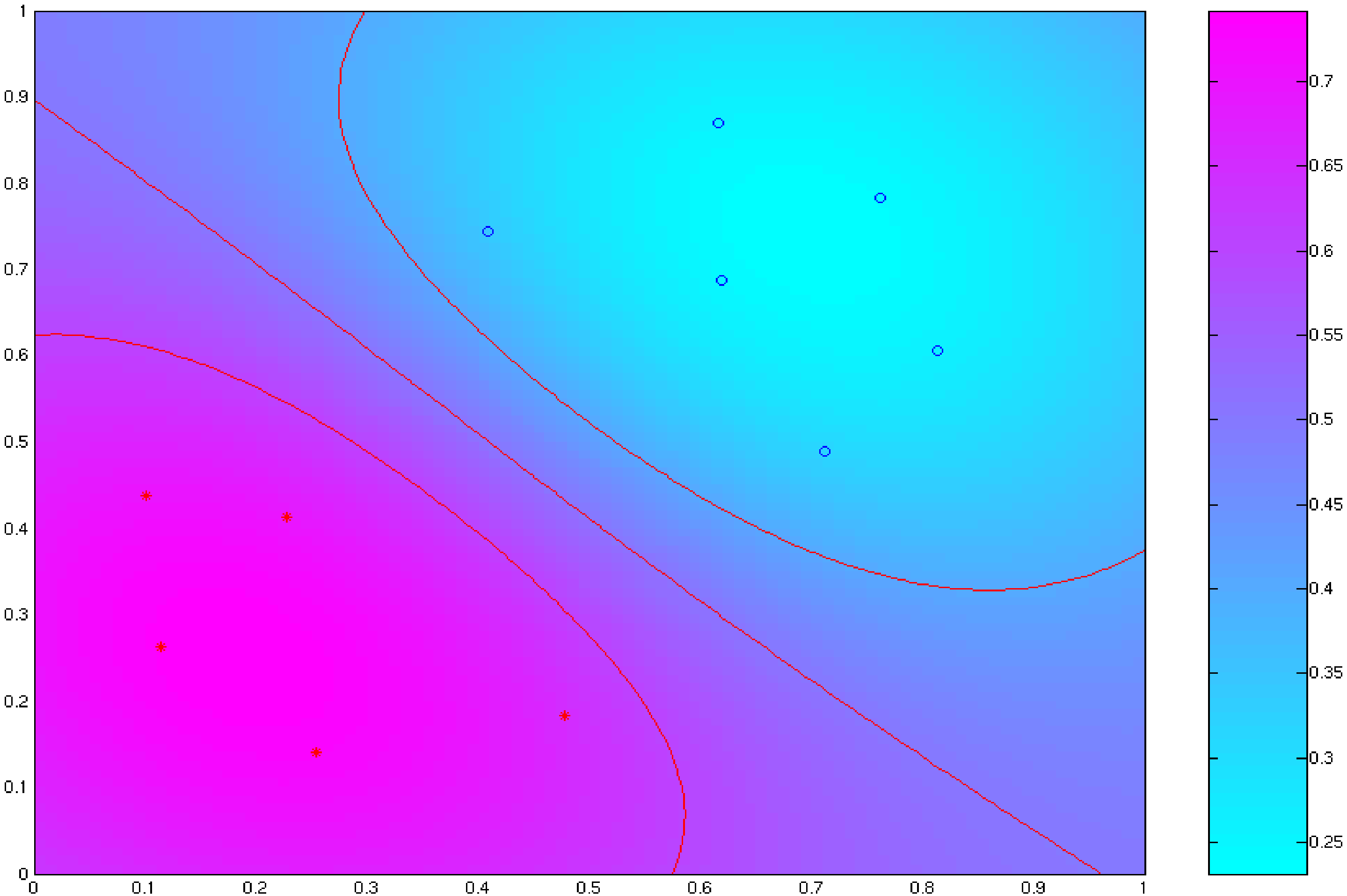
By the representer theorem we get that

$$\theta = \sum_{i=1}^{m} \sum_{y} \alpha_{iy} \phi(x_i, y)$$

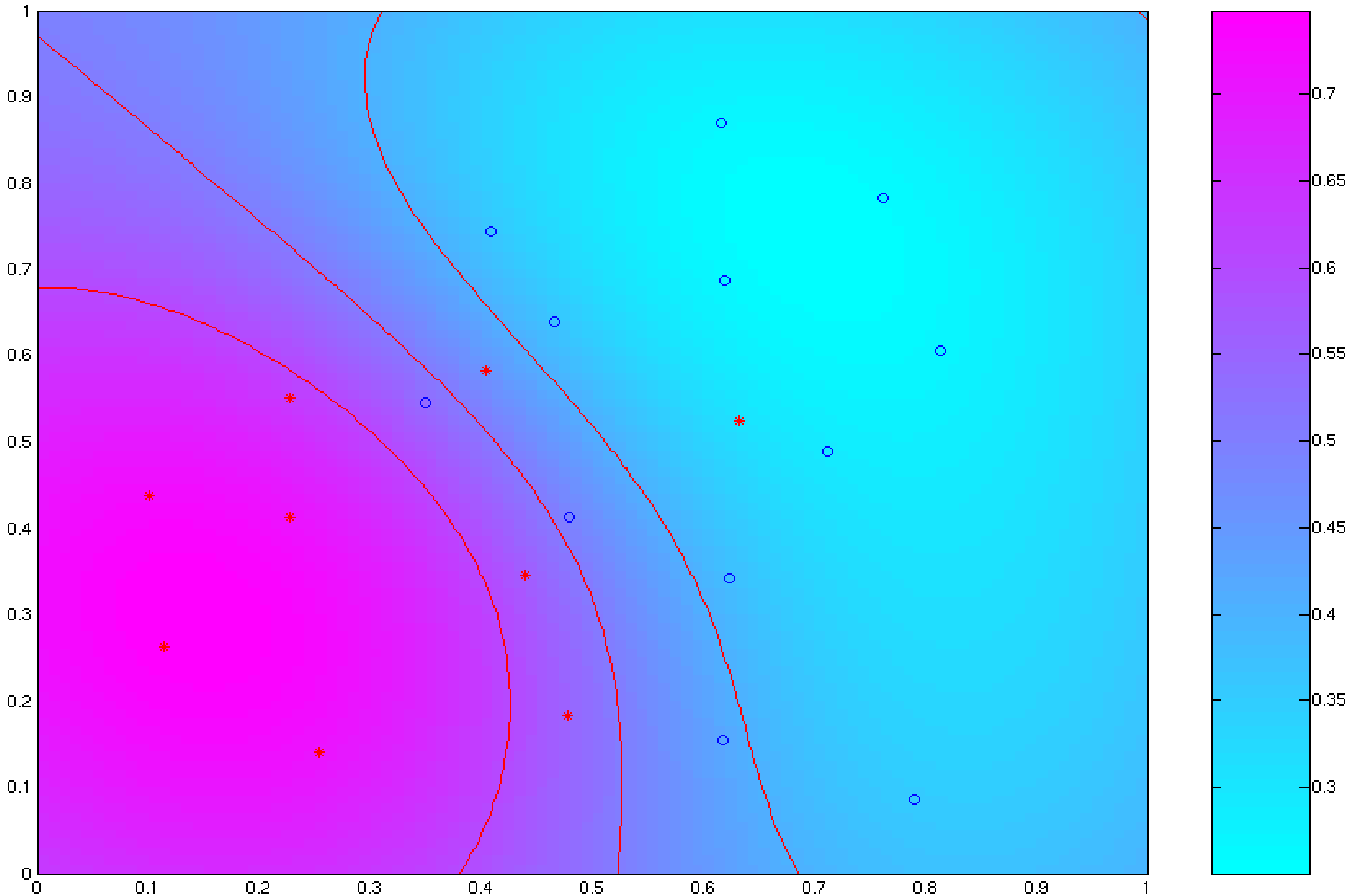
Optimization Problem

- Big mess . . . but convex.
- Solve by Newton or Block-Jacobi method.

A Toy Example



Noisy Data



1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3

SVM Connection

Problems with GP Classification

- Optimize even where classification is good
- Only sign of classification needed
- Only “strongest” wrong class matters
- Want to classify with a margin

Optimization Problem

MAP $$\sum_{i=1}^{m} -\log p(y_i|x_i,\theta) + \frac{1}{2\sigma^2}\|\theta\|^2$$

