An Introduction to Machine Learning with Kernels

Lecture 7

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

Statistical Machine Learning Program
National ICT Australia, Canberra

Day 2

Text analysis and bioinformatics

Text categorization, biological sequences, kernels on strings, efficient computation, examples

Optimization

Sequential minimal optimization, convex subproblems, convergence, SVMLight, SimpleSVM

Regression and novelty detection

SVM regression, regularized least mean squares, adaptive margin width, novel observations

Practical tricks

Crossvalidation, ν-trick, median trick, data scaling, smoothness and kernels

L7 Novelty Detection and Regression

Novelty Detection

- Basic idea
- Optimization problem
- Stochastic Approximation
- Examples

LMS Regression

- Additive noise
- Regularization
- Examples
- SVM Regression

Novelty Detection

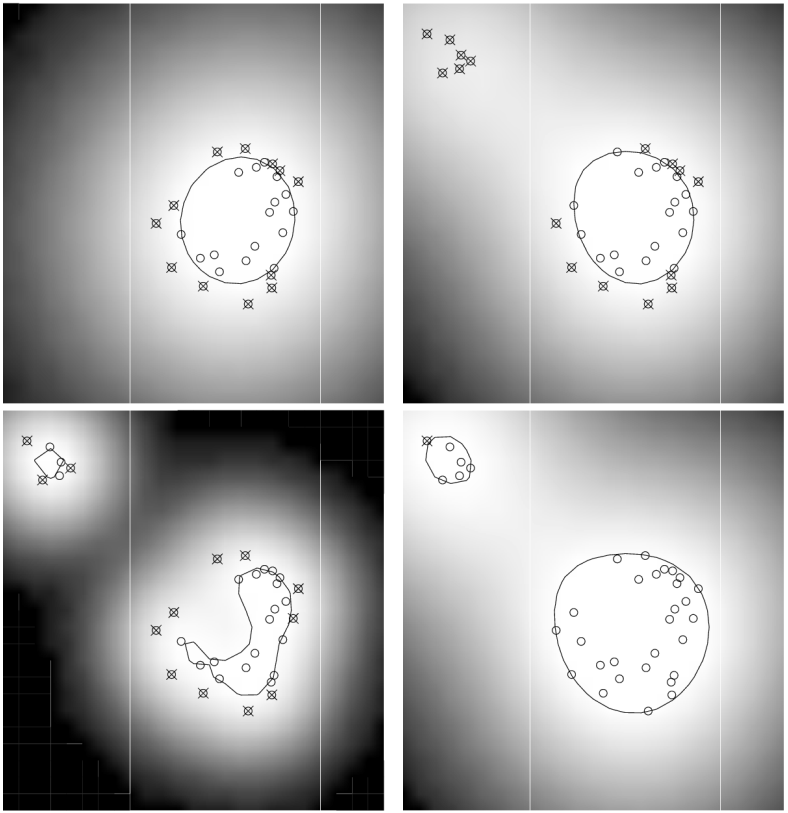
Data

Observations (x_i) generated from some $\mathrm{P}(x)$, e.g.,

- network usage patterns
- handwritten digits
- alarm sensors
- factory status

Task

Find unusual events, clean database, distinguish typical examples.



Applications

Network Intrusion Detection

Detect whether someone is trying to hack the network, downloading tons of MP3s, or doing anything else *unusual* on the network.

Jet Engine Failure Detection

You can't destroy jet engines just to see *how* they fail.

Database Cleaning

We want to find out whether someone stored bogus information in a database (typos, etc.), mislabelled digits, ugly digits, bad photographs in an electronic album.

Fraud Detection

Credit Cards, Telephone Bills, Medical Records

Self calibrating alarm devices

Car alarms (adjusts itself to where the car is parked), home alarm (furniture, temperature, windows, etc.)

Novelty Detection via Densities

Key Idea

- Novel data is one that we don't see frequently.
- It must lie in low density regions.

Step 1: Estimate density

- Observations $x_1, \ldots, x_m$
- Density estimate via Parzen windows

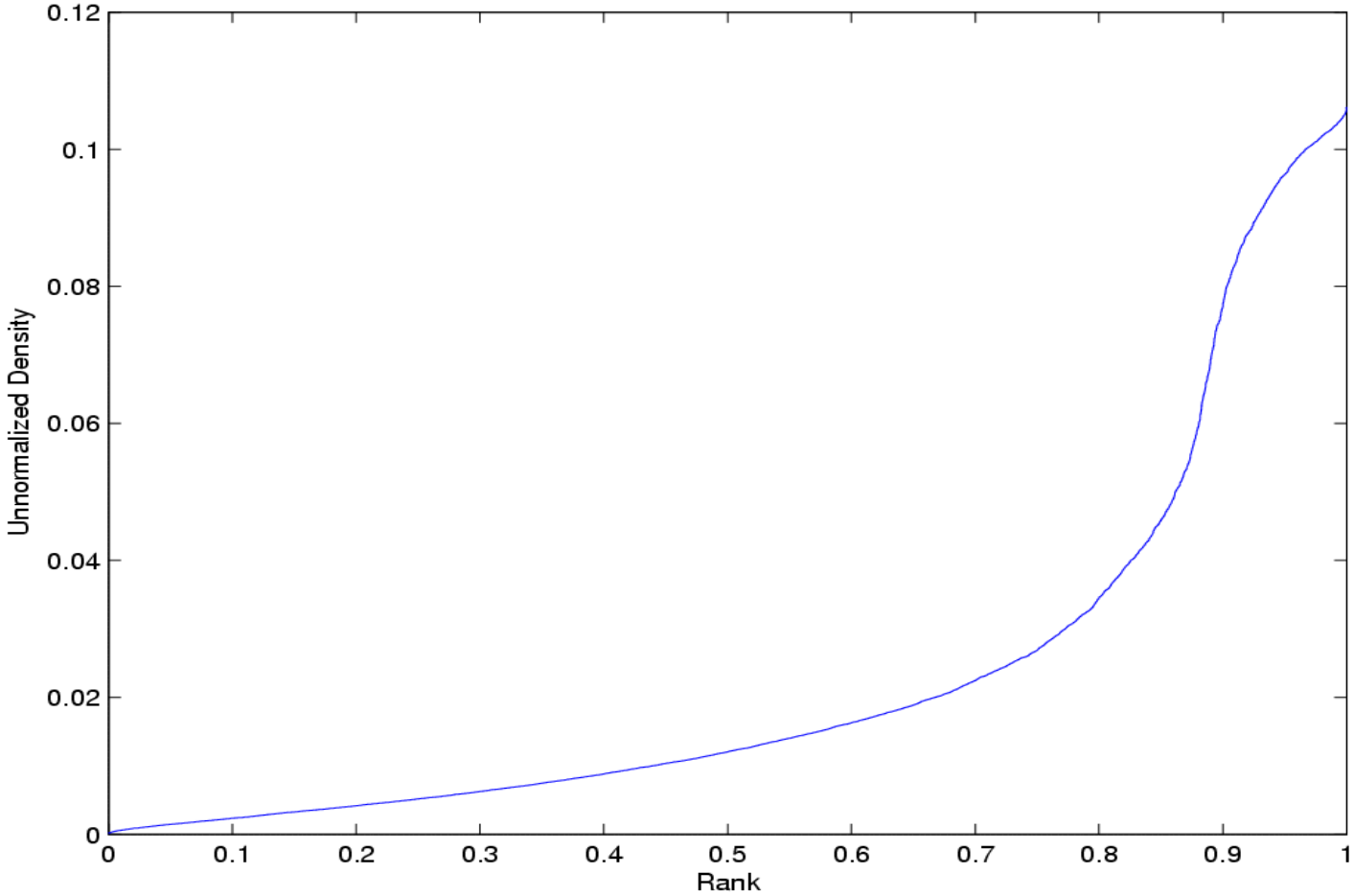
Step 2: Thresholding the density

- Sort data according to density and use it for rejection
- Practical implementation: compute

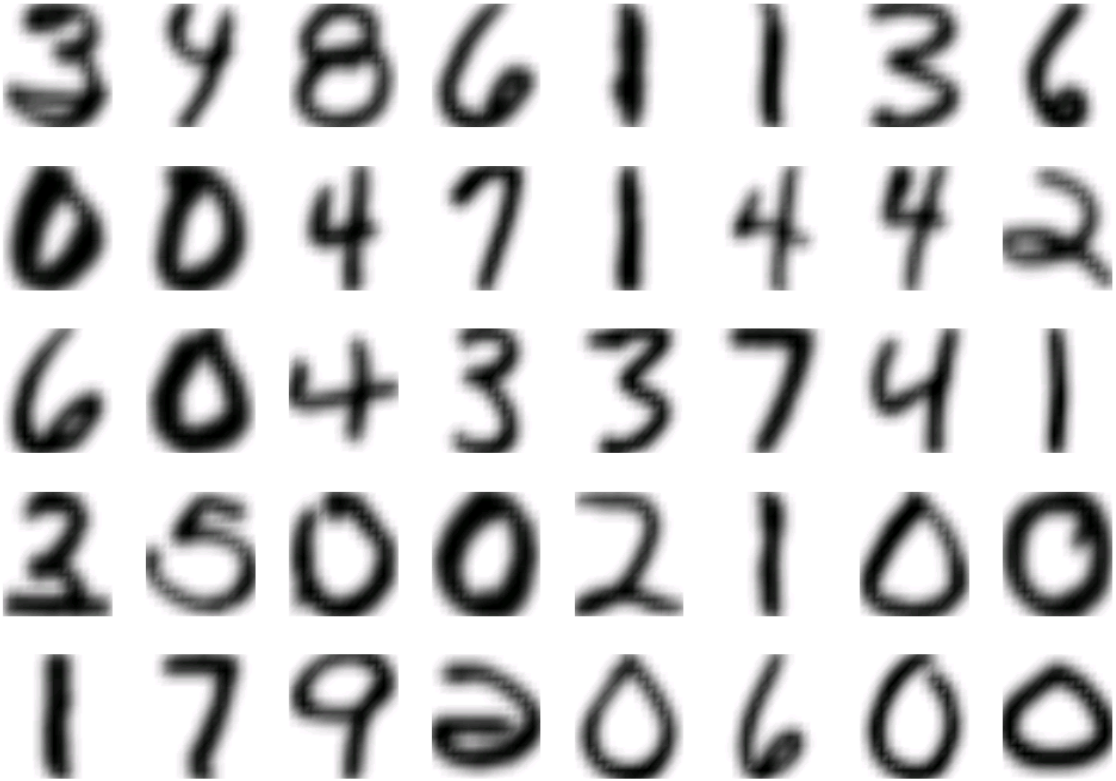
$$p(x_i) = \frac{1}{m} \sum_j k(x_i, x_j) \text{ for all } i$$

  and sort according to magnitude.
- Pick smallest $p(x_i)$ as novel points.

Order Statistic of Densities



Typical Data



Outliers

A better way . . .

