Exponential Families and Kernels
Lecture 2

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 2

Clifford Hammersley Theorem and Graphical Models

- Decomposition results
- Key connection

Conditional Distributions

- Log partition function
- Expectations and derivatives
- Inner product formulation and kernels
- Gaussian Processes

Applications

- Classification + Regression
- Conditional Random Fields
- Spatial Poisson Models

Graphical Model

Conditional Independence

- x, x' are conditionally independent given c, if

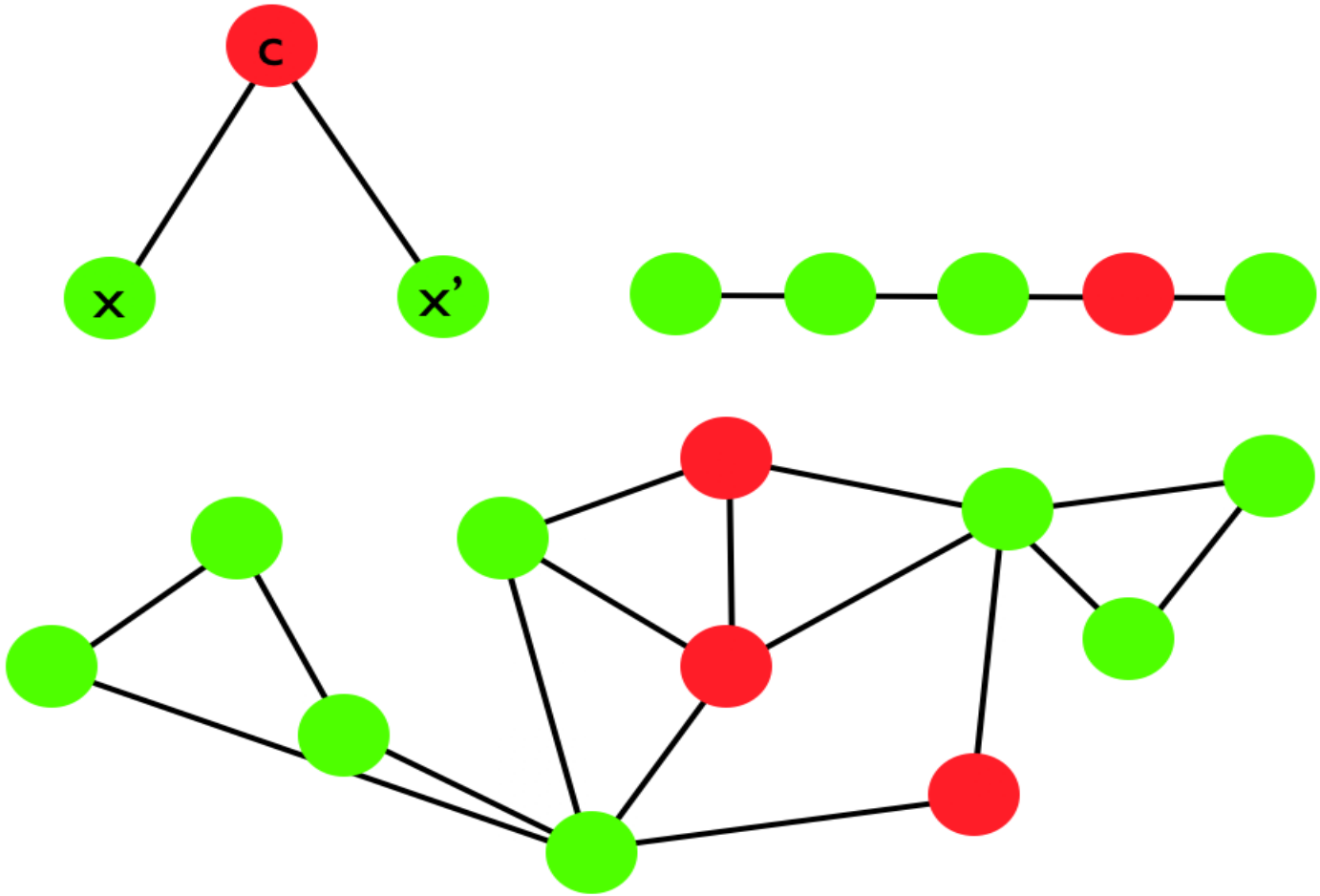
$$p(x, x'|c) = p(x|c)p(x'|c)$$

- Distributions can be simplified greatly by conditional independence assumptions.

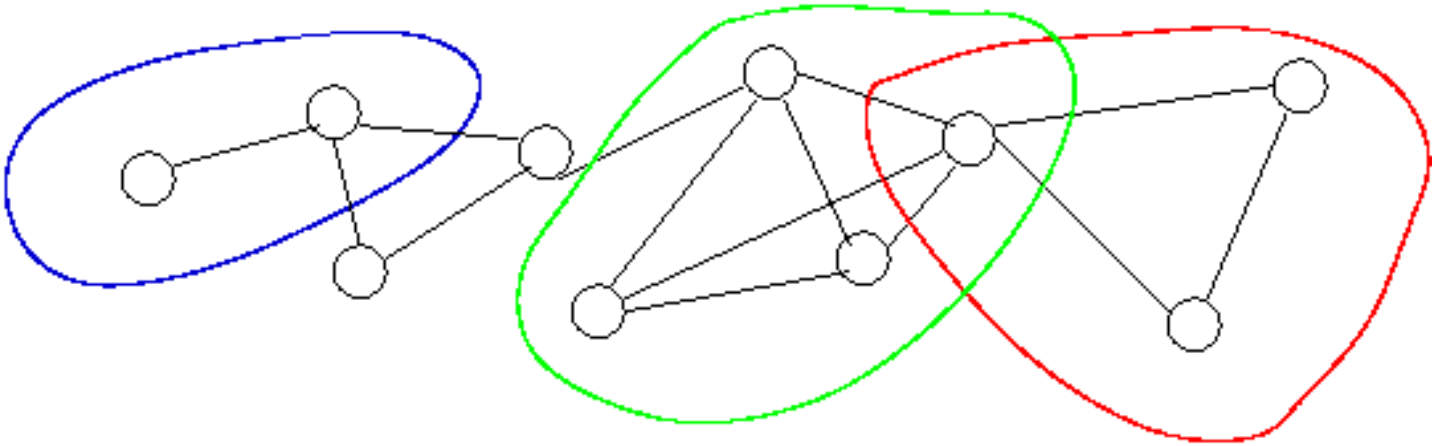
Markov Network

- Given a graph $G(V, E)$ with vertices V and edges E associate a random variable $x \in \mathbb{R}^{|V|}$ with G.
- Subsets of random variables $x_S, x_{S'}$ are conditionally independent given x_C if removing the vertices C from $G(V, E)$ decomposes the graph into disjoint subsets containing S, S'.

Conditional Independence



Cliques



Definition

- Subset of the graph which is fully connected
- Maximal Cliques (they define the graph)

Advantage

- Easy to specify dependencies between variables
- Use graph algorithms for inference

Hammersley Clifford Theorem

Problem

Specify $p(x)$ with conditional independence properties.

Theorem

$$p(x) = \frac{1}{Z} \exp\left(\sum_{c \in \mathcal{C}} \psi_c(x_c)\right)$$

whenever $p(x)$ is nonzero on the entire domain.

Application

Apply decomposition for exponential families where $p(x) = \exp(\langle \phi(x), \theta \rangle - g(\theta))$.

Corollary

The sufficient statistics $\phi(x)$ decompose according to

$$\phi(x) = (\ldots, \phi_c(x_c), \ldots) \Longrightarrow \langle \phi(x), \phi(x') \rangle = \sum_c \langle \phi_c(x_c), \phi_c(x'_c) \rangle$$

Proof

Step 1: Obtain linear functional

Combing the exponential setting with the CH theorem:

$$\langle \Phi(\mathbf{x}), \theta \rangle = \sum_{c \in C} \psi_c(x_c) - \log Z + g(\theta) \text{ for all } \mathbf{x}, \theta.$$

Step 2: Orthonormal basis in θ

Pick an orthonormal basis and swallow Z, g. This gives

$$\langle \Phi(\mathbf{x}), e_i \rangle = \sum_{c \in C} \eta_c^i(x_c) \text{ for some } \eta_c^i(x_c).$$

Step 3: Reconstruct sufficient statistics

$$\Phi_c(x_c) := (\eta_c^1(x_c), \eta_c^2(x_c), \ldots)$$

which allows us to compute

$$\langle \Phi(\mathbf{x}), \theta \rangle = \sum_{c \in C} \sum_i \theta_i \Phi_c^i(x_c).$$

Example: Normal Distributions

Sufficient Statistics

Recall that for normal distributions $\phi(x) = (x, xx^\top)$.

Clifford Hammersley Application

- $\phi(x)$ must decompose into subsets involving only variables from each maximal clique.
- The linear term x is OK by default.
- The only nonzero terms coupling $x_i x_j$ are those corresponding to an edge in the graph $G(V, E)$.