SVM $$\sum_{i=1}^{m} \max\left(\rho - \log \frac{p(y_i|x_i,\theta)}{\max_{y\neq y_i} p(y|x_i,\theta)}, 0\right) + \frac{1}{2}\|\theta\|^2$$

$$\sum_{i=1}^{m} \max(\rho - \langle \phi(x_i,y_i),\theta\rangle + \max_{y\neq y_i}\langle \phi(x_i,y),\theta\rangle, 0) + \frac{1}{2}\|\theta\|^2$$

Binary Classification

Sufficient Statistics

- Offset in $\phi(x, y)$ can be arbitrary
- Pick such that $\phi(x, y) = y\phi(x)$ where $y \in \{\pm 1\}$.
- Kernel matrix becomes

$$K_{ij} = k((x_i, y_i), (x_j, y_j)) = y_i y_j k(x_i, x_j)$$

Optimization Problem

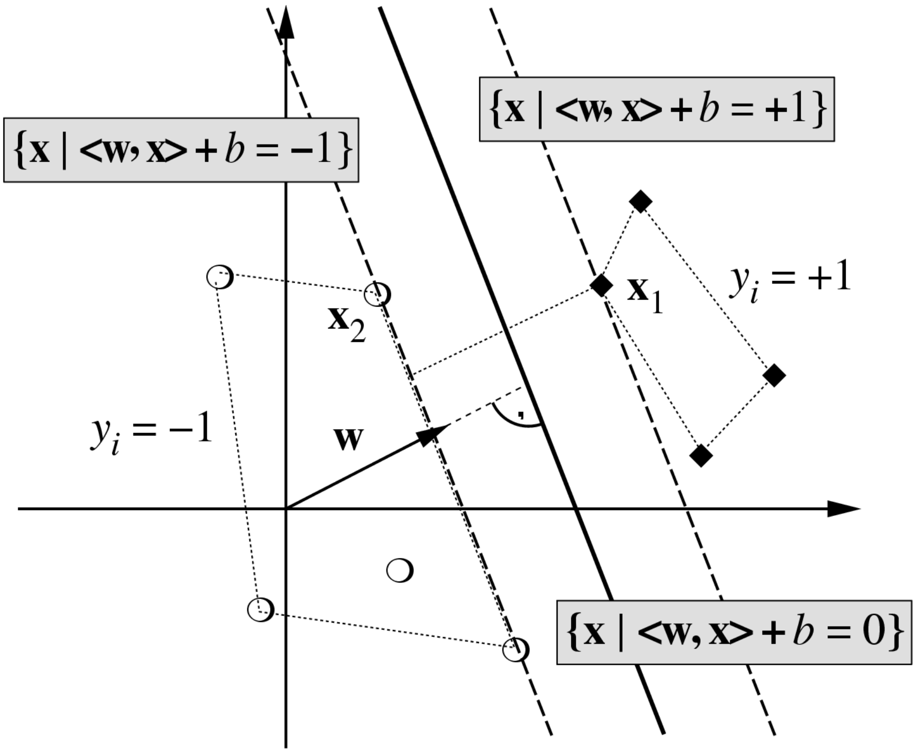
- The max over other classes becomes

$$\max_{y \neq y_i} \langle \phi(x_i, y), \theta \rangle = -y \langle \phi(x_i), \theta \rangle$$