Problems

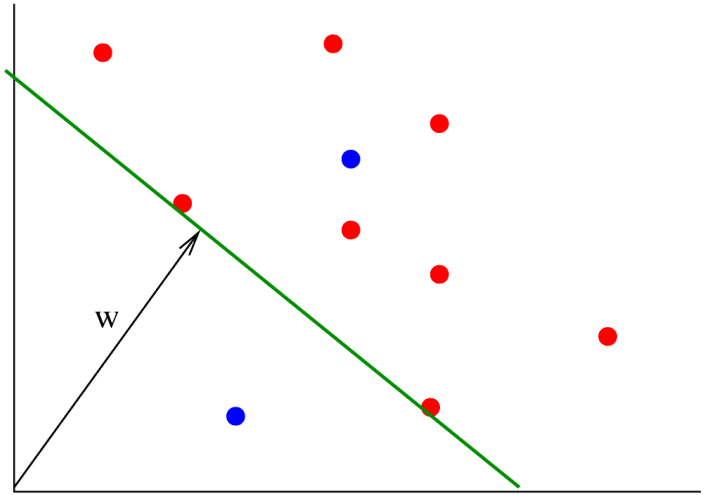
- We do not care about estimating the density properly in regions of high density (waste of capacity).
- We only care about the relative density for thresholding purposes.
- We want to eliminate a certain fraction of observations and tune our estimator specifically for this fraction.

Solution

- Areas of low density can be approximated as the **level set** of an auxiliary function. No need to estimate $p(x)$ directly — use proxy of $p(x)$.
- Specifically: find $f(x)$ such that x is novel if $f(x) \leq c$ where c is some constant, i.e. $f(x)$ describes the amount of novelty.

Maximum Distance Hyperplane

Idea Find hyperplane, given by $f(x) = \langle w, x \rangle + b = 0$ that has **maximum distance from origin** yet is still closer to the origin than the observations.



Hard Margin

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

Soft Margin

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

Dual Problem

Primal Problem

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

$$\text{subject to} \quad \langle w, x_i\rangle - 1 + \xi_i \geq 0 \text{ and } \xi_i \geq 0$$

Lagrange Function L

- Subtract constraints, multiplied by Lagrange multipliers (α_i and η_i), from Primal Objective Function.
- Lagrange function L has **saddlepoint** at optimum.

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

subject to $\alpha_i, \eta_i \geq 0$.

Dual Problem

Optimality Conditions

$$\partial_w L = w - \sum_{i=1}^m \alpha_i x_i = 0 \implies w = \sum_{i=1}^m \alpha_i x_i$$
$$\partial_{\xi_i} L = C - \alpha_i - \eta_i = 0 \implies \alpha_i \in [0, C]$$

Now **substitute** the optimality conditions **back into** L.

Dual Problem

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

All this is only possible due to the convexity of the primal problem.

The ν-Trick

Problem

- Depending on C, the number of novel points will vary.
- We would like to **specify the fraction** ν beforehand.

Solution

Use hyperplane separating data from the origin

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

where the threshold ρ is **adaptive**.

Intuition

- Let the hyperplane shift by shifting ρ
- Adjust it such that the 'right' number of observations is considered novel.
- Do this automatically

The ν-Trick

Primal Problem

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

$$\text{where } \langle w, x_i \rangle - \rho + \xi_i \geq 0$$

$$\xi_i \geq 0$$

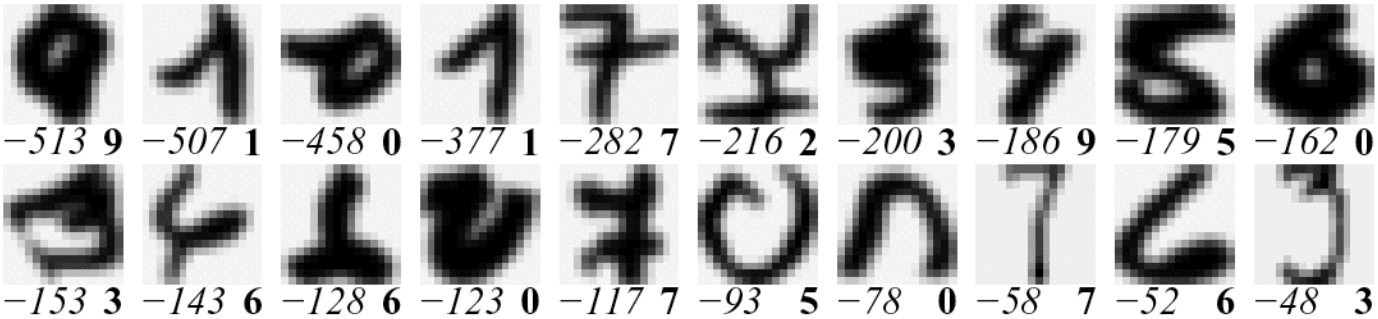
Dual Problem

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

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

Similar to SV classification problem, use standard optimizer for it.

USPS Digits



- Better estimates since we only optimize in low density regions.
- Specifically tuned for small number of outliers.
- Only estimates of a level-set.
- For $\nu = 1$ we get the Parzen-windows estimator back.

A Simple Online Algorithm

Objective Function

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

Stochastic Approximation

$$\frac{1}{2}\|w\|^2 \max(0, \rho - \langle w, \phi(x_i)\rangle) - \nu\rho$$

Gradient

$$\partial_w[\ldots] = \begin{cases} w - \phi(x_i) & \text{if } \langle w, \phi(x_i)\rangle < \rho \\ w & \text{otherwise} \end{cases}$$

$$\partial_\rho[\ldots] = \begin{cases} (1-\nu) & \text{if } \langle w, \phi(x_i)\rangle < \rho \\ -\nu & \text{otherwise} \end{cases}$$

Practical Implementation

Update in coefficients

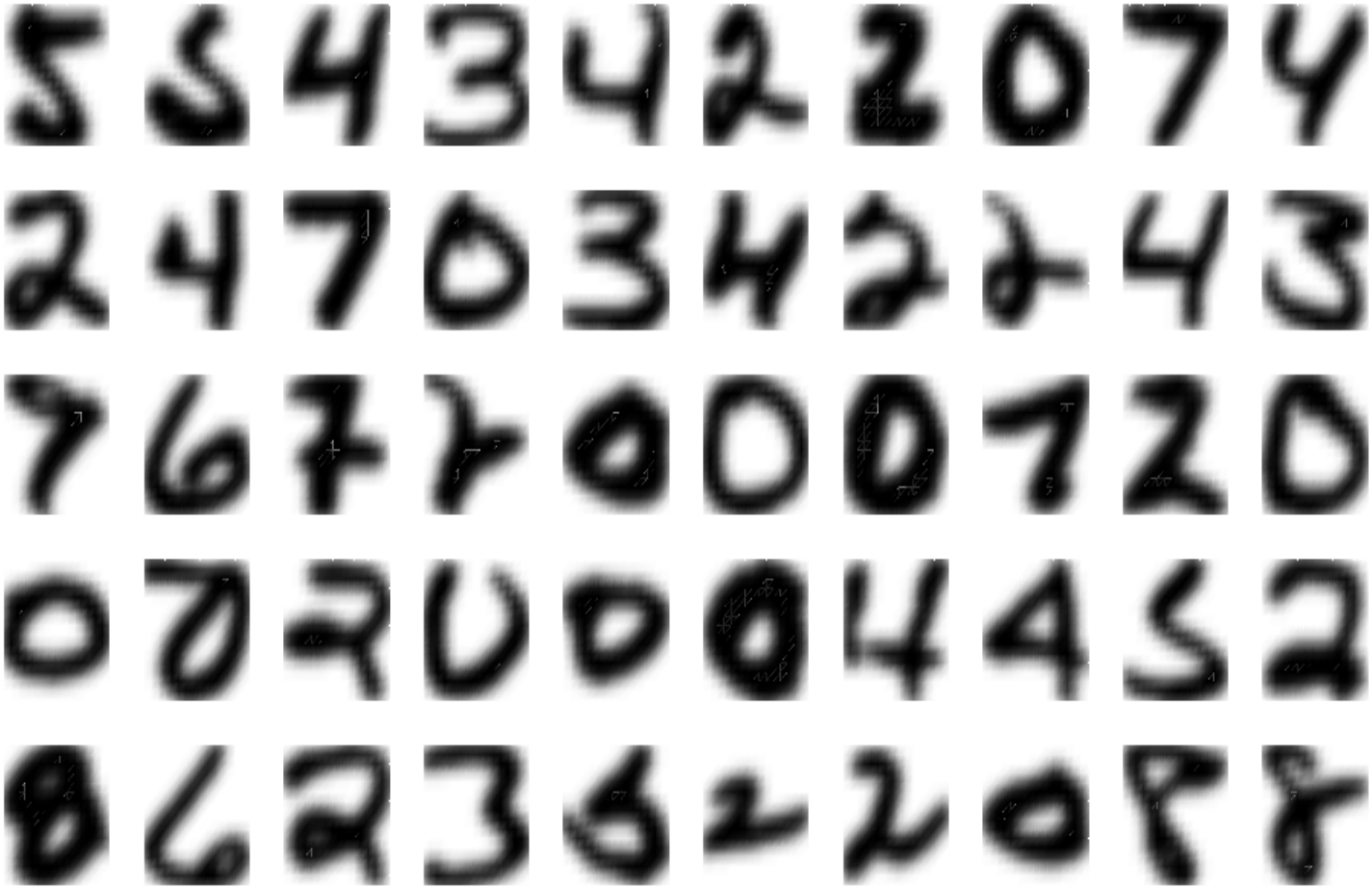
$$\alpha_j \leftarrow (1-\eta)\alpha_j \text{ for } j \neq i$$

$$\alpha_i \leftarrow \begin{cases} \eta_i & \text{if } \sum_{j=1}^{i-1} \alpha_i k(x_i, x_j) < \rho \\ 0 & \text{otherwise} \end{cases}$$

$$\rho = \begin{cases} \rho + \eta(\nu - 1) & \text{if } \sum_{j=1}^{i-1} \alpha_i k(x_i, x_j) < \rho \\ \rho + \eta\nu & \text{otherwise} \end{cases}$$