Inverse Covariance Matrix

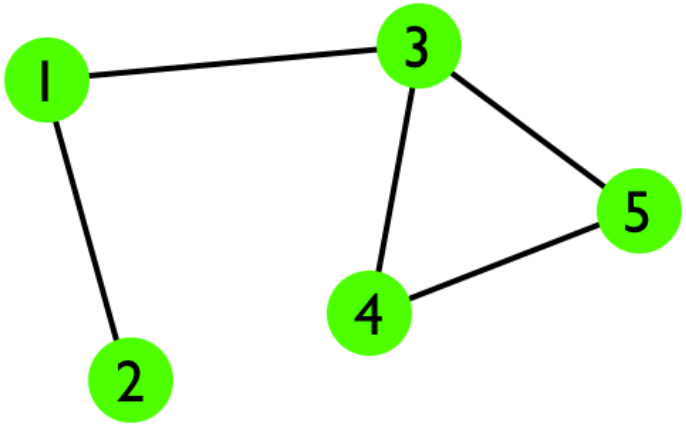
- The natural parameter aligned with $xx^\top$ is the inverse covariance matrix.
- Its sparsity mirrors $G(V, E)$.
- Hence a sparse inverse kernel matrix corresponds to graphical model!

Example: Normal Distributions

Density

$$p(x|\theta) = \exp\left(\sum_{i=1}^{n} x_i\theta_{1i} + \sum_{i,j=1}^{n} x_i x_j \theta_{2ij} - g(\theta)\right)$$

Here $\theta_2 = \Sigma^{-1}$, is the inverse covariance matrix. We have that $(\Sigma^{-1})_{[ij]} \neq 0$ only if (i, j) share an edge.



	1	2	3	4	5
1	■	■			
2	■	■	■		
3		■	■	■	■
4			■	■	■
5			■	■	■

Conditional Distributions

Conditional Density

$$p(x|\theta) = \exp(\langle \phi(x), \theta \rangle - g(\theta))$$
$$p(y|x, \theta) = \exp(\langle \phi(x, y), \theta \rangle - g(\theta|x))$$

Log-partition function

$$g(\theta|x) = \log \int_y \exp(\langle \phi(x, y), \theta \rangle) dy$$

Sufficient Criterion

$p(x, y|\theta)$ is a member of the exponential family itself.

Key Idea

Avoid computing $\phi(x, y)$ directly, only evaluate inner products via

$$k((x, y), (x', y')) := \langle \phi(x, y), \phi(x', y') \rangle$$

Conditional Distributions

Maximum a Posteriori Estimation

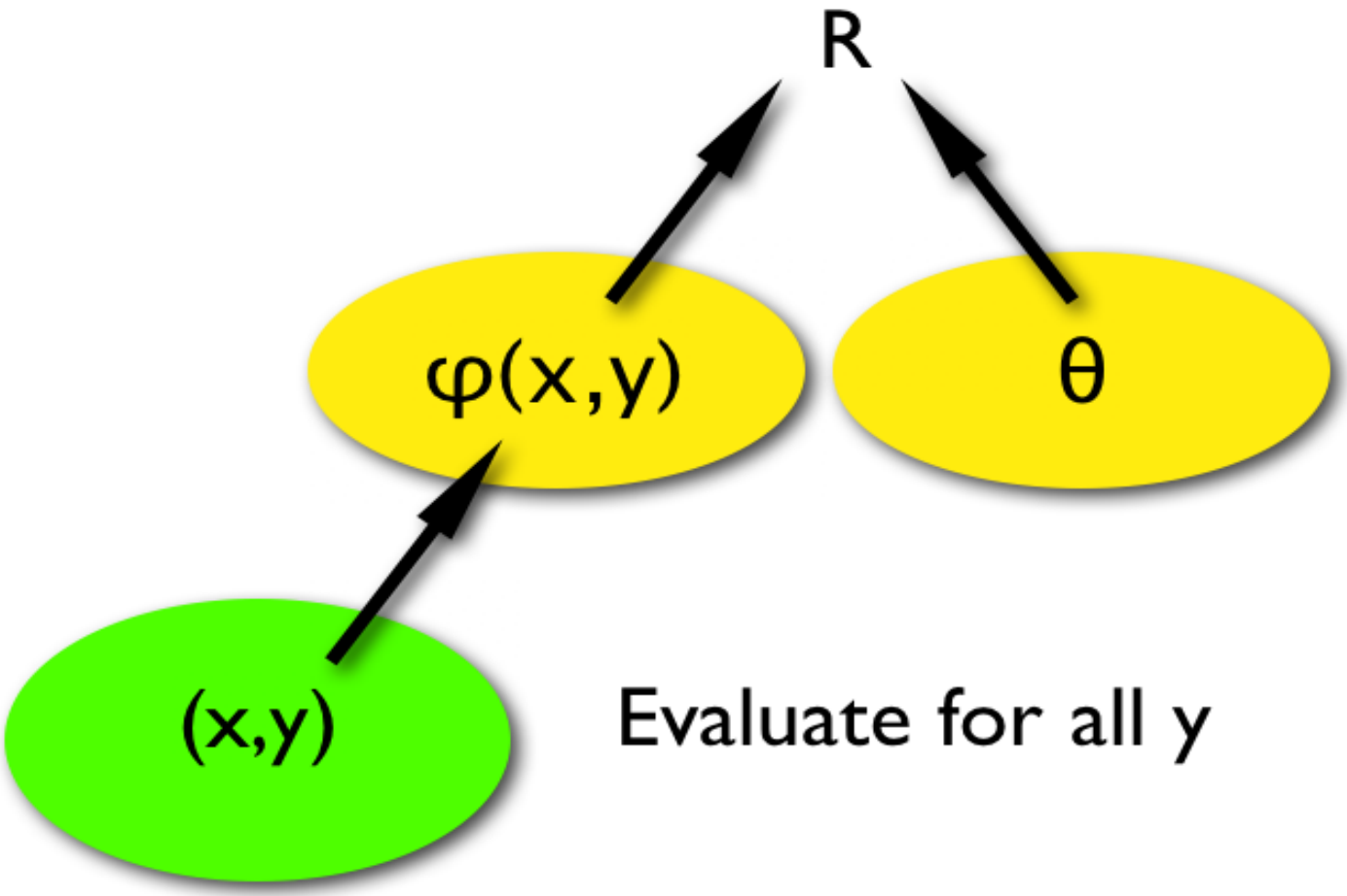
$$-\log p(\theta|X) = \sum_{i=1}^{m} -\langle \phi(x_i), \theta \rangle + m g(\theta) + \frac{1}{2\sigma^2}\|\theta\|^2 + c$$

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

Solving the Problem

- The problem is strictly convex in θ.
- Direct solution is impossible if we cannot compute $\phi(x, y)$ directly.
- Solve convex problem in expansion coefficients.
- Expand θ in a linear combination of $\phi(x_i, y)$.

Joint Feature Map



Representer Theorem

Objective Function

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

Decomposition

- Decompose θ into $\theta = \theta_{\|} + \theta_{\perp}$ where

$$\theta_{\|} \in \operatorname{span}\{\phi(x_i, y) \text{ where } 1 \leq i \leq m \text{ and } y \in \mathcal{Y}\}$$

- Both $g(\theta|x_i)$and $\langle \phi(x_i, y_i), \theta \rangle$are independent of $\theta_{\perp}$.

Theorem

$-\log p(\theta|X,Y)$ is minimized for $\theta_{\perp} = 0$, hence $\theta = \theta_{\|}$.

Consequence

If $\operatorname{span}\{\phi(x_i, y)$ where $1 \leq i \leq m$ and $y \in \mathcal{Y}\}$ is finite dimensional, we have a parametric optimization problem.

Using It

Expansion

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

Inner Product

$$\langle \phi(x, y), \theta \rangle = \sum_{i=1}^{m} \sum_{y \in \mathcal{Y}} \alpha_{iy} k((x, y), (x_i, y))$$

Norm

$$\|\theta\|^2 = \sum_{i,j=1}^{m} \sum_{y,y' \in \mathcal{Y}} \alpha_{iy} \alpha_{jy'} k((x_i, y), (x_j, y'))$$

Log-partition function

$$g(\theta|x) = \log \sum_{y \in \mathcal{Y}} \exp\left(\langle \phi(x, y), \theta \rangle\right)$$

The Gaussian Process Link

Normal Prior on θ . . .