- Overall problem

$$\sum_{i=1}^{m} \max(\rho - 2y_i \langle \phi(x_i), \theta \rangle, 0) + \frac{1}{2} \|\theta\|^2$$

Geometrical Interpretation



Note:

$$\langle \mathbf{w}, \mathbf{x}_1 \rangle + b = +1$$
$$\langle \mathbf{w}, \mathbf{x}_2 \rangle + b = -1$$

$$\Rightarrow \quad \langle \mathbf{w}, (\mathbf{x}_1 - \mathbf{x}_2) \rangle = 2$$

$$\Rightarrow \quad \left\langle \frac{\mathbf{w}}{\|\mathbf{w}\|}, (\mathbf{x}_1 - \mathbf{x}_2) \right\rangle = \frac{2}{\|\mathbf{w}\|}$$

Minimize $\frac{1}{2}\|\theta\|^2$ subject to $y_i(\langle \theta, x_i \rangle + b) \geq 1$ for all i.

Optimization Problem

Linear Function

$$f(x) = \langle \theta, x \rangle + b$$

Mathematical Programming Setting

If we require error-free classification with a margin, i.e., $yf(x) \geq 1$, we obtain:

$$\text{minimize} \quad \frac{1}{2}\|\theta\|^2$$
$$\text{subject to} \quad y_i(\langle \theta, x_i \rangle + b) - 1 \geq 0 \text{ for all } 1 \leq i \leq m$$

Result

The dual of the optimization problem is a simple quadratic program (more later ...).

Connection back to conditional probabilities

Offset b takes care of bias towards one of the classes.

Regression

Maximum a Posteriori Estimation

$$-\log p(\theta|X,Y) = \sum_{i=1}^{m} -\langle \phi(x_i, y_i), \theta\rangle + g(\theta|x_i) + \frac{1}{2\sigma^2}\|\theta\|^2 + c$$

Domain

- Continuous domain of observations $\mathcal{Y} = \mathbb{R}$
- Log-partition function $g(\theta|x)$ easy to compute **in closed form** as normal distribution.

Gaussian Process Connection

Inner product $t(x,y) = \langle \phi(x,y), \theta$ is drawn from Gaussian process. In particular also rescaled mean and covariance.

Regression

Sufficient Statistic (Standard Model)

We pick $\phi(x, y) = (y\phi(x), y^2)$, that is

$$k((x, y), (x', y')) = k(x, x')yy' + y^2 y'^2 \text{ where } y, y' \in \mathbb{R}$$

Traditionally the variance is fixed, that is $\theta_2 = \text{const.}$.

Sufficient Statistic (Fancy Model)

We pick $\phi(x, y) = (y\phi_1(x), y^2\phi_2(x))$, that is

$$k((x, y), (x', y')) = k_1(x, x')yy' + k_2(x, x')y^2 y'^2 \text{ where } y, y' \in \mathbb{R}$$

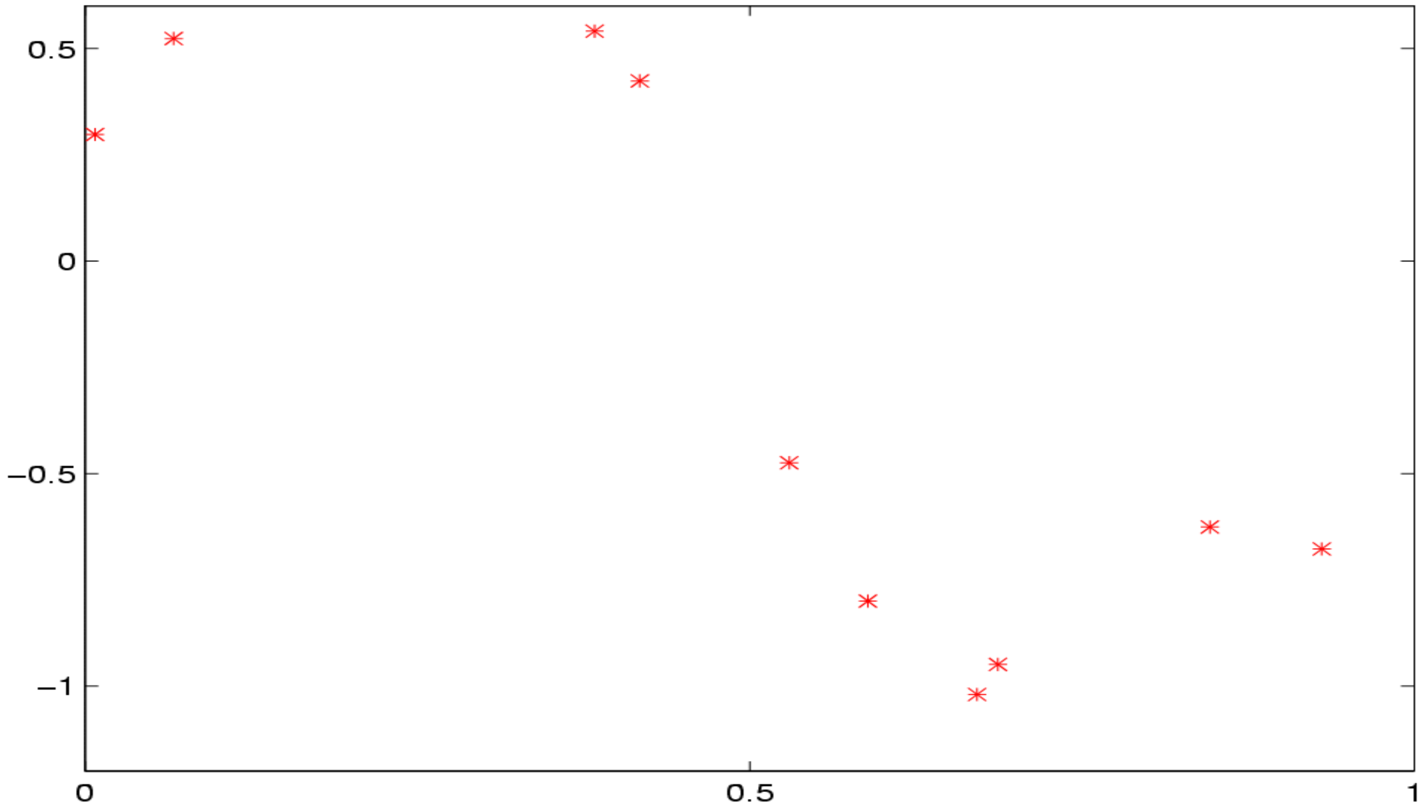
We estimate mean and variance **simultaneously**.