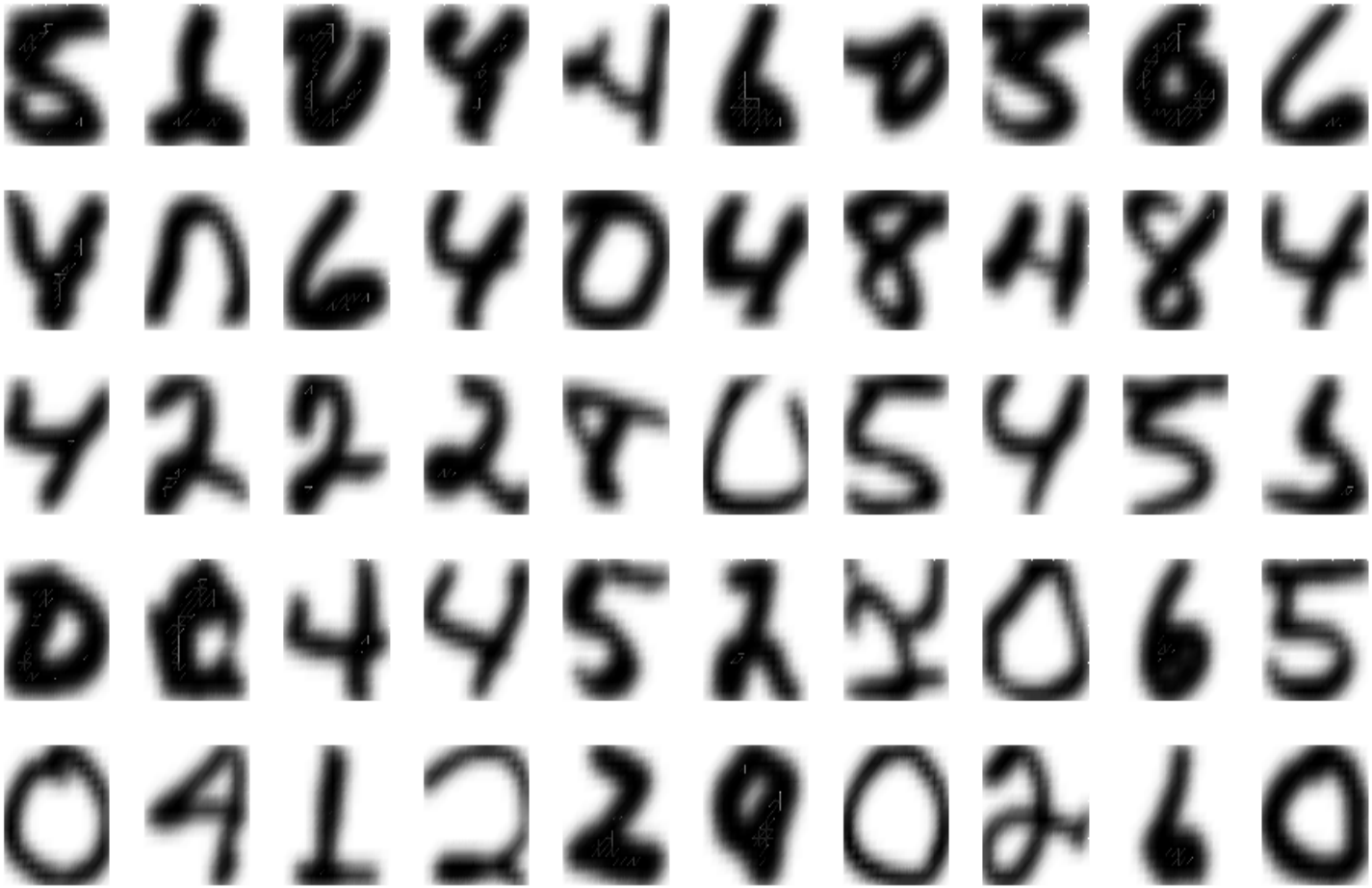
Using learning rate η.

Online Training Run



Worst Training Examples

Worst Test Examples



Mini Summary

Novelty Detection via Density Estimation

- Estimate density e.g. via Parzen windows
- Threshold it at level and pick low-density regions as novel

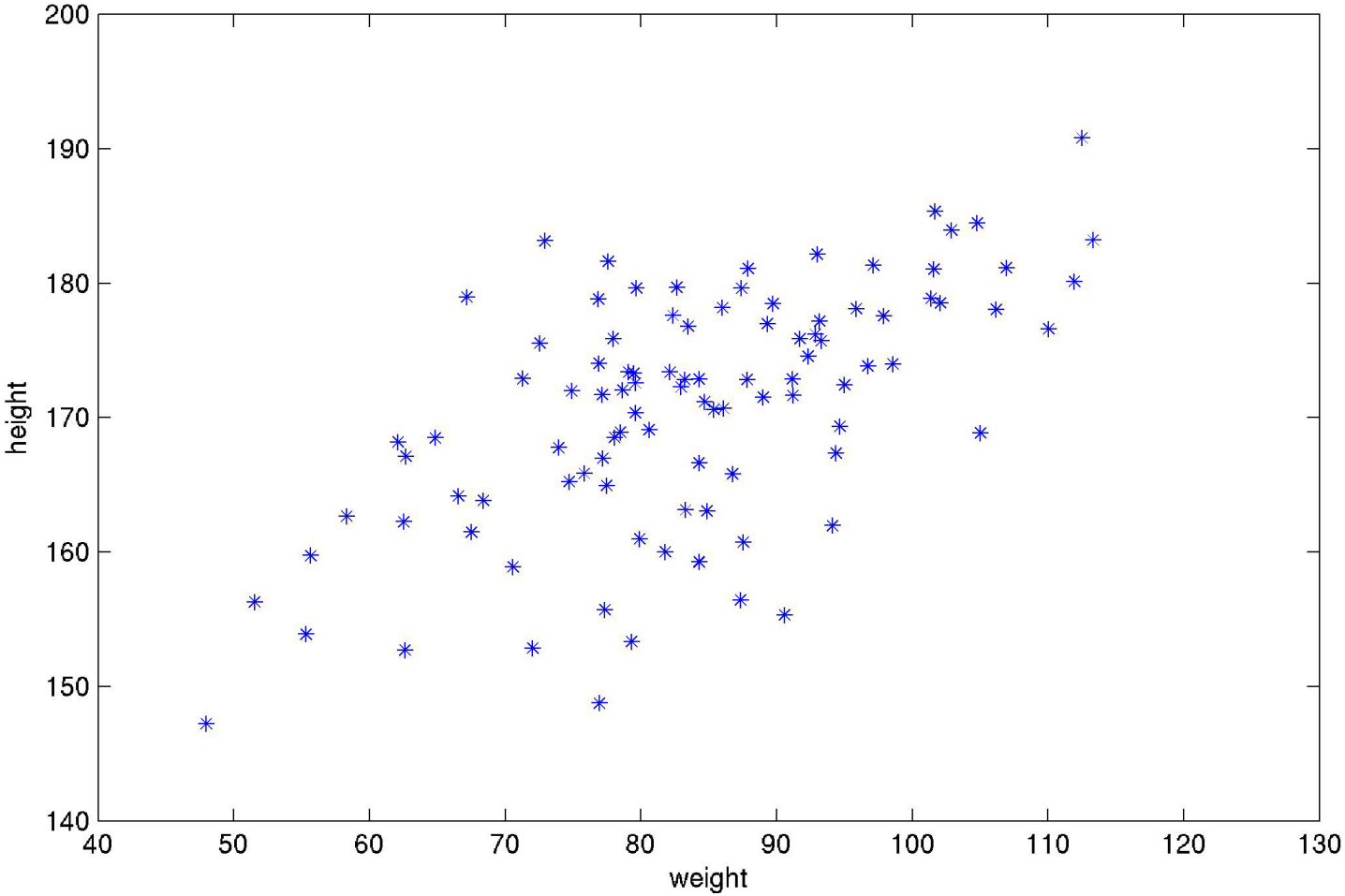
Novelty Detection via SVM

- Find halfspace bounding data
- Quadratic programming solution
- Use existing tools

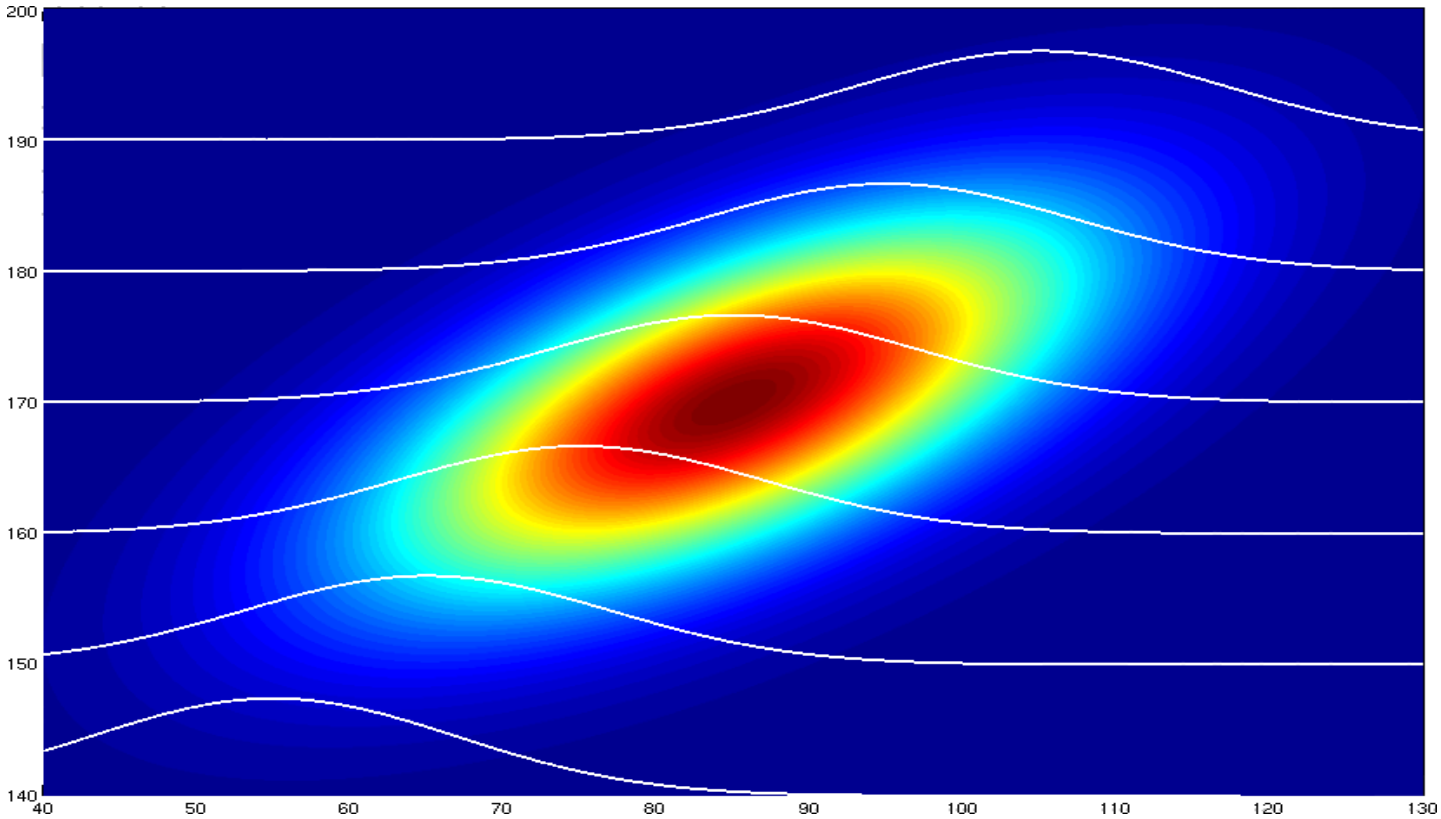
Online Version

- Stochastic gradient descent
- Simple update rule: keep data if novel, but only with fraction ν and adjust threshold.
- Easy to implement

A simple problem



Inference



$$p(\text{weight}|\text{height}) = \frac{p(\text{height}, \text{weight})}{p(\text{height})} \propto p(\text{height}, \text{weight})$$

Bayesian Inference HOWTO

Conditional probability

- If we have conditional probability $p(y|x)$ we can estimate y (here x are the observations and y is what we want to compute).
- For instance, we can get the regression by computing the mean of $p(y|x)$.