$$\theta \sim \mathcal{N}(0, \sigma^2 \mathbf{1})$$

. . . yields Normal Prior on $t(x, y) = \langle \phi(x, y), \theta \rangle$

- Distribution of projected Gaussian is Gaussian.
- The mean vanishes

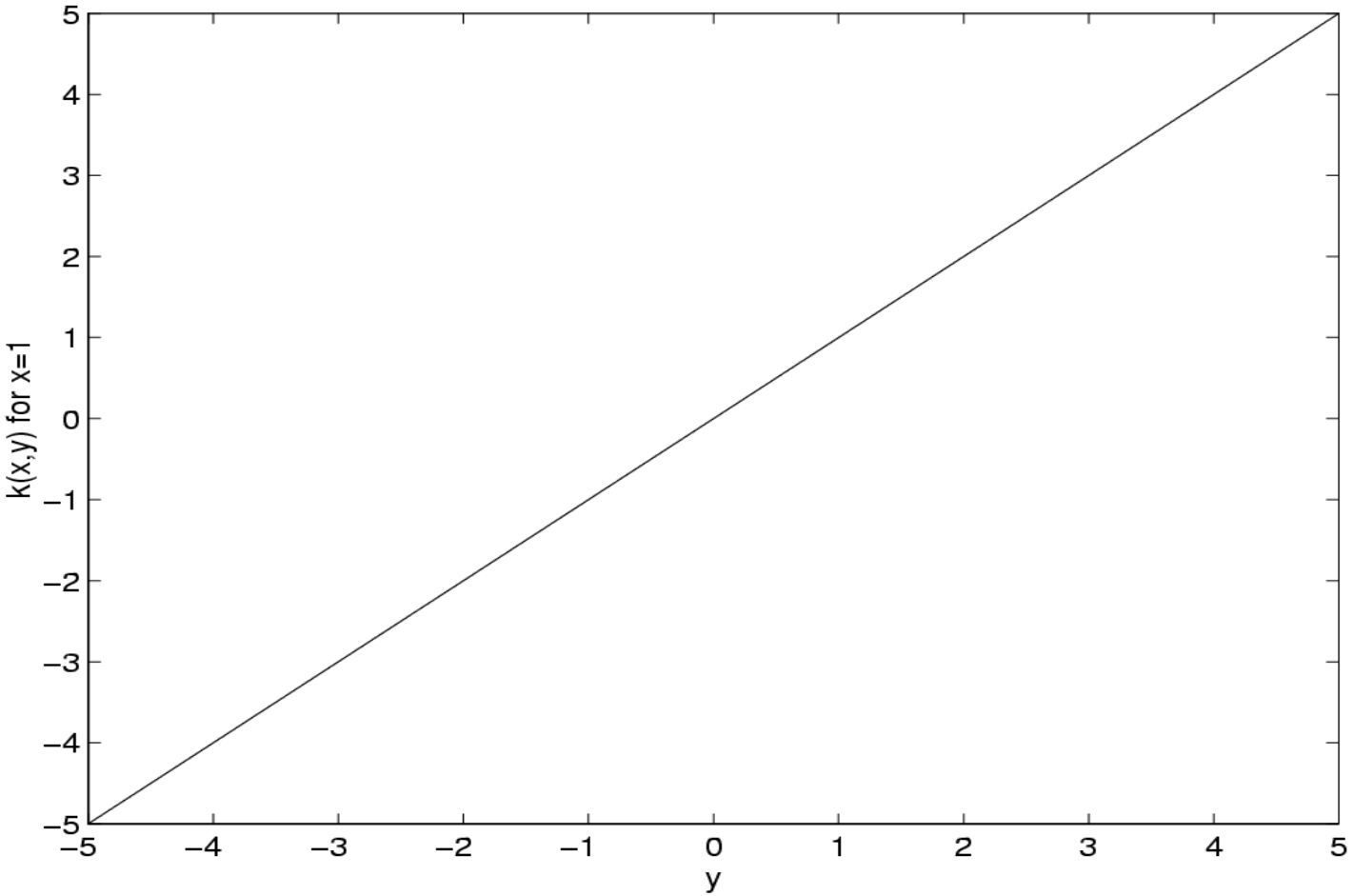
$$\mathbf{E}_\theta[t(x, y)] = \langle \phi(x, y), \mathbf{E}_\theta[\theta] \rangle = 0$$

- The covariance yields

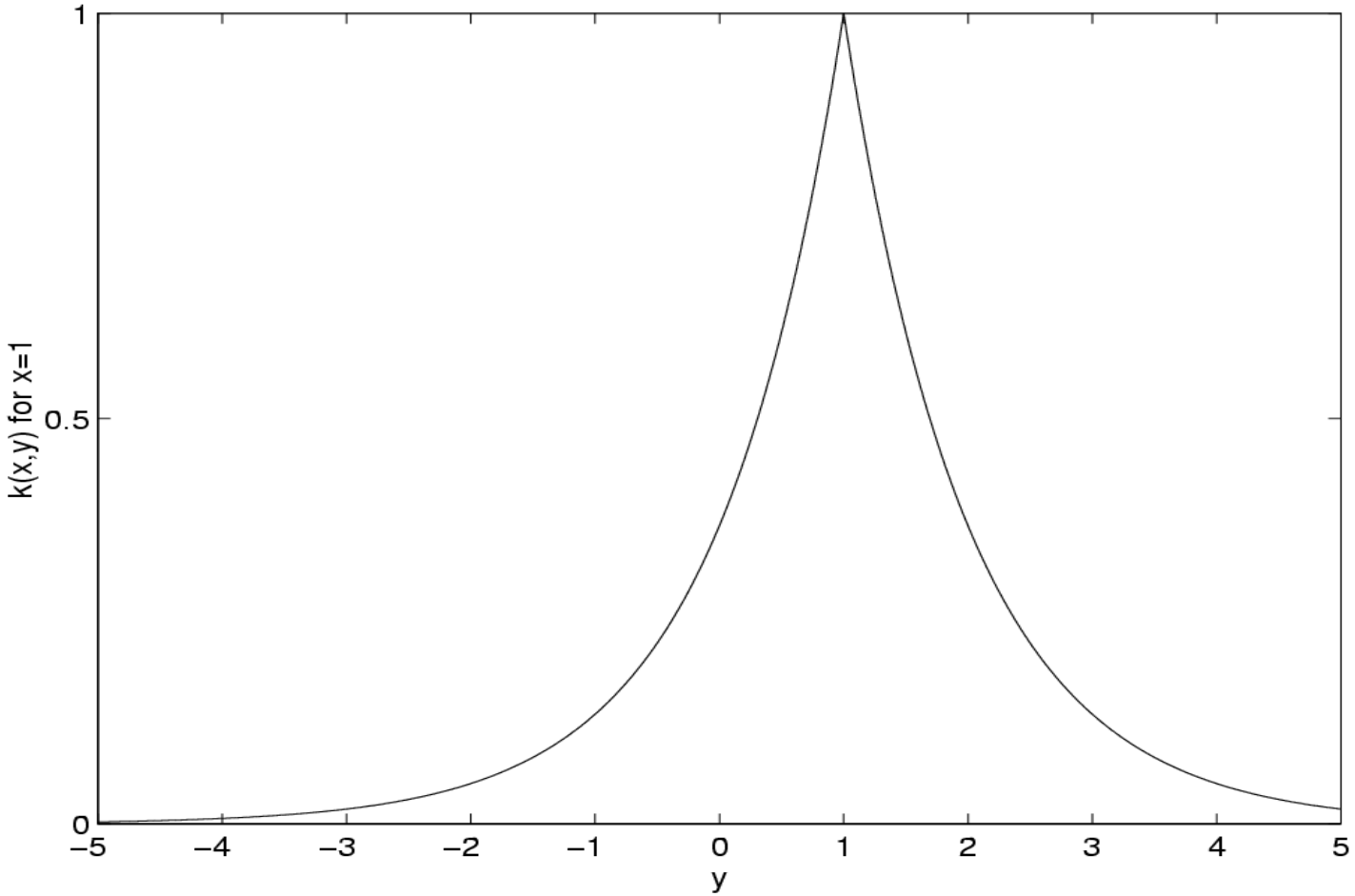
$$\begin{aligned} \mathrm{Cov}[t(x, y), t(x', y')] &= \mathbf{E}_\theta \left[\langle \phi(x, y), \theta \rangle \langle \theta, \phi(x', y') \rangle \right] \\ &= \underbrace{\sigma^2 \langle \phi(x, y), \phi(x', y') \rangle}_{:= k((x,y),(x',y'))} \end{aligned}$$

. . . so we have a Gaussian Process on x . . .
with kernel $k((x, y), (x', y')) = \sigma^2 \langle \phi(x, y), \phi(x', y') \rangle$.