Kernel Expansion

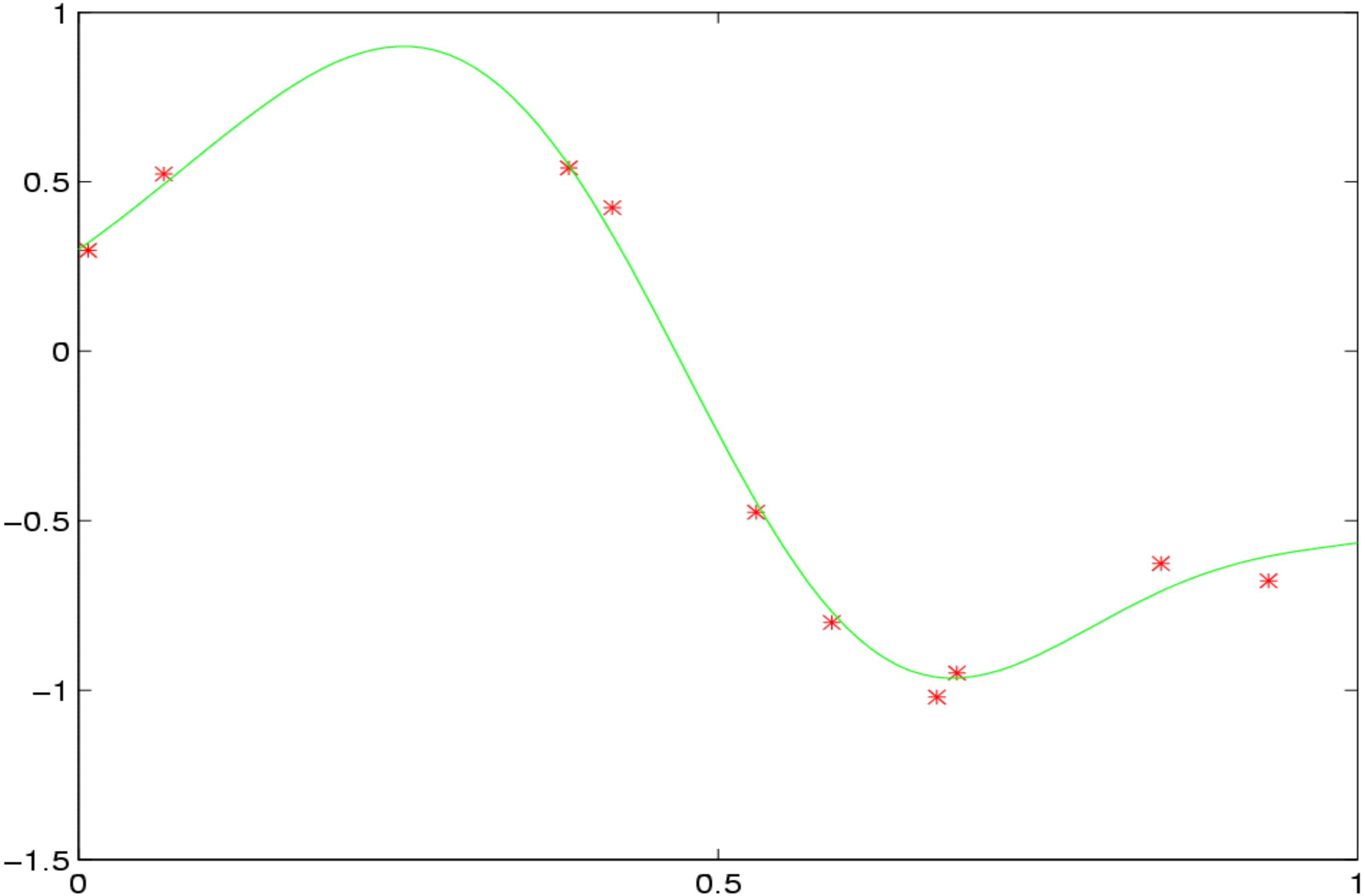
By the representer theorem (and more algebra) we get

$$\theta = \left(\sum_{i=1}^{m} \alpha_{i1}\phi_1(x_i), \sum_{i=1}^{m} \alpha_{i2}\phi_2(x_i) \right)$$