Joint to conditional probability

- Joint can be used to get conditional, via Bayes rule

$$p(x,y) = p(y|x)p(x) \text{ and hence } p(y|x) = \frac{p(x,y)}{p(x)} \propto p(x,y)$$

- Expression only depends on y for fixed x in $p(x,y)$.

Normal Distribution in $\mathbb{R}^n$

Normal Distribution in $\mathbb{R}$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}$.

Normal Distribution in $\mathbb{R}^n$

$$p(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

Parameters

- $\mu \in \mathbb{R}^n$ is the mean.
- $\Sigma \in \mathbb{R}^{n\times n}$ is the covariance. Note that this is now a matrix.
- Σ has only nonnegative eigenvalues (i.e. the variance is never negative).

Inference in Normal Distributions

Correlated Observations

Assume that the random variables $t \in \mathbb{R}^n, t' \in \mathbb{R}^{n'}$ are jointly normal with mean (μ, μ') and covariance matrix K

$$p(t, t') \propto \exp\left(-\frac{1}{2}\begin{bmatrix} t-\mu \\ t'-\mu' \end{bmatrix}^\top \begin{bmatrix} K_{tt} & K_{tt'} \\ K_{tt'}^\top & K_{t't'} \end{bmatrix}^{-1} \begin{bmatrix} t-\mu \\ t'-\mu' \end{bmatrix}\right).$$

Inference

Given t, estimate t' via $p(t'|t)$. Translation into machine learning language: we learn t' from t.

Practical Solution

Since $t'|t \sim \mathcal{N}(\tilde{\mu}, \tilde{K})$, we only need to collect all terms in $p(t, t')$ depending on t' by matrix inversion, hence

$$\tilde{K} = K_{t't'} - K_{tt'}^\top K_{tt}^{-1} K_{tt'} \text{ and } \tilde{\mu} = \mu' + K_{tt'}^\top \underbrace{\left[K_{tt}^{-1}(t-\mu)\right]}_{\text{independent of } t'}$$

Gaussian Process

Key Idea

Instead of a fixed set of random variables t, t' we assume a stochastic process $t : \mathcal{X} \to \mathbb{R}$, e.g. $\mathcal{X} = \mathbb{R}^n$.

Previously we had $\mathcal{X} = \{\text{age}, \text{height}, \text{weight}, \ldots\}$.

Definition of a Gaussian Process

A stochastic process $t : \mathcal{X} \to \mathbb{R}$, where all $(t(x_1), \ldots, t(x_m))$ are normally distributed.

Parameters of a GP

Mean	$\mu(x) := \mathbf{E}[t(x)]$
Covariance Function	$k(x, x') := \mathrm{Cov}(t(x), t(x'))$

Simplifying Assumption

We assume knowledge of $k(x, x')$ and set $\mu = 0$.

Some Covariance Functions

Observation

Any function k leading to a symmetric matrix with non-negative eigenvalues is a valid covariance function.

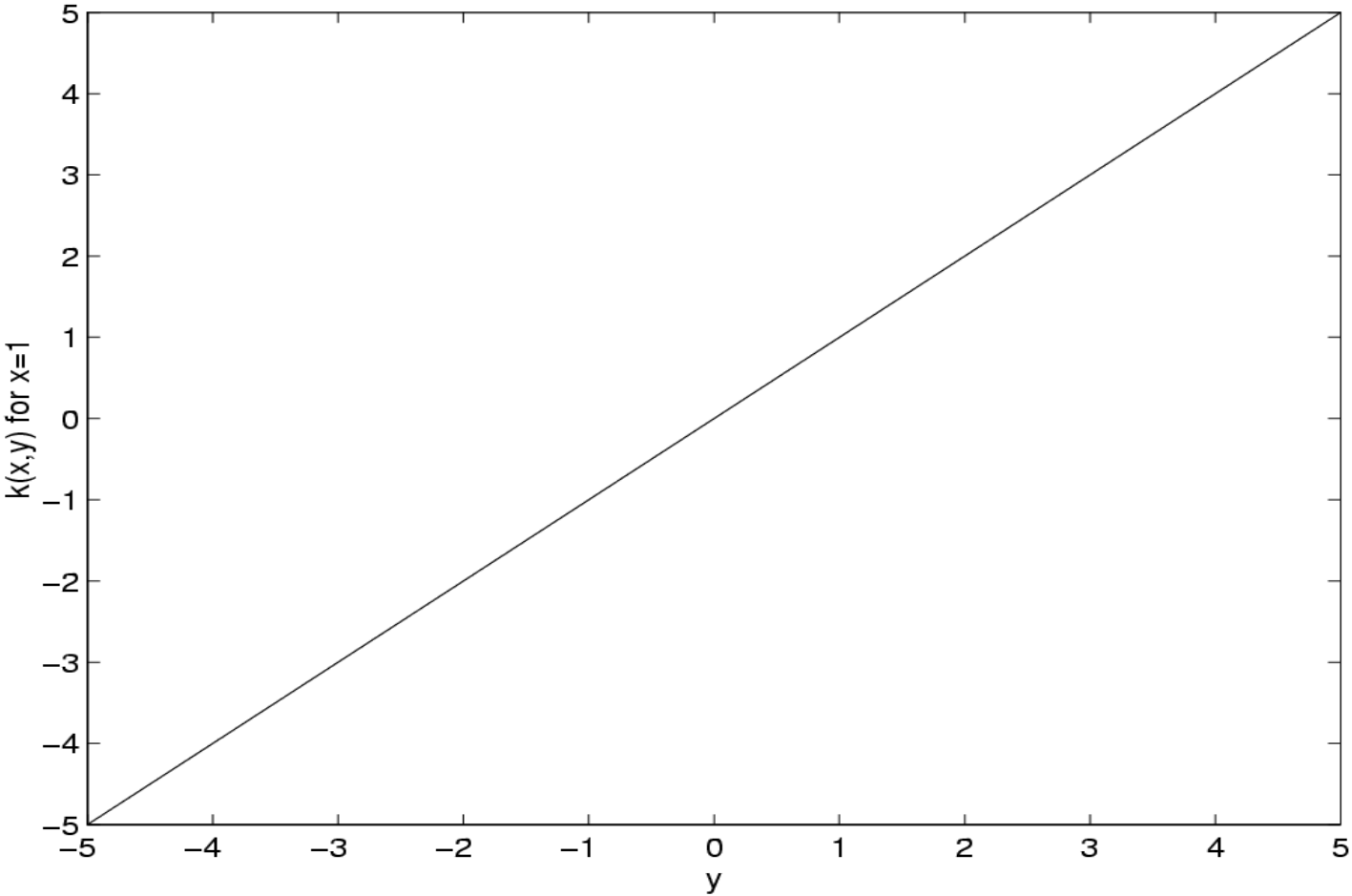
Necessary and sufficient condition (Mercer's Theorem)

k needs to be a nonnegative integral kernel.

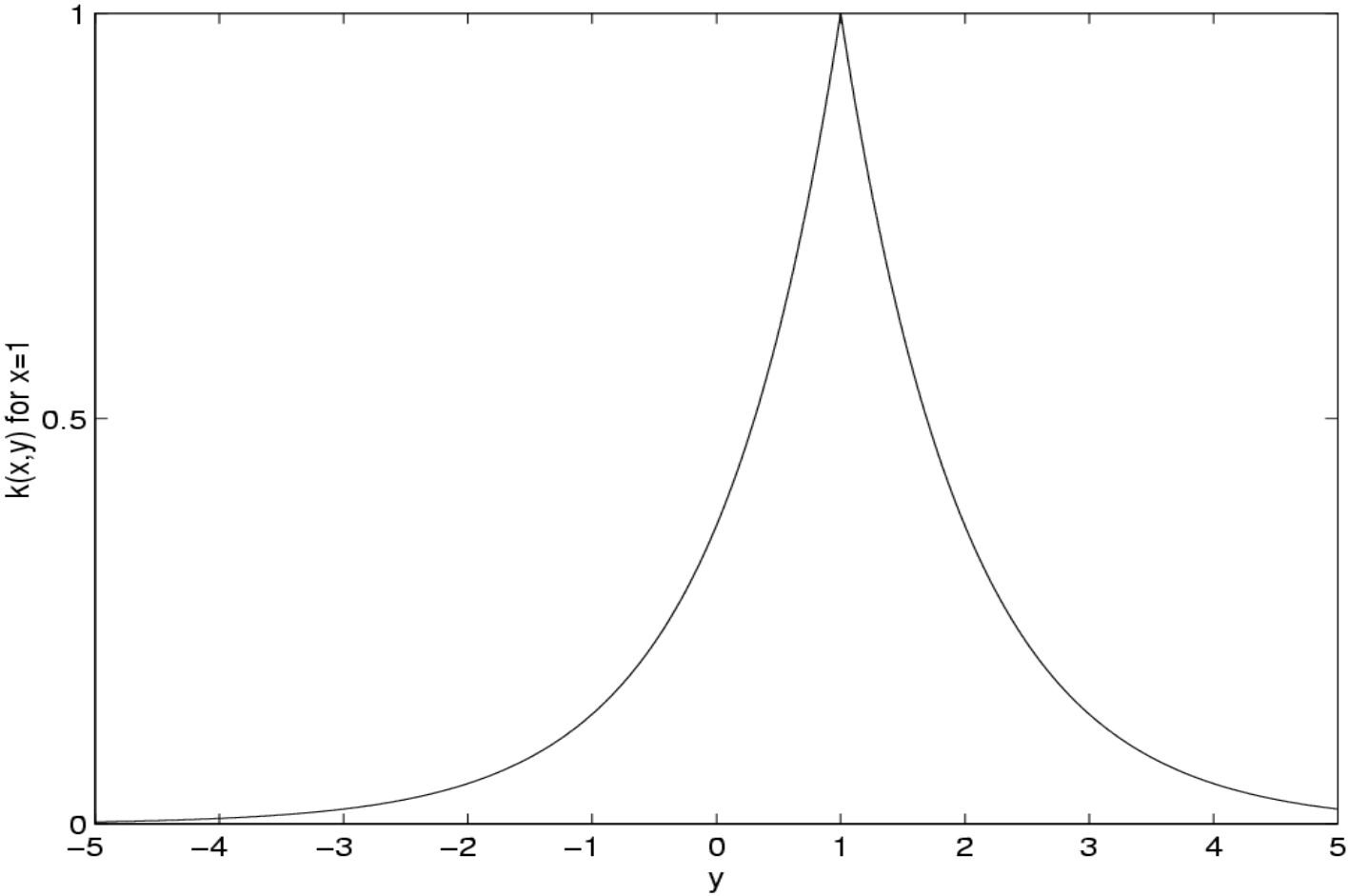
Examples of kernels $k(x, x')$

Linear	$\langle x, x' \rangle$
Laplacian RBF	$\exp\left(-\lambda \|x - x'\|\right)$
Gaussian RBF	$\exp\left(-\lambda \|x - x'\|^2\right)$
Polynomial	$\left(\langle x, x' \rangle + c\rangle\right)^d, c \geq 0,\ d \in \mathbb{N}$
B-Spline	$B_{2n+1}(x - x')$
Cond. Expectation	$\mathbf{E}_c[p(x\|c)p(x'\|c)]$