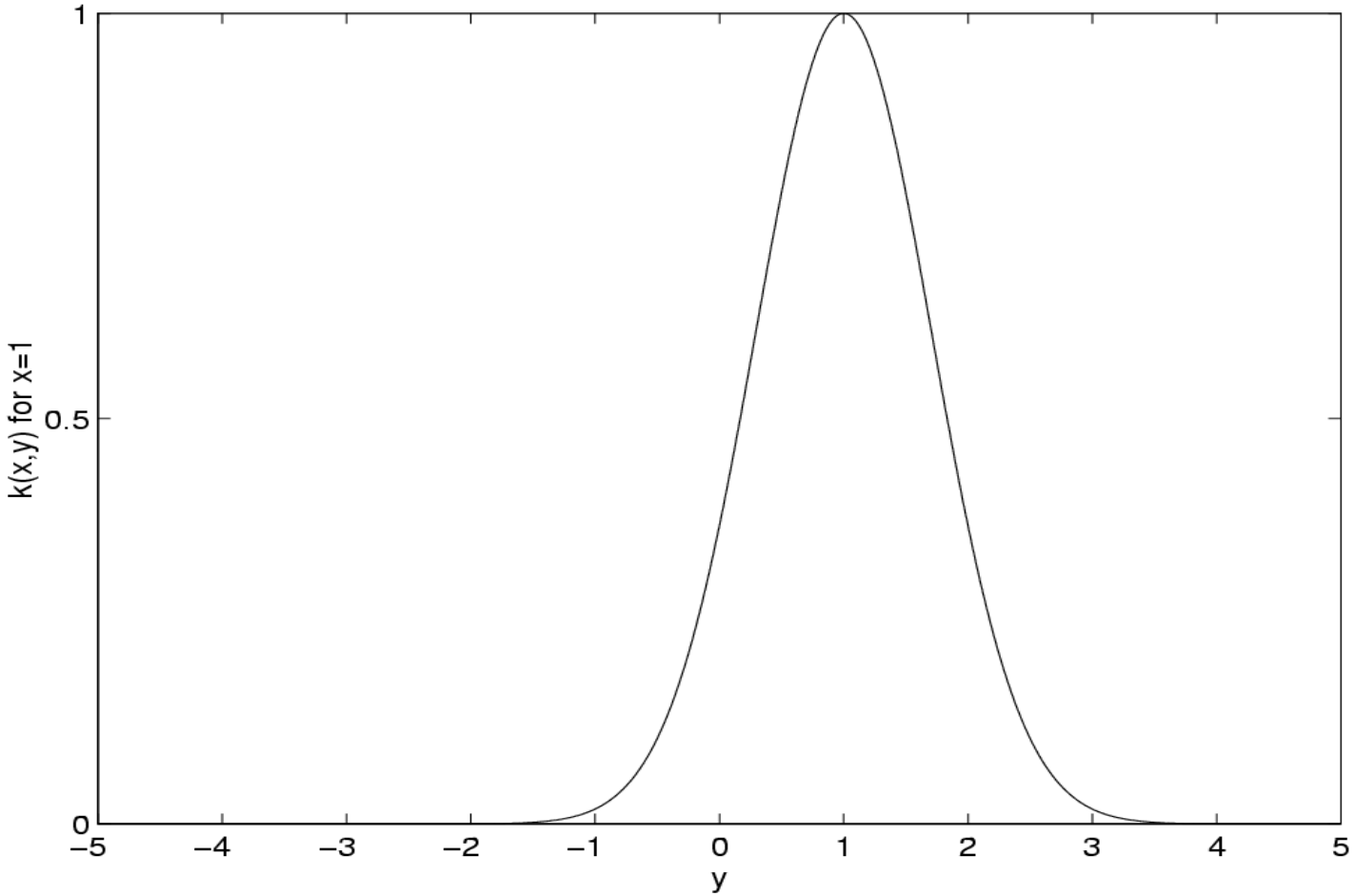
Linear Covariance



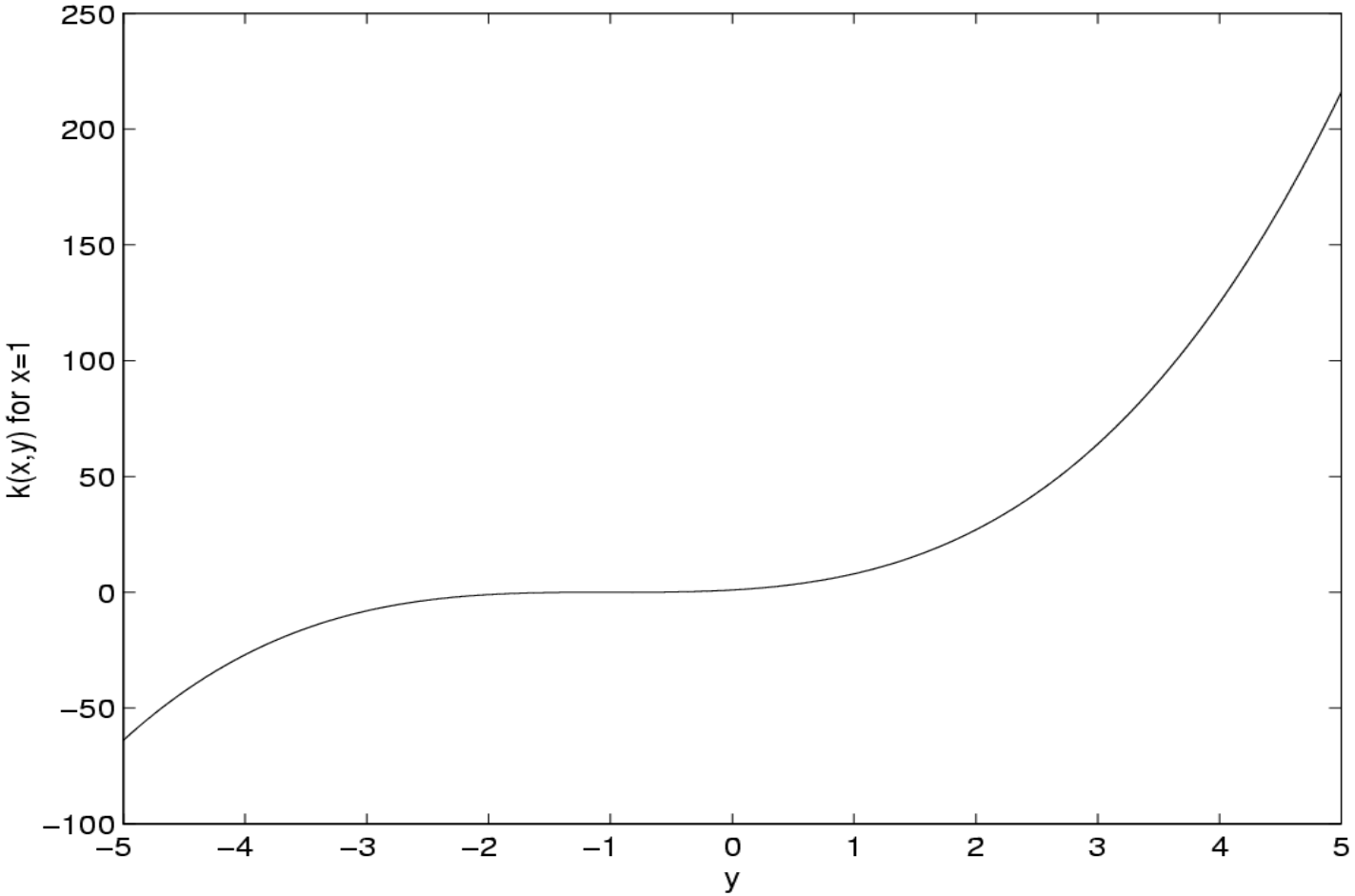
Laplacian Covariance