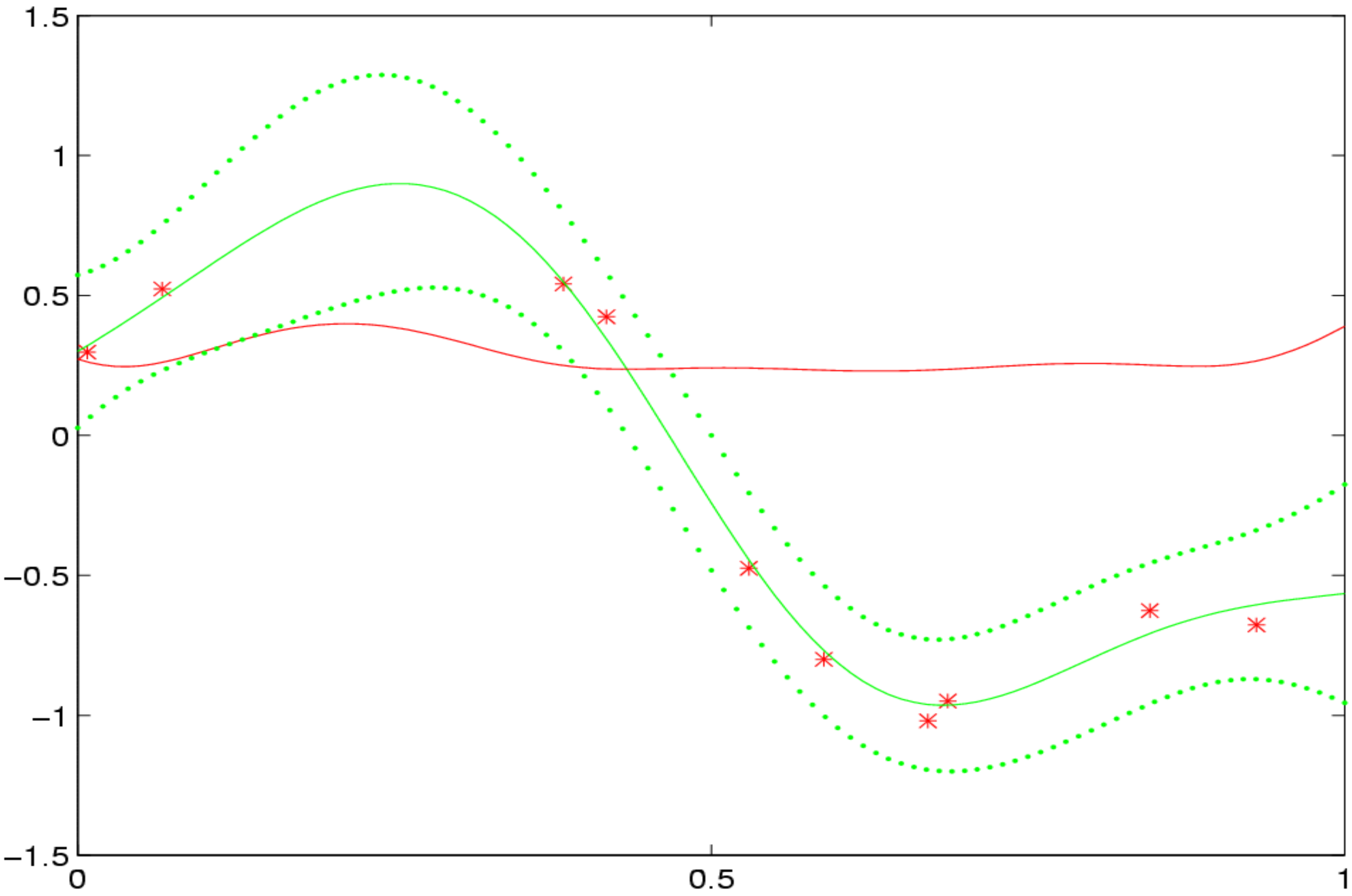
Training Data



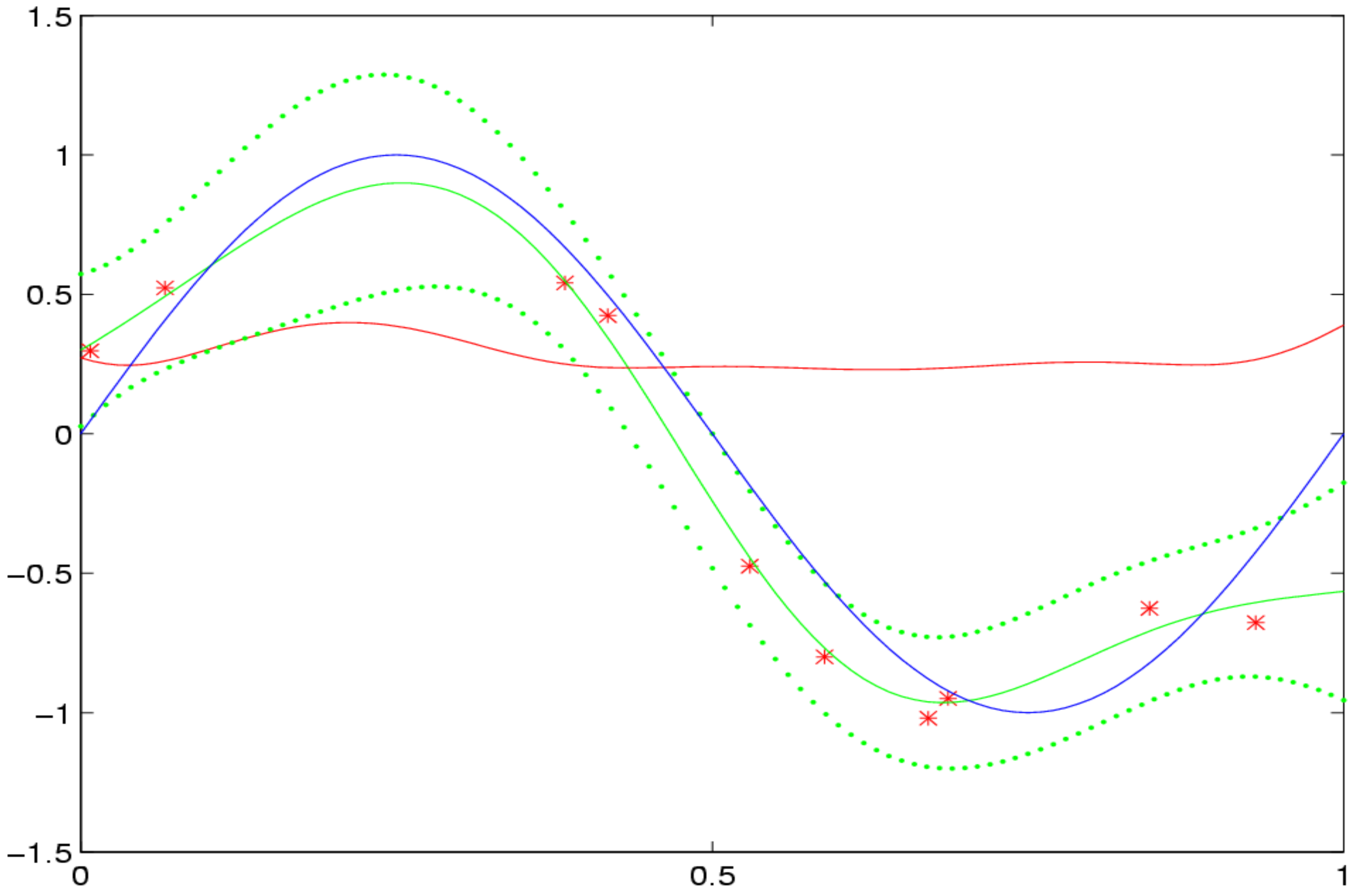
Mean $\vec{k}^\top(x)(K + \sigma^2\mathbf{1})^{-1}y$