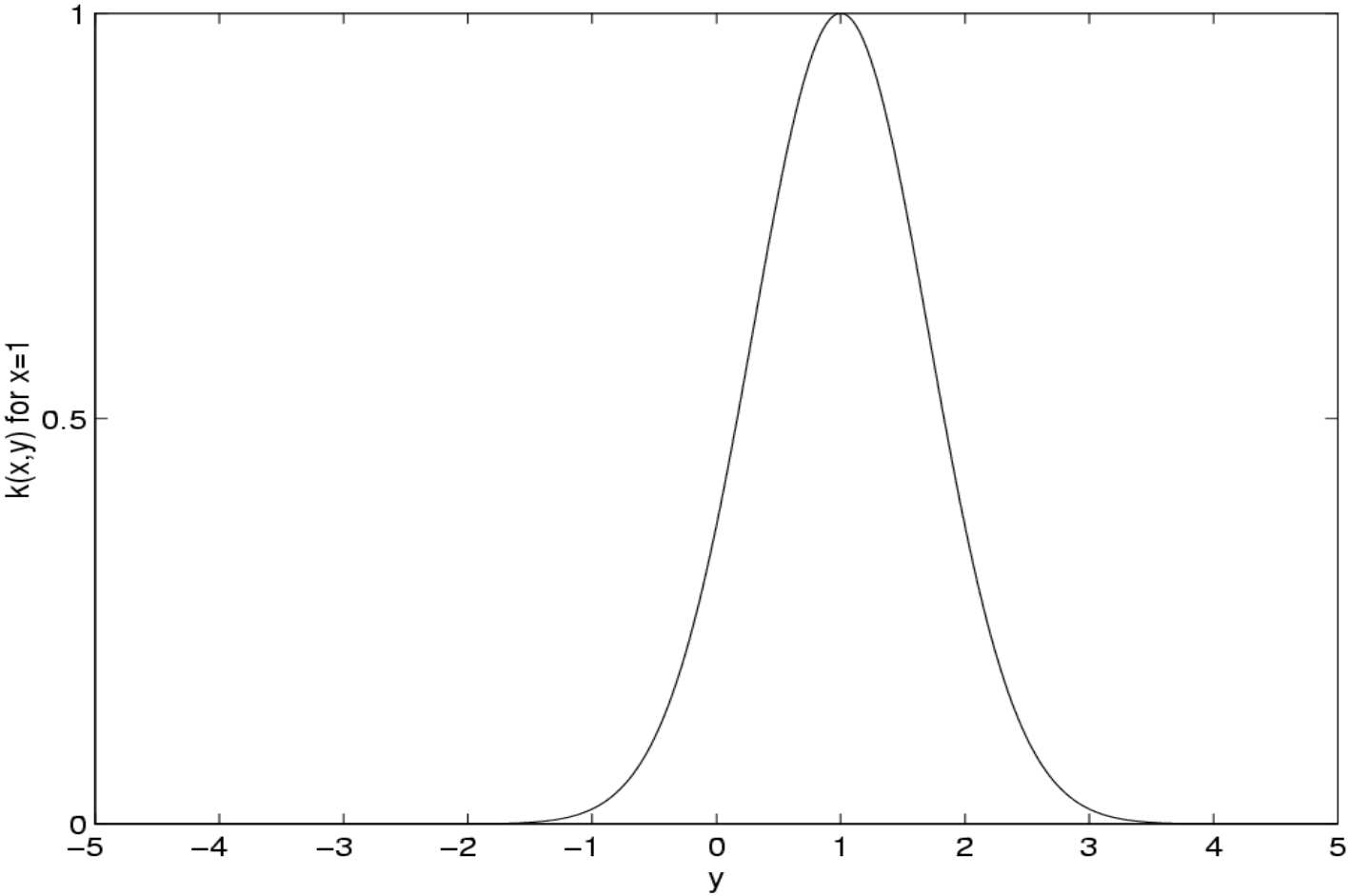
Linear Covariance



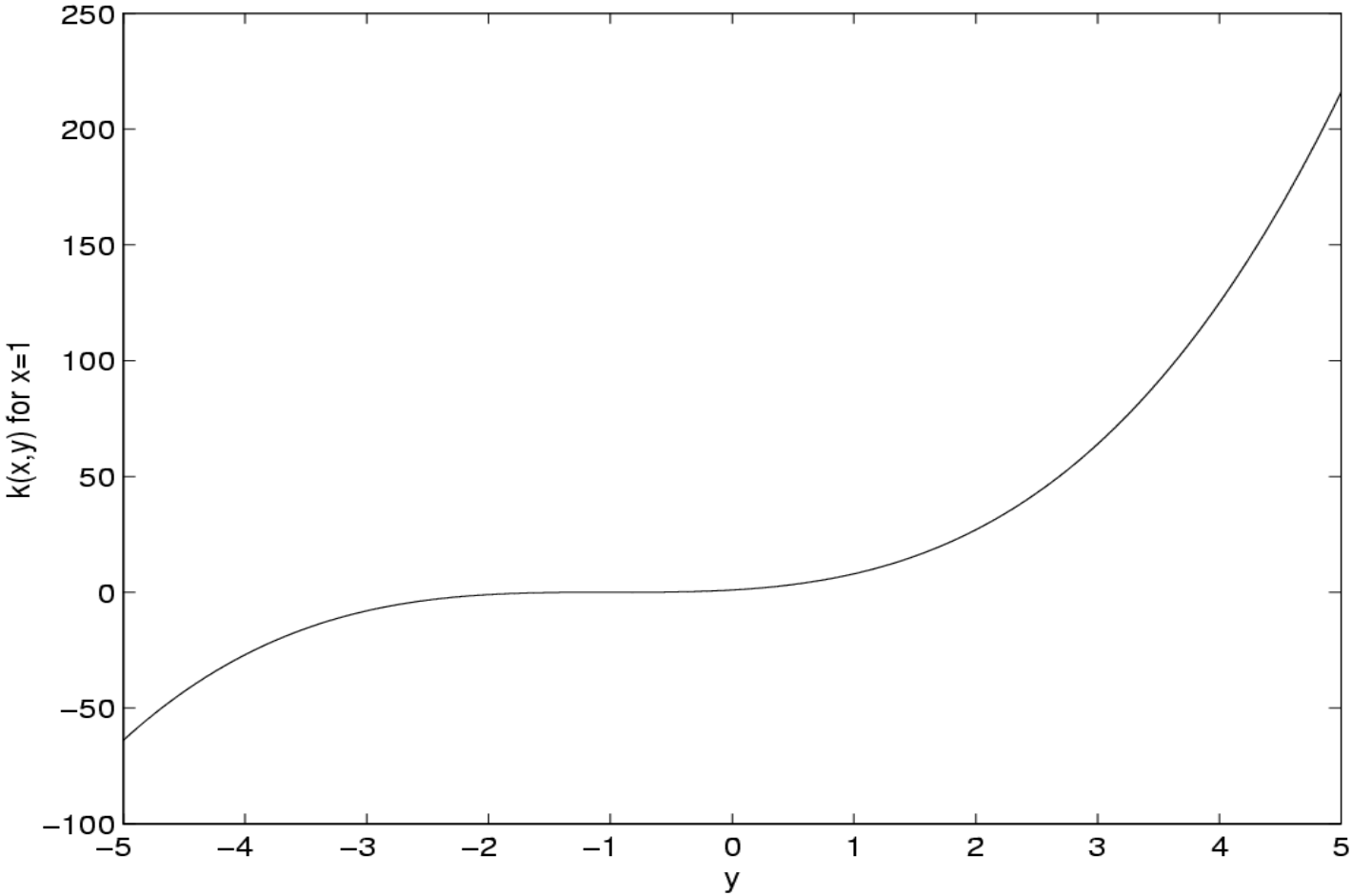
Laplacian Covariance



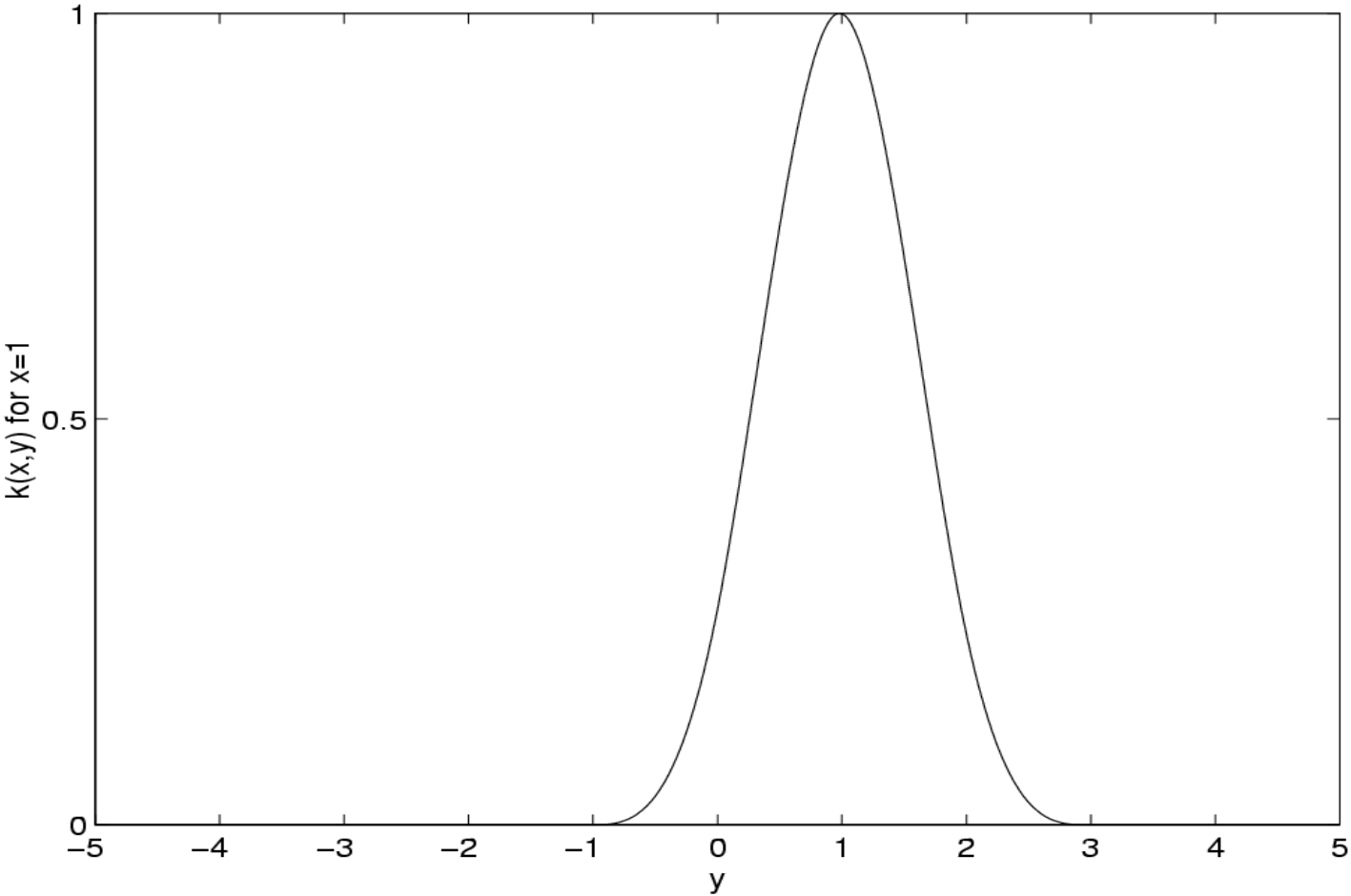
Gaussian Covariance



Polynomial (Order 3)



B_3-Spline Covariance



Gaussian Processes and Kernels

Covariance Function

- Function of two arguments
- Leads to matrix with nonnegative eigenvalues
- Describes correlation between pairs of observations

Kernel

- Function of two arguments
- Leads to matrix with nonnegative eigenvalues
- Similarity measure between pairs of observations

Lucky Guess

- We suspect that kernels and covariance functions are the same . . .