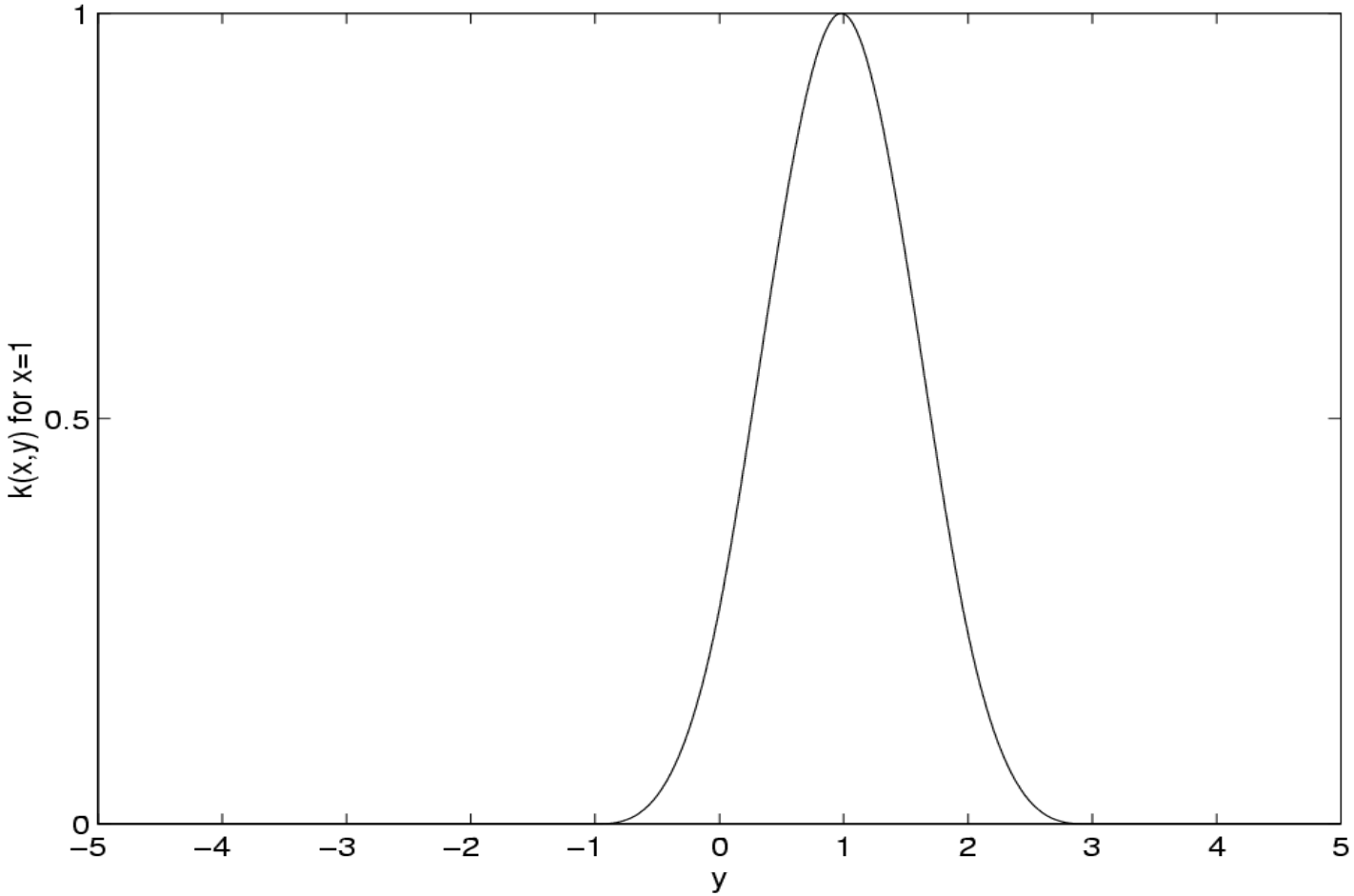
Gaussian Covariance



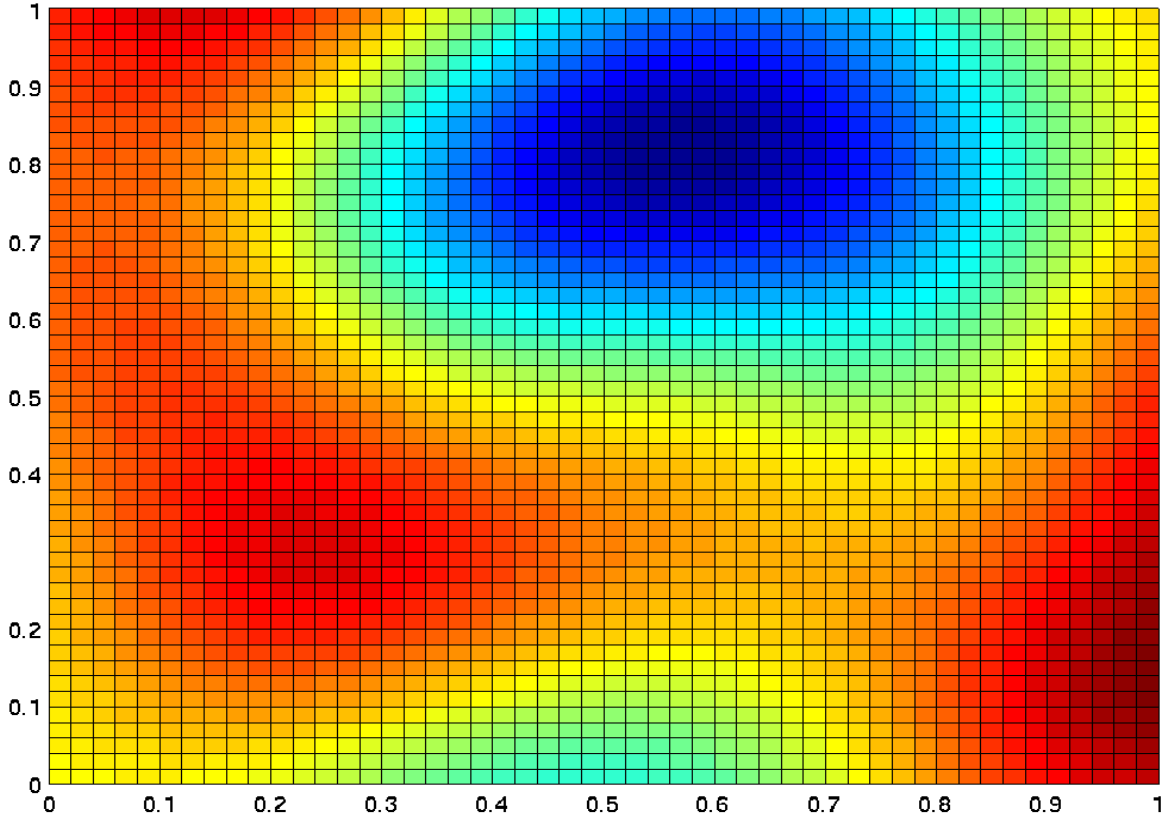
Polynomial (Order 3)