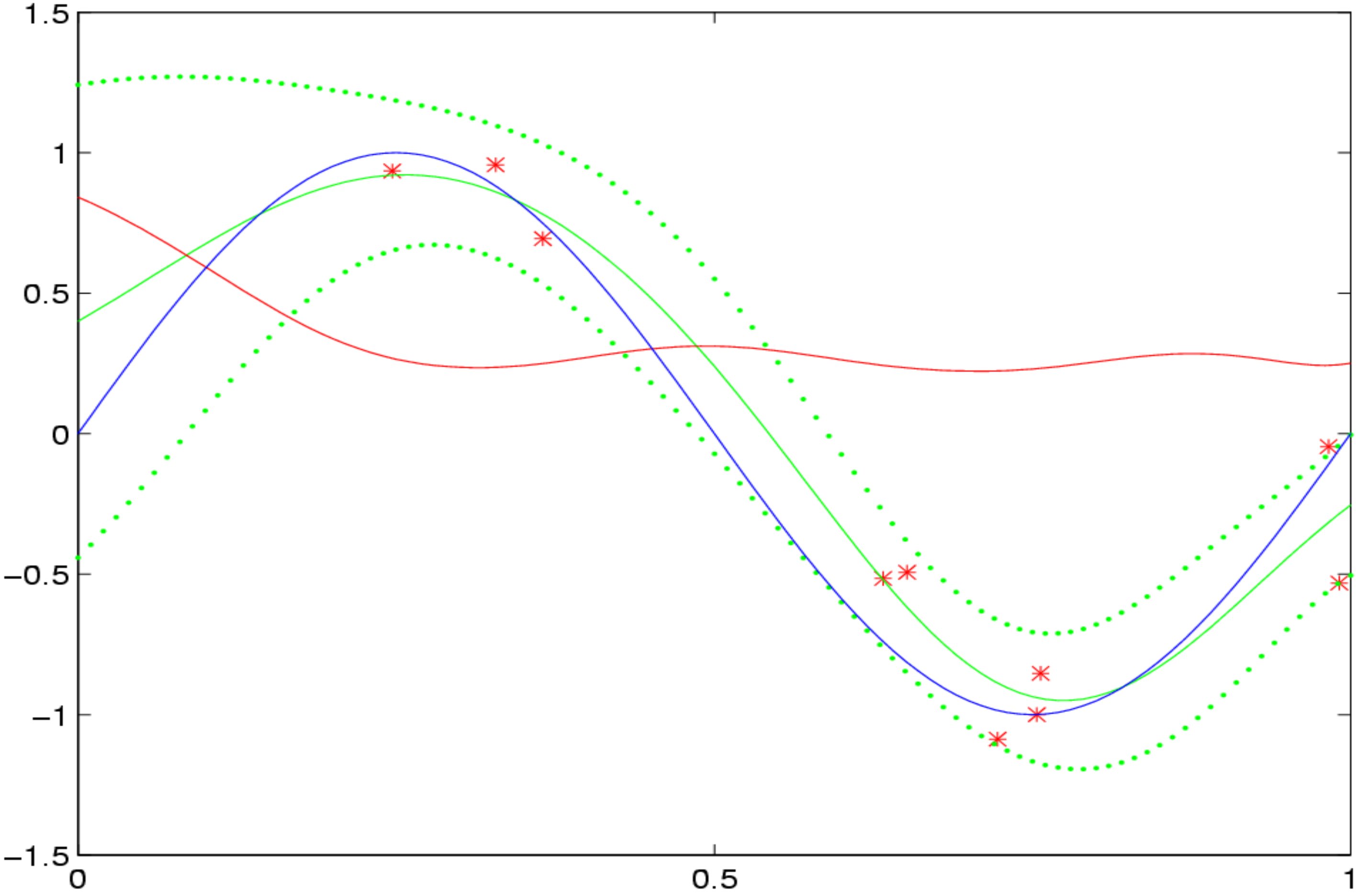
Variance $k(x,x) + \sigma^2 - \vec{k}^\top(x)(K + \sigma^2\mathbf{1})^{-1}\vec{k}(x)$