The Support Vector Connection

Gaussian Process on Parameters

$$t \sim \mathcal{N}(\mu, K) \text{ where } K_{ij} = k(x_i, x_j)$$

Linear Model in Feature Space

$$t(x) = \langle \Phi(x), w \rangle + \mu(x) \text{ where } w \sim \mathcal{N}(0, \mathbf{1})$$

The covariance between $t(x)$ and $t(x')$ is then given by

$$\mathbf{E}_w \left[\langle \Phi(x), w \rangle \langle w, \Phi(x') \rangle \right] = \langle \Phi(x), \Phi(x') \rangle = k(x, x')$$

Conclusion

A small weight vector in "feature space", as commonly used in SVM amounts to observing t with high $p(t)$.

Log prior $-\log p(t) \iff$ **Margin** $\|w\|^2$

Will get back to this later again.

Regression

Simple Model

Assume correlation between $t(x)$ and $t(x')$ via $k(x, x')$, so we can perform regression on $t(x')$, given $t(x)$.

Recall

$$p(t, t') \propto \exp\left(-\frac{1}{2}\begin{bmatrix} t \\ t' \end{bmatrix}^\top \begin{bmatrix} K_{tt} & K_{tt'} \\ K_{tt'}^\top & K_{t't'} \end{bmatrix}^{-1} \begin{bmatrix} t \\ t' \end{bmatrix}\right)$$

yields $t'|t \sim \mathcal{N}(\tilde{\mu}, \tilde{K})$, where

$$\tilde{K} = K_{t't'} - K_{tt'}^\top K_{tt}^{-1} K_{tt'} \text{ and } \tilde{\mu} = K_{tt'}^\top K_{tt}^{-1} t$$

Proof Idea

- $t'|t$ is normally distributed, hence we need only get the linear and quadratic terms in t'.
- Quadratic term via inverse of big covariance matrix.
- Linear term (for the mean) has cross terms with t.

Example: Linear Regression

Linear kernel: $k(x, x') = \langle x, x' \rangle$

- Kernel matrix $X^\top X$
- Mean and covariance

$$\tilde{K} = X'^\top X' - X'^\top X (X^\top X)^{-1} X^\top X' = X'^\top (\mathbf{1} - P_X) X'.$$

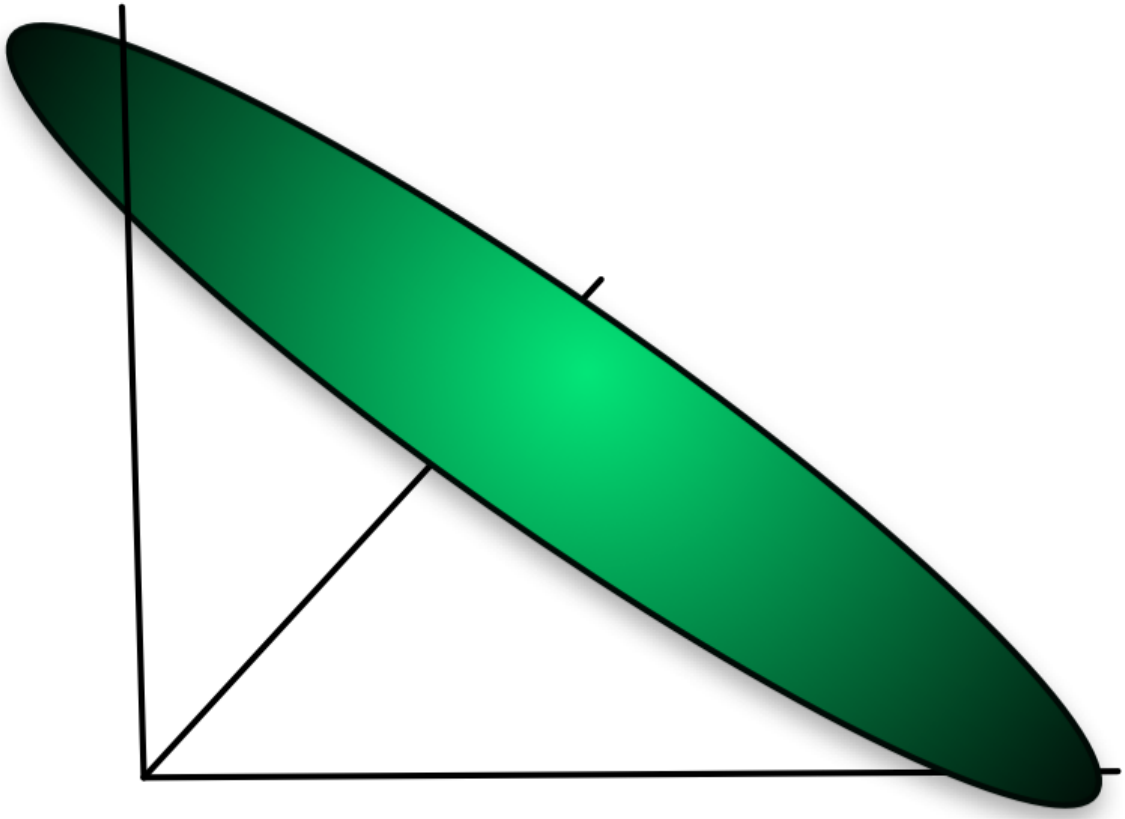
$$\tilde{\mu} = X'^\top \left[X (X^\top X)^{-1} t \right]$$

- $\tilde{\mu}$ is a **linear function of** X'.

Problem

- The covariance matrix $X^\top X$ has at most rank n.
- After n observations ($x \in \mathbb{R}^n$) the **variance vanishes**. This is **not realistic**.
- "Flat pancake" or "cigar" distribution.

Degenerate Covariance



Additive Noise

Indirect Model

Instead of observing $t(x)$ we observe $y = t(x) + \xi$, where ξ is a nuisance term. This yields

$$p(Y|X) = \int \prod_{i=1}^{m} p(y_i|t_i)p(t|X)dt$$

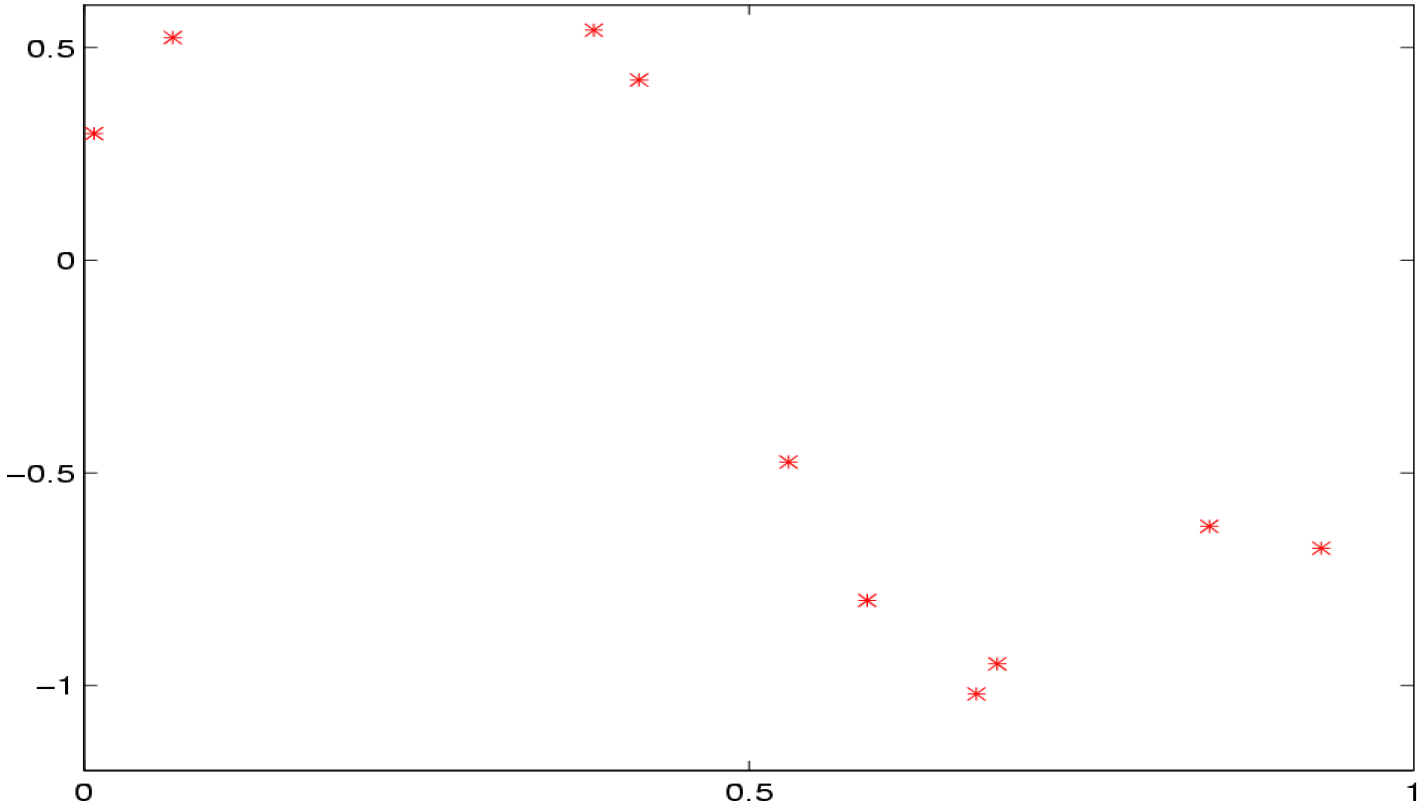
where we can now find a maximum a posteriori solution for t by maximizing the integrand (we will use this later).