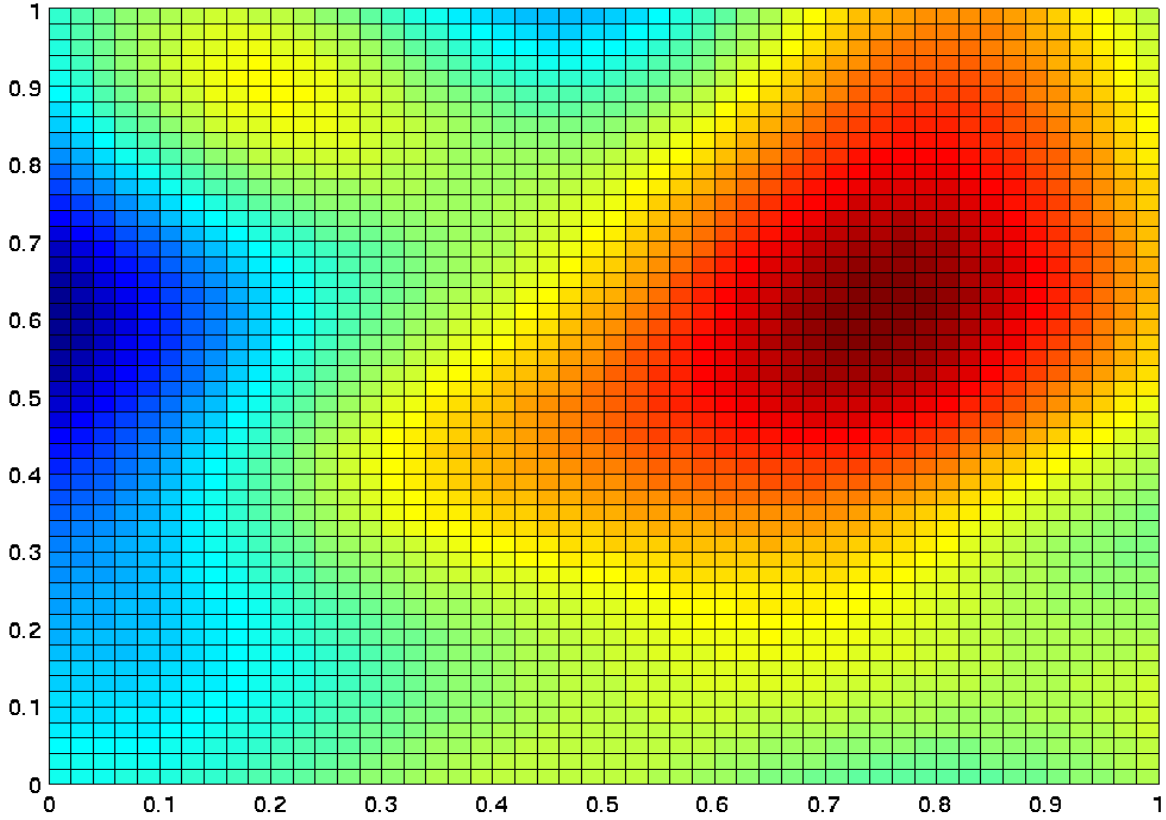
B_3-Spline Covariance



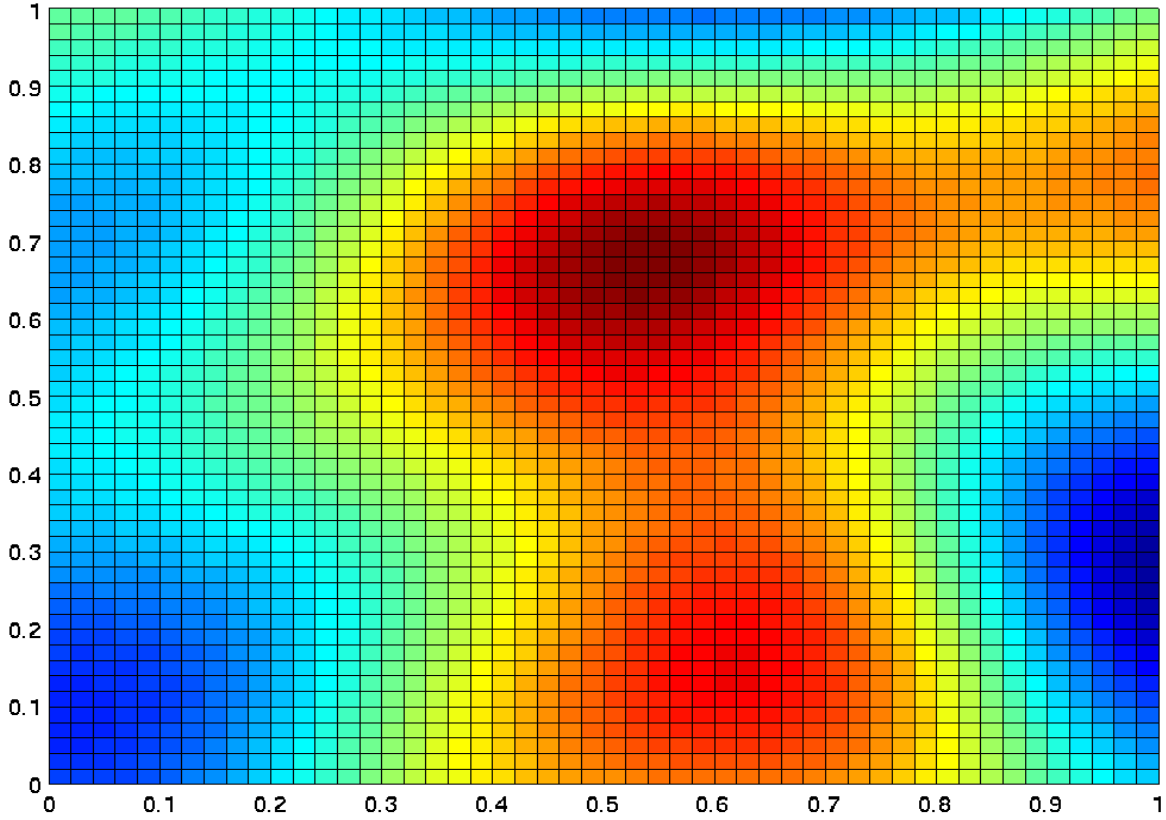
Sample from Gaussian RBF



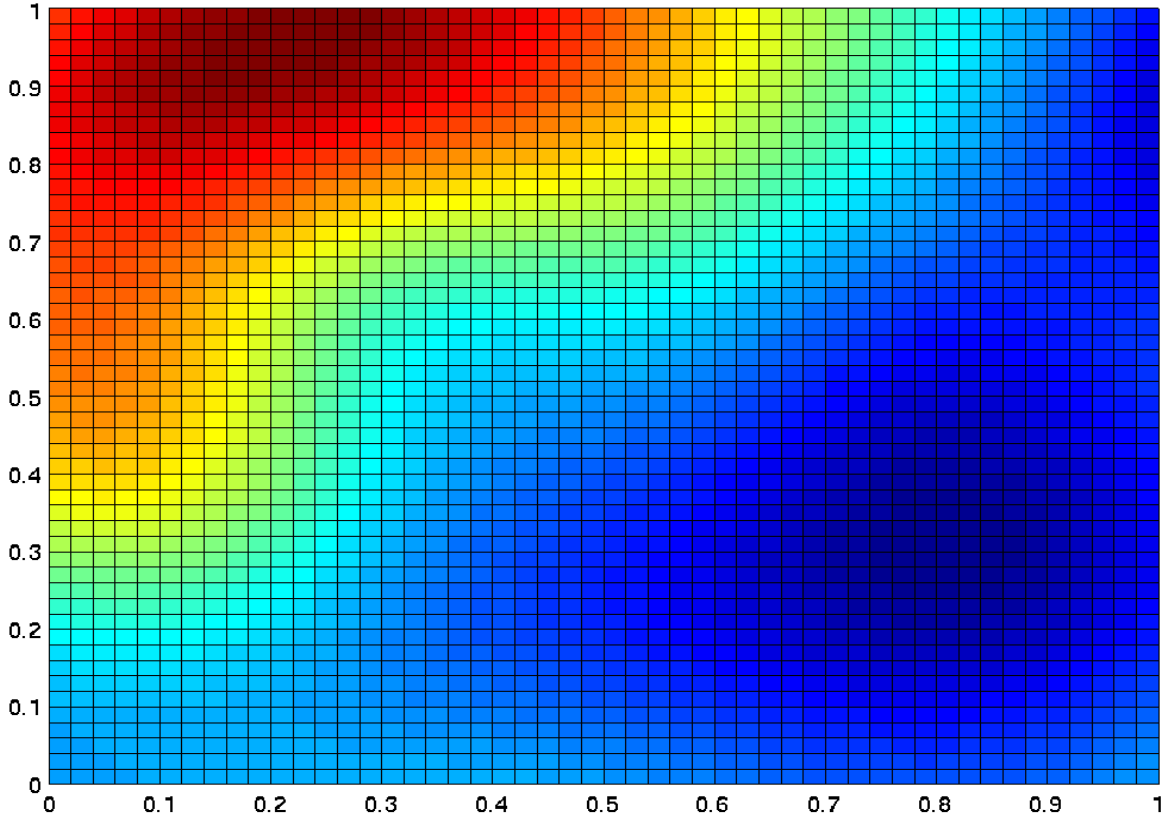
Sample from Gaussian RBF



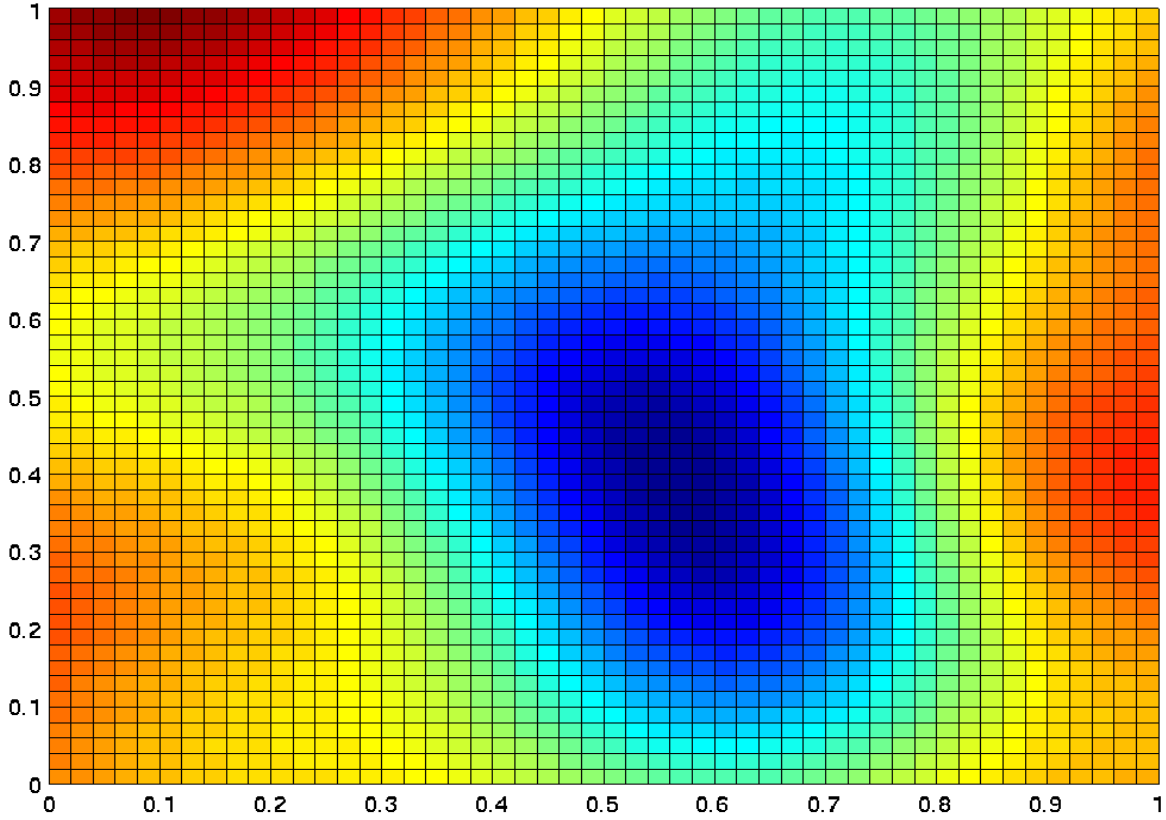
Sample from Gaussian RBF



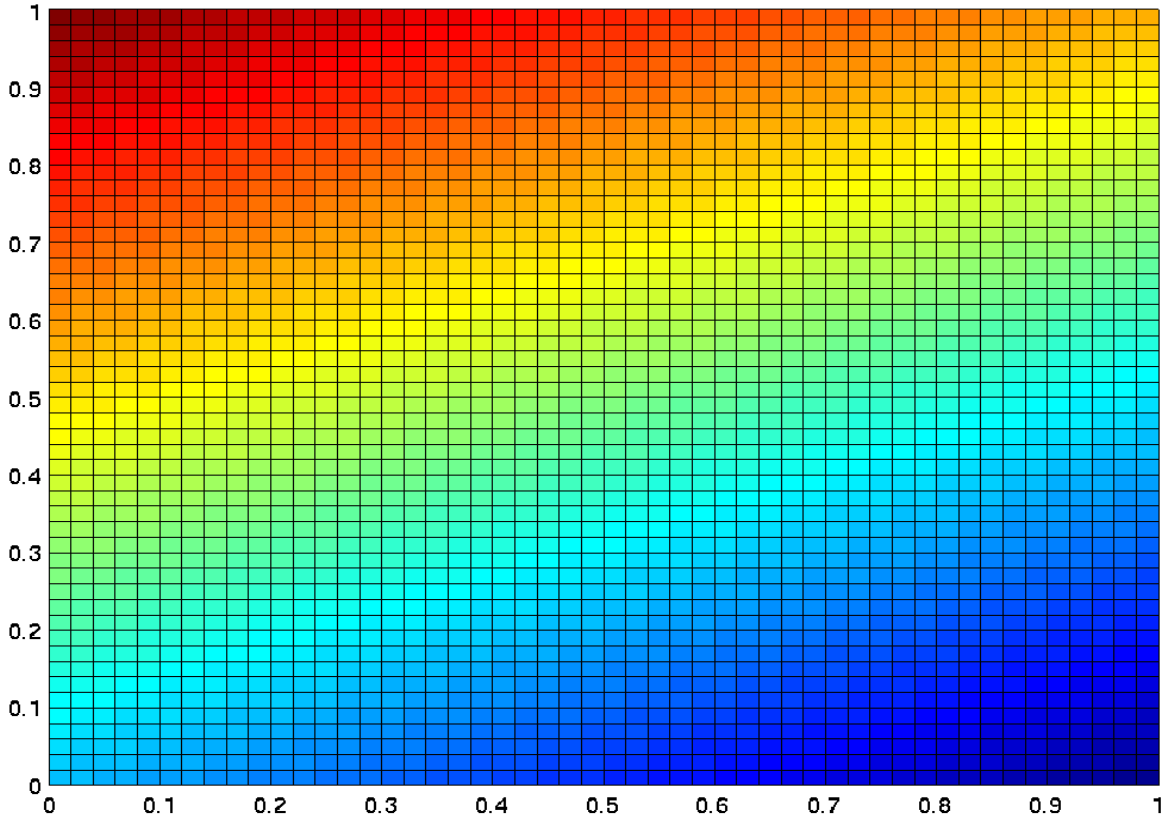
Sample from Gaussian RBF



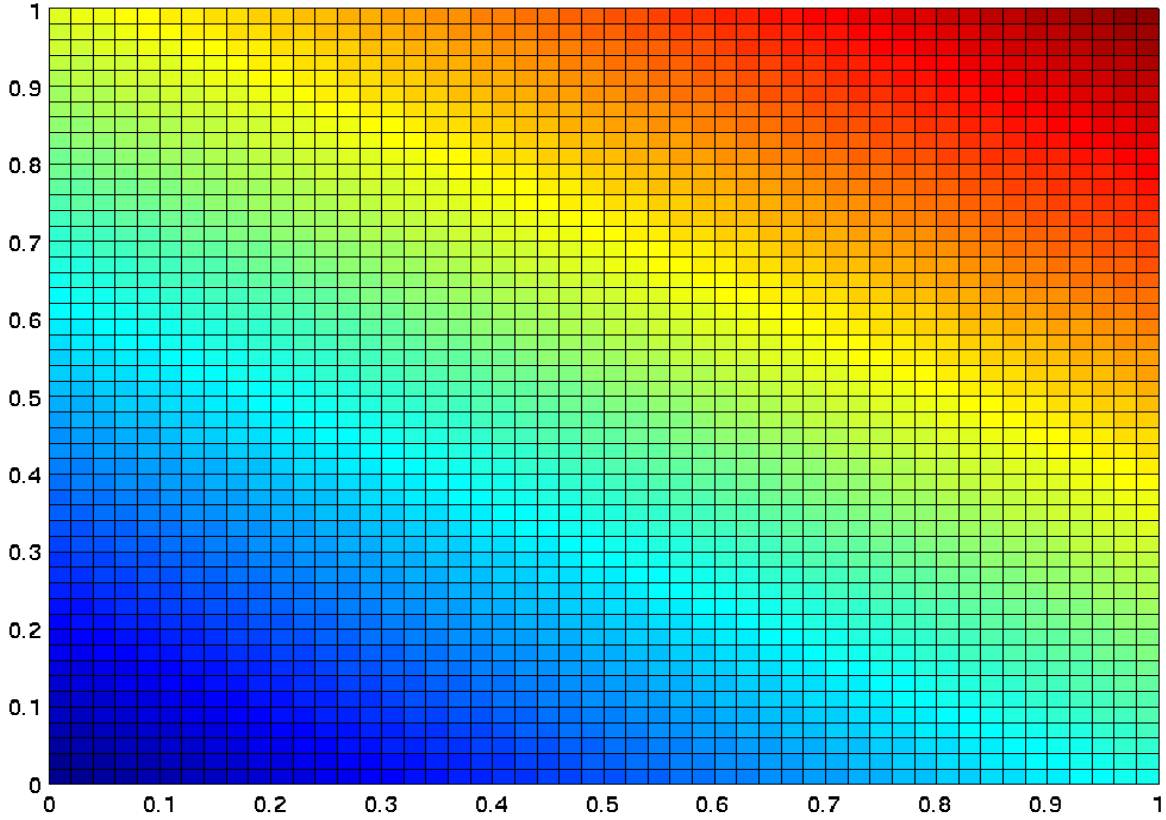
Sample from Gaussian RBF



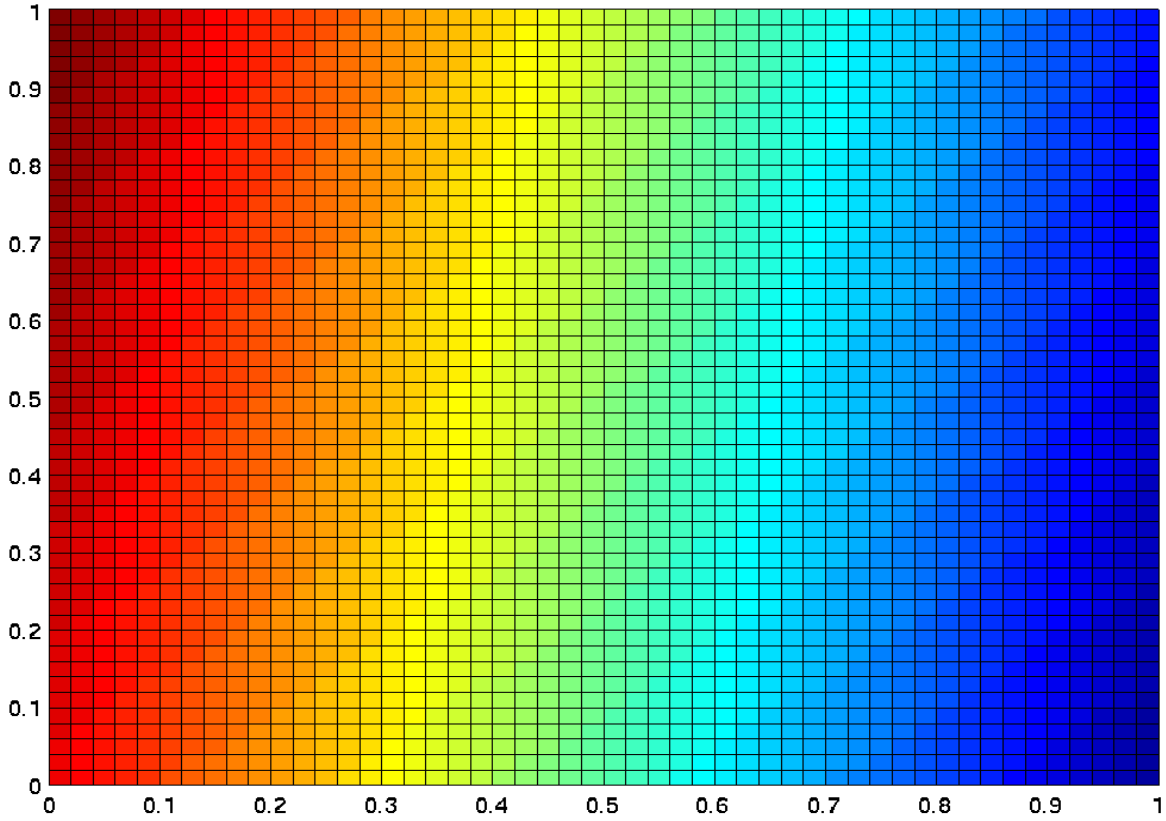
Sample from linear kernel



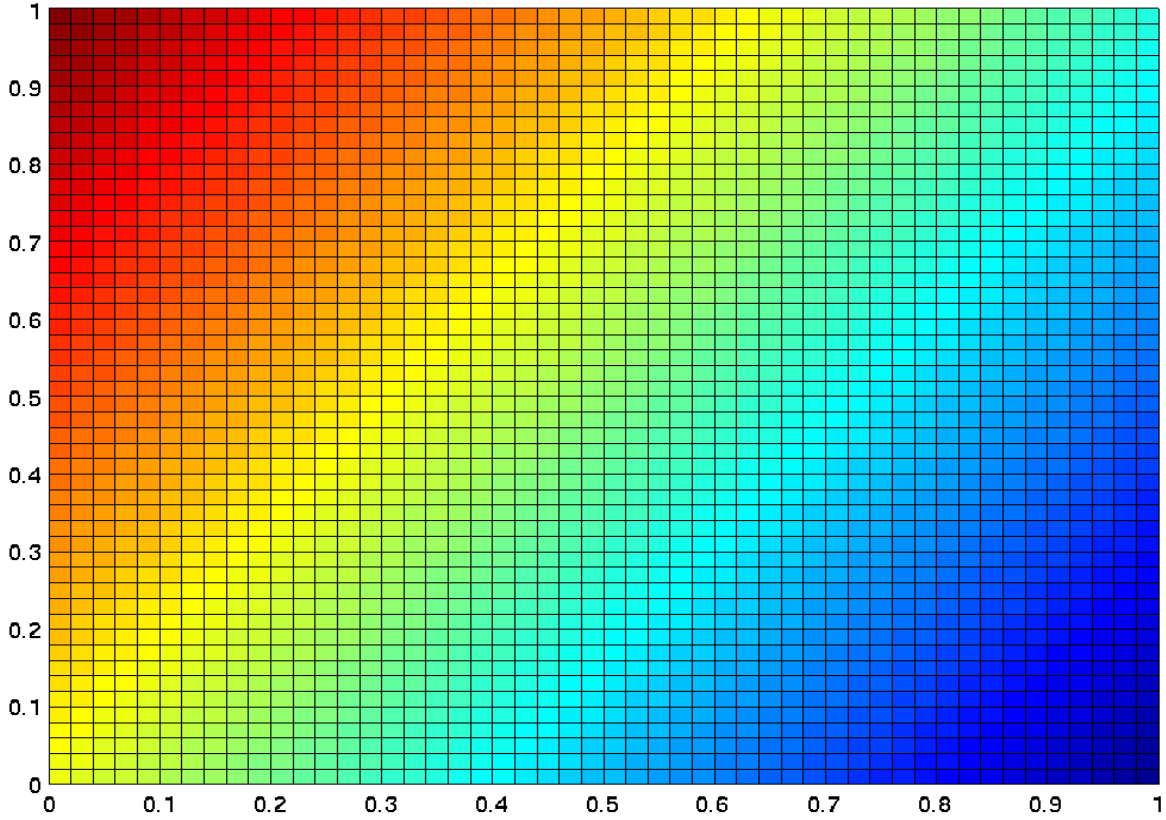