Putting everything together . . .



Another Example



Adaptive Variance Method

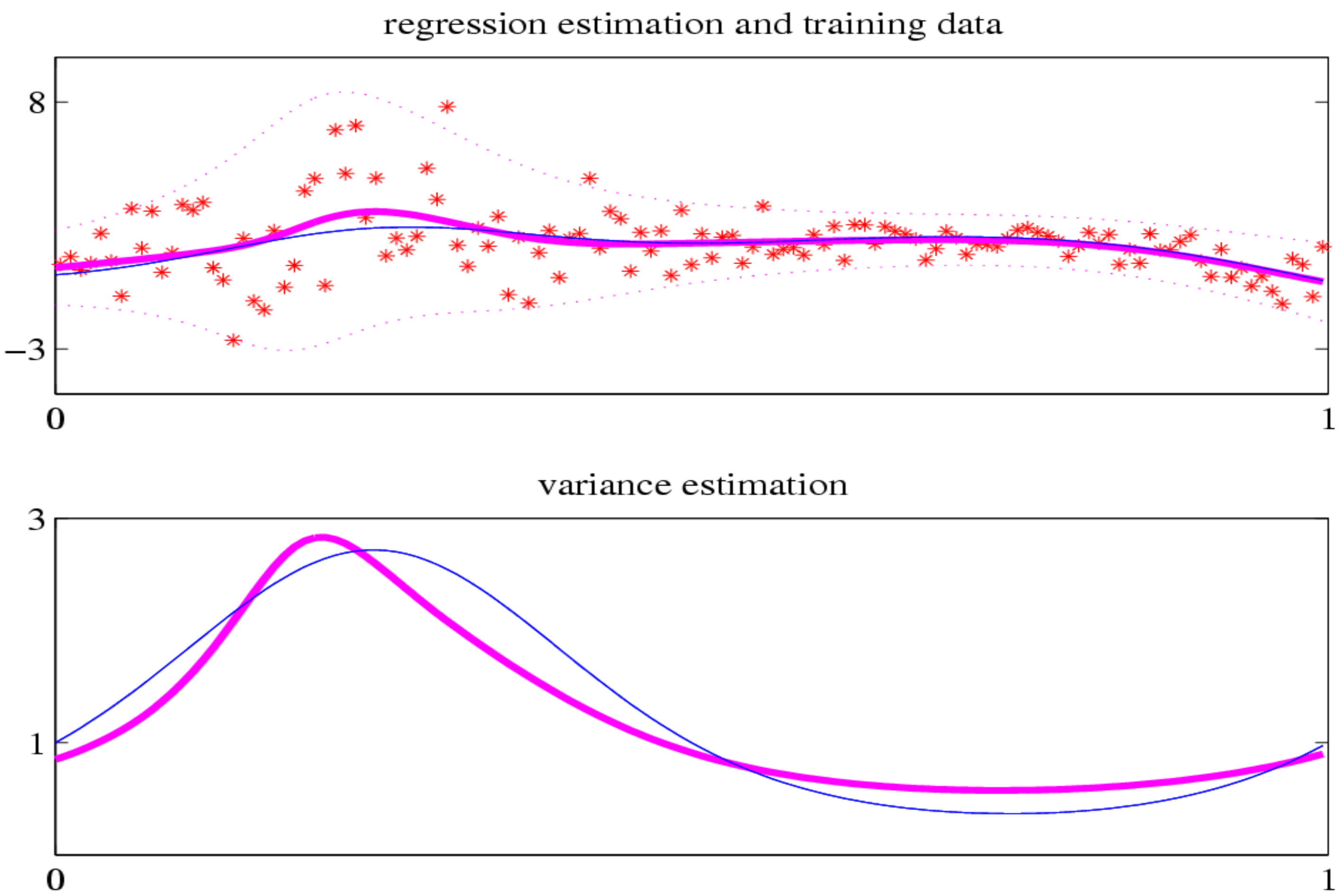
Optimization Problem:

$$\text{minimize} \sum_{i=1}^{m} \left[-\frac{1}{4} \left[\sum_{j=1}^{m} \alpha_{1j} k_1(x_i, x_j)\right]^\top \left[\sum_{j=1}^{m} \alpha_{2j} k_2(x_i, x_j)\right]^{-1} \left[\sum_{j=1}^{m} \alpha_{1j} k_1(x_i, x_j)\right] \right.$$
$$\left. -\frac{1}{2} \log\det -2 \left[\sum_{j=1}^{m} \alpha_{2j} k_2(x_i, x_j)\right] - \sum_{j=1}^{m} \left[y_i^\top \alpha_{1j} k_1(x_i, x_j) + (y_j^\top \alpha_{2j} y_j) k_2(x_i, x_j)\right] \right]$$
$$+ \frac{1}{2\sigma^2} \sum_{i,j} \alpha_{1i}^\top \alpha_{1j} k_1(x_i, x_j) + \operatorname{tr} \left[\alpha_{2i} \alpha_{2j}^\top\right] k_2(x_i, x_j).$$
$$\text{subject to } 0 \succ \sum_{i=1}^{m} \alpha_{2i} k(x_i, x_j)$$

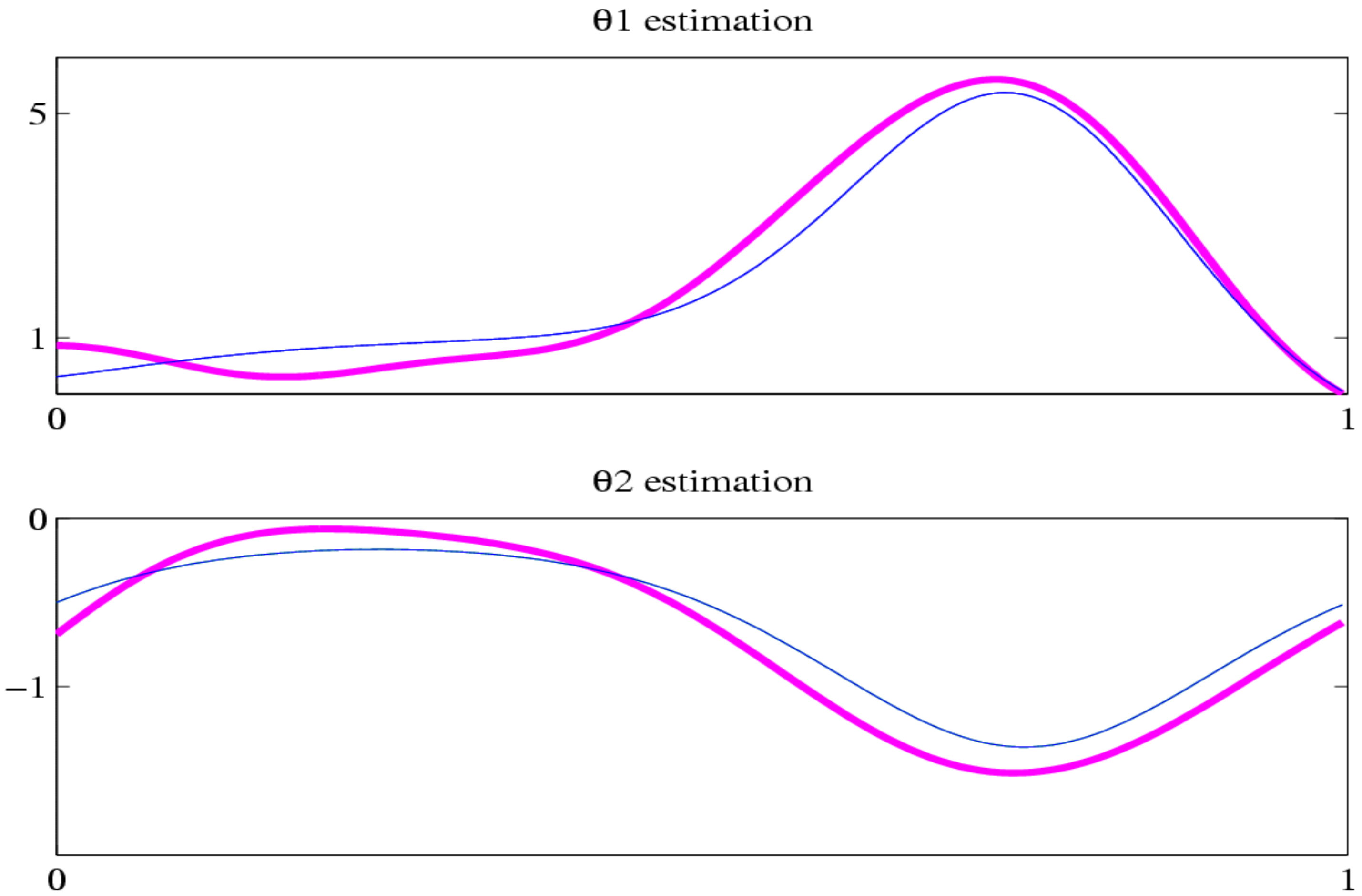
Properties of the problem:

- The problem is convex
- The log-determinant from the normalization of the Gaussian acts as a **barrrier function**.
- We get a semidefinite program.

Heteroscedastic Regression



Natural Parameters



Lecture 3

Novelty Detection

- Density estimation
- Thresholding and likelihood ratio

Classification

- Log partition function
- Optimization problem
- Examples
- Clustering and transduction

Regression

- Conditional normal distribution
- Estimating the covariance
- Heteroscedastic estimators