Additive Normal Noise

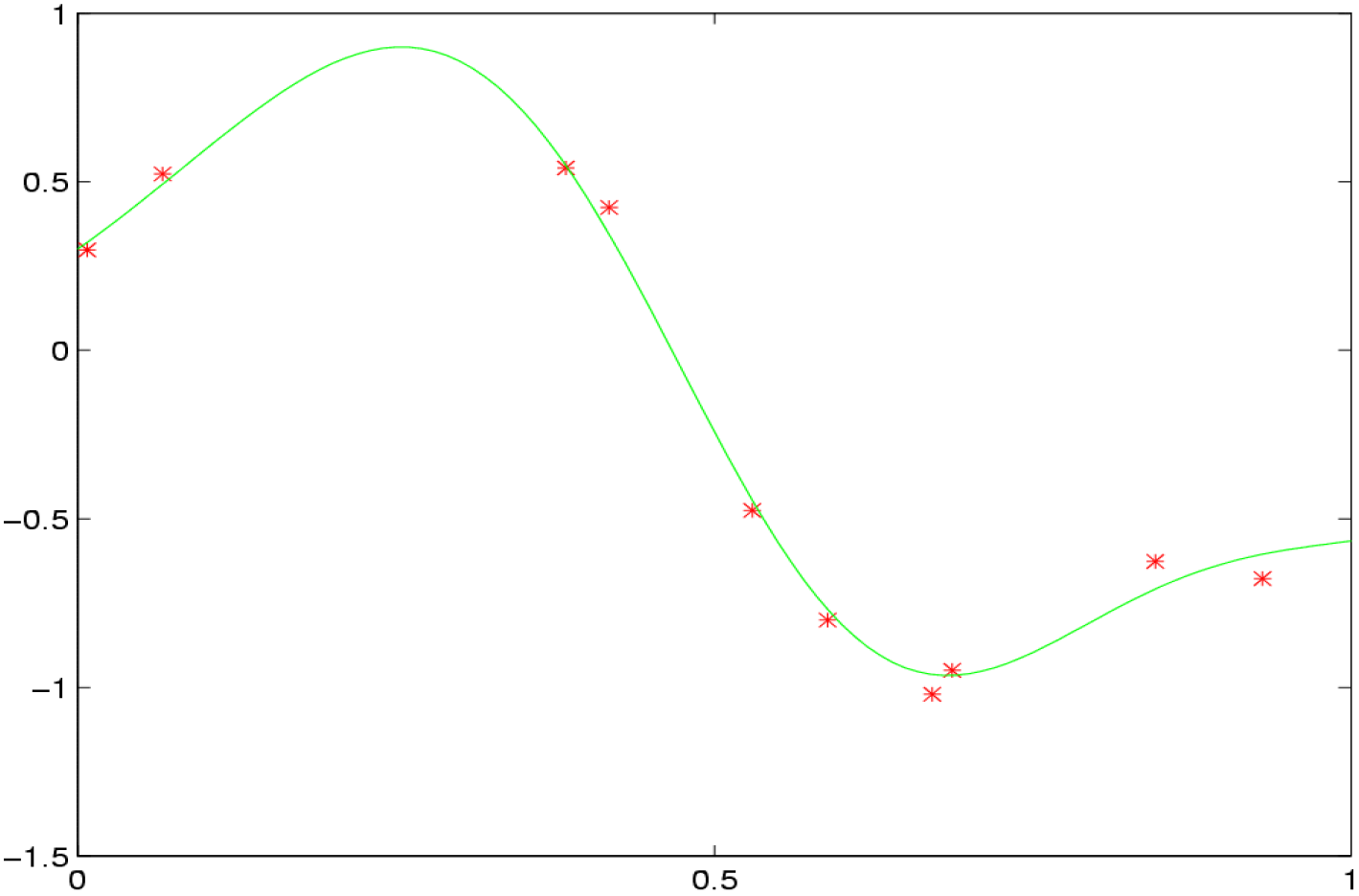
- If $\xi \sim \mathcal{N}(0, \sigma^2)$ then y is the sum of two Gaussian random variables.
- Means and variances **add up**.

$$y \sim \mathcal{N}(\mu, K + \sigma^2 \mathbf{1}).$$

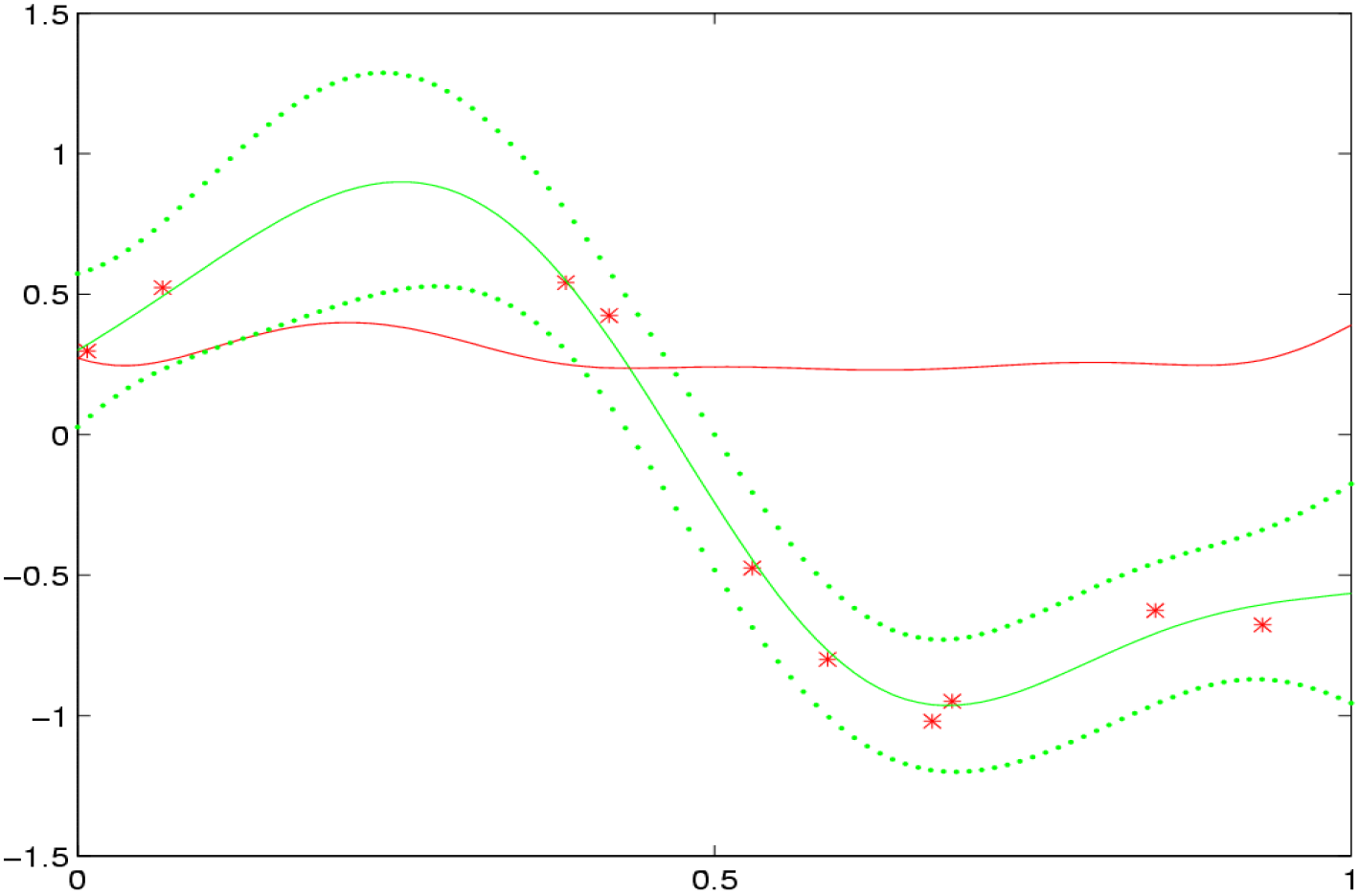
Training Data



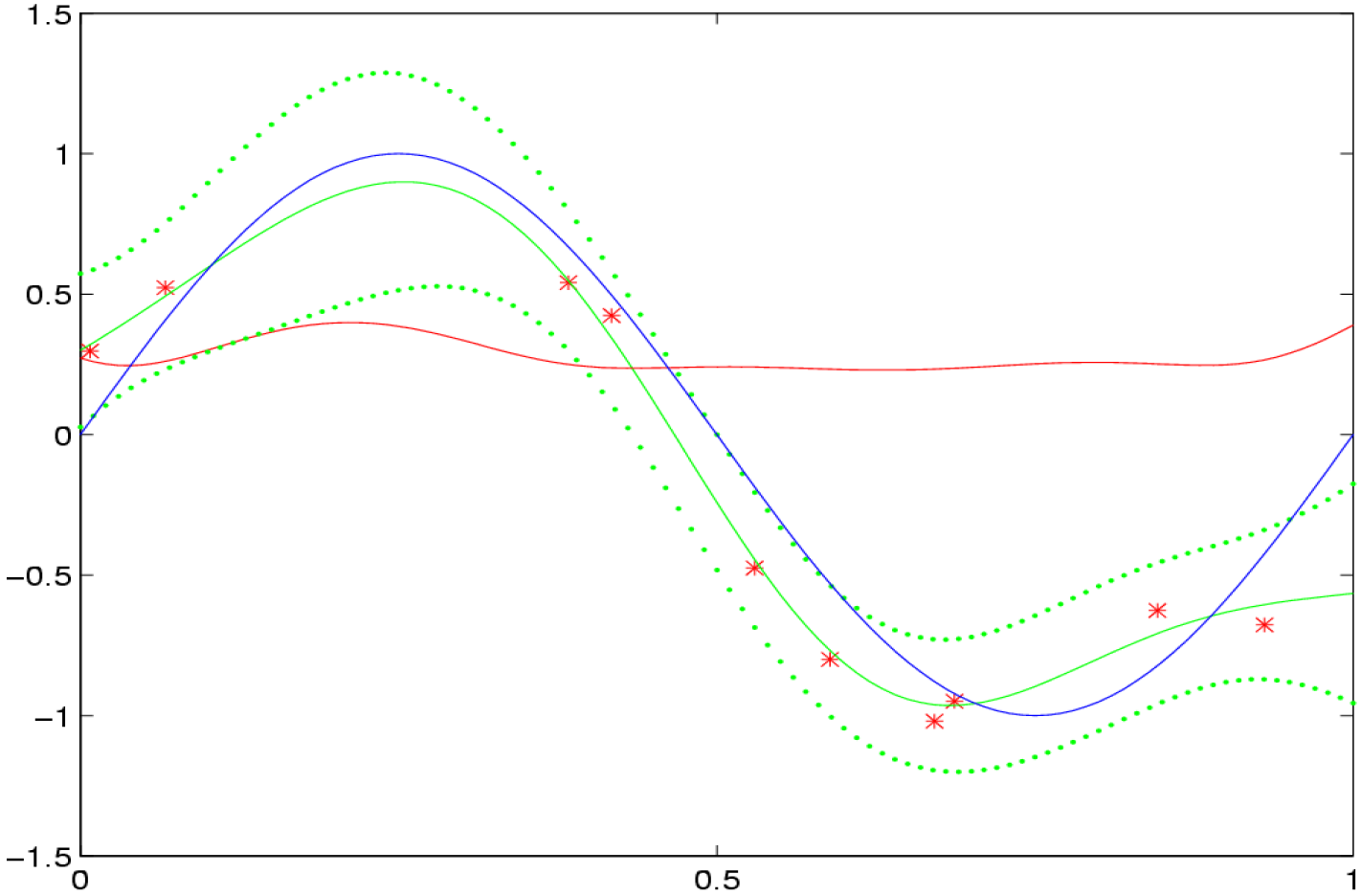
Mean $\vec{k}^\top(x)(K+\sigma^2\mathbf{1})^{-1}y$