Sample from linear kernel



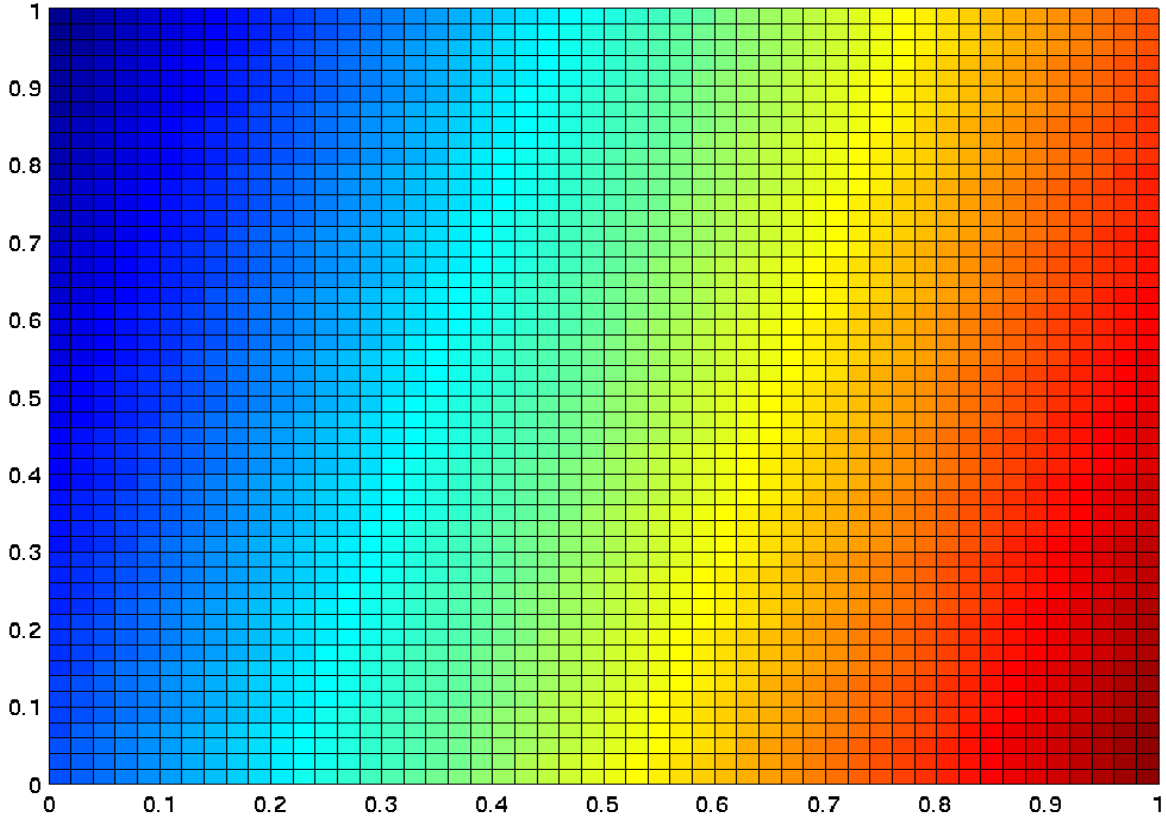
Sample from linear kernel



Sample from linear kernel



Sample from linear kernel



General Strategy

Choose a suitable sufficient statistic $\phi(x, y)$

- Conditionally multinomial distribution leads to Gaussian Process multiclass estimator: we have a distribution over n classes which depends on x.
- Conditionally Gaussian leads to Gaussian Process regression: we have a normal distribution over a random variable which depends on the location.