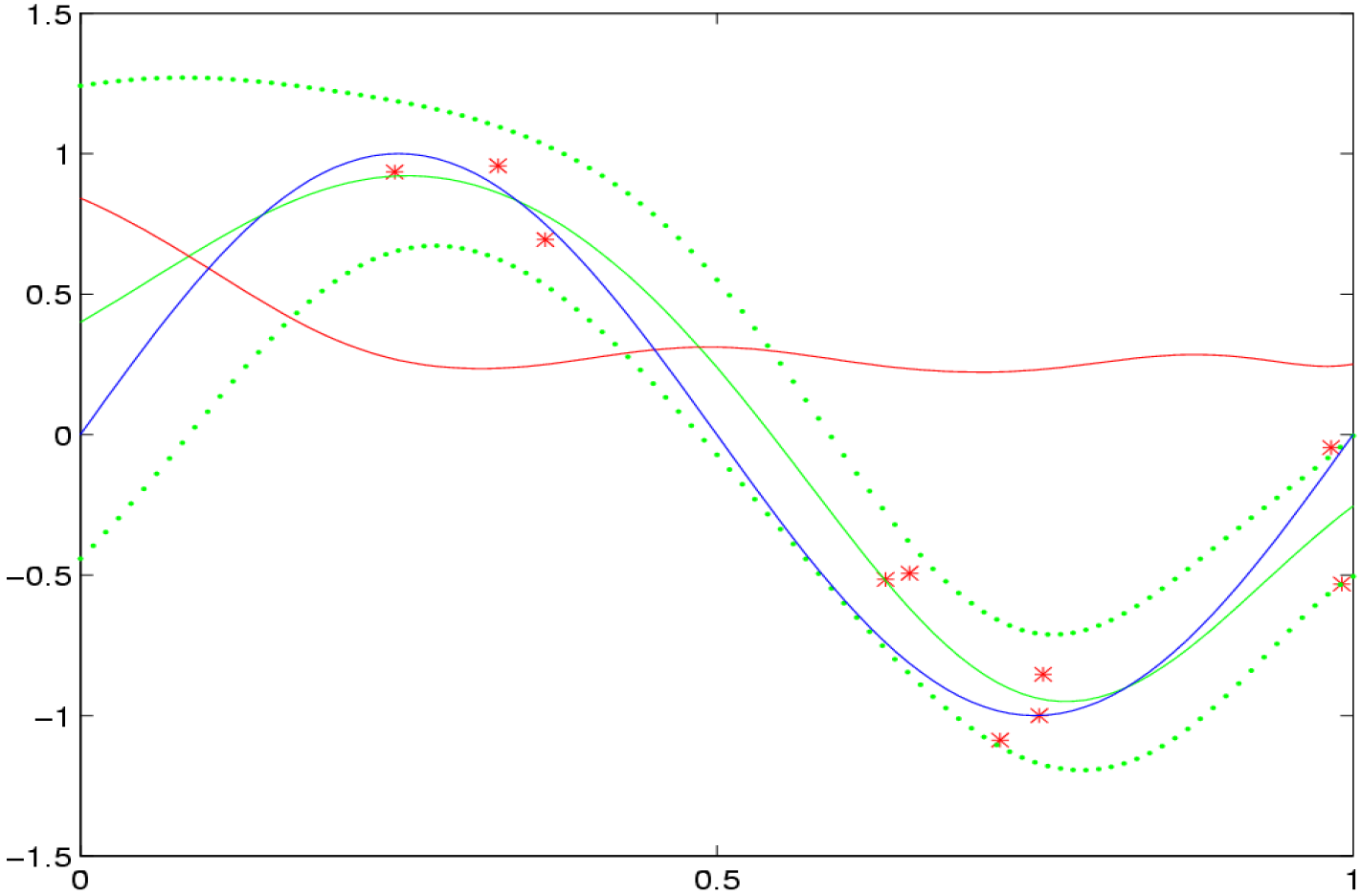
Variance $k(x,x) + \sigma^2 - \vec{k}^\top(x)(K + \sigma^2 \mathbf{1})^{-1}\vec{k}(x)$



Putting everything together . . .



Another Example



The ugly details

Covariance Matrices

- Additive noise

$$K = K_{\text{kernel}} + \sigma^2 \mathbf{1}$$

- Predictive mean and variance

$$\tilde{K} = K_{t't'} - K_{tt'}^\top K_{tt}^{-1} K_{tt'} \text{ and } \tilde{\mu} = K_{tt'}^\top K_{tt}^{-1} t$$

Pointwise prediction

$$\begin{aligned} K_{tt} &= K + \sigma^2 \mathbf{1} \\ K_{t't'} &= k(x, x) + \sigma^2 \\ K_{tt'} &= (k(x_1, x), \ldots, k(x_m, x)) \end{aligned}$$

Plug this into the mean and covariance equations.

The Support Vector Connection

SV Optimization Problem

$$\text{minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \text{loss}(x_i, y_i, w)$$

Quadratic Loss

- Least mean squares regression

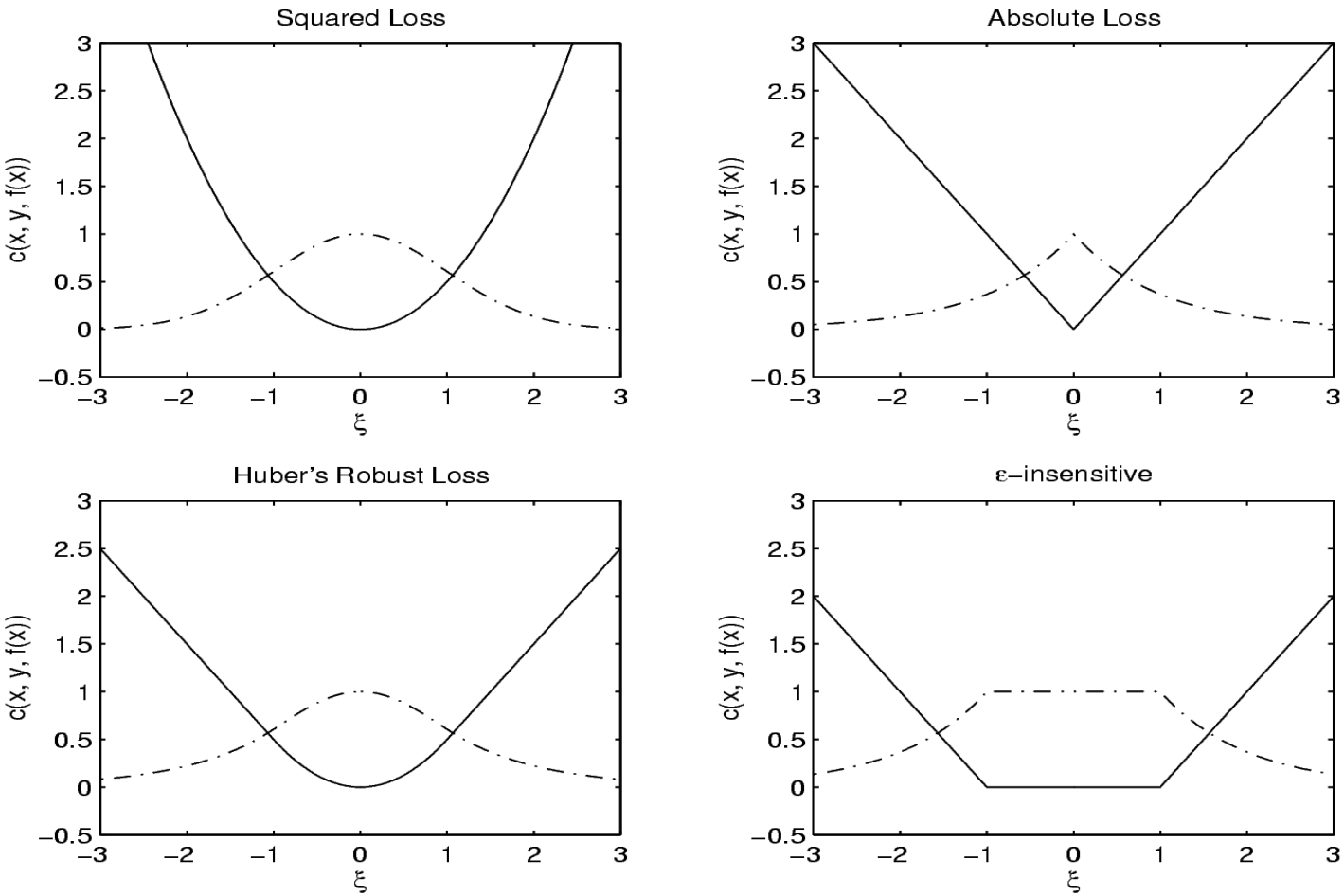
$$\text{minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \text{loss}(y_i - \langle\phi(x_i), w\rangle)^2$$

- Solution

$$w = \sum_{i=1}^{m} \alpha_i \phi(x_i) \text{ where } \alpha = (K + C^{-1}\mathbf{1})y$$

This is identical to the GP regression (where $C = \sigma^{-2}$).

Regression loss functions



Mini Summary

Gaussian Process

- Like function, just random
- Mean and covariance determine the process
- Can use it for estimation

Regression

- Jointly normal model
- Additive noise to deal with error in measurements
- Estimate for mean and uncertainty

SV and GP connection

- GP kernel and SV kernel are the same
- Just different loss functions

Summary

Novelty Detection

- Basic idea
- Optimization problem
- Stochastic Approximation
- Examples

LMS Regression

- Additive noise
- Regularization
- Examples
- SVM Regression

Shameless Plugs

We are hiring. For details contact

- Alex.Smola@nicta.com.au (http://www.nicta.com.au)

Positions

- PhD scholarships
- Postdoctoral positions, Senior researchers
- Long-term visitors (sabbaticals etc.)

More details on kernels

- **http://www.kernel-machines.org**
- **http://www.learning-with-kernels.org**
  Schölkopf and Smola: Learning with Kernels

Machine Learning Summer School

- **http://canberra05.mlss.cc**
- MLSS'05 Canberra, Australia, 23/1-5/2/2005