  Note: we estimate mean and variance.
- Conditionally Poisson distributions yield locally varying Poisson processes. This has no name yet ...

Solve the optimization problem

This is typically convex.

The bottom line

Instead of choosing $k(x, x')$ choose $k((x, y), (x', y'))$.

Example: GP 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

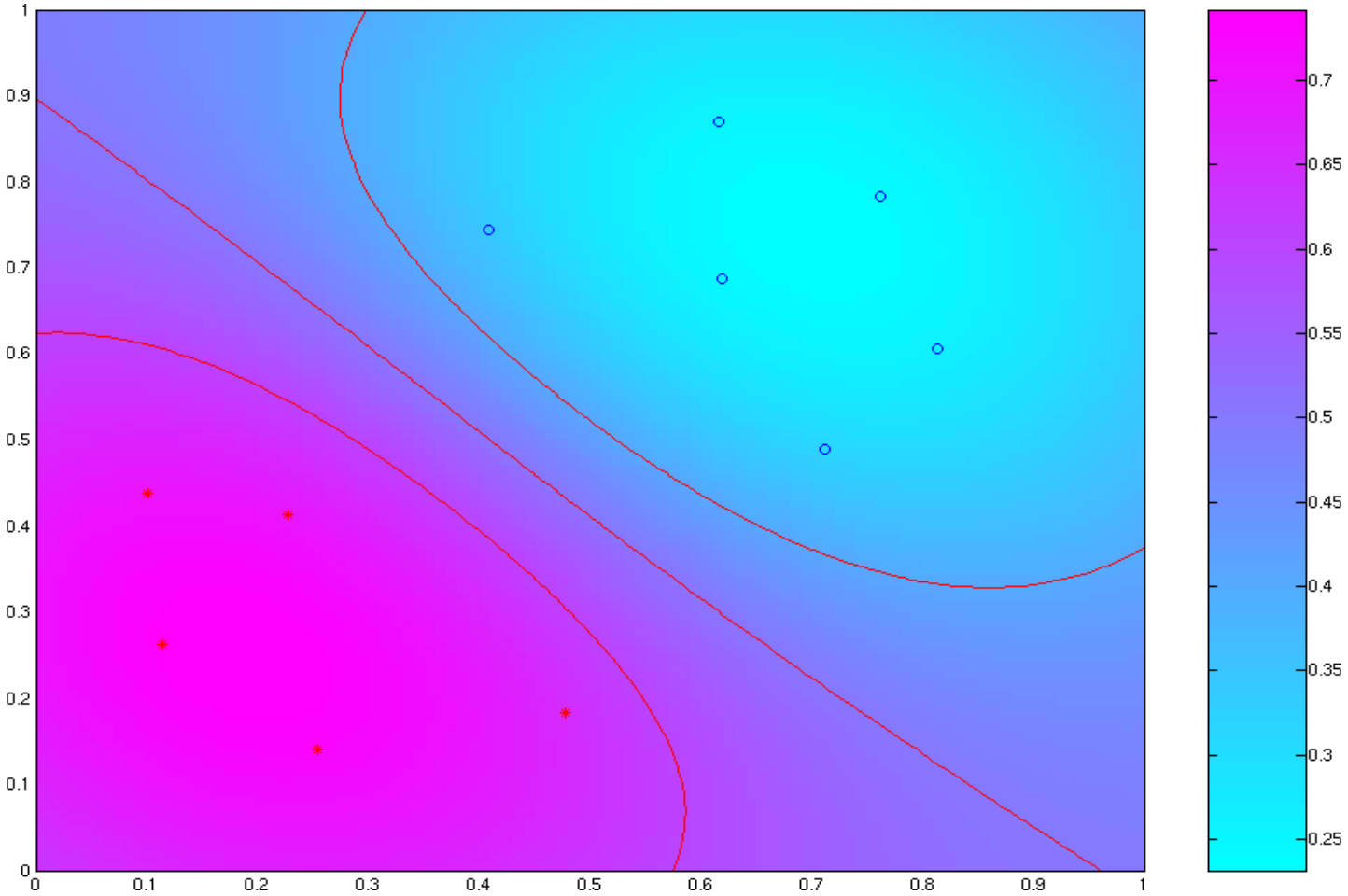
By the representer theorem we get that

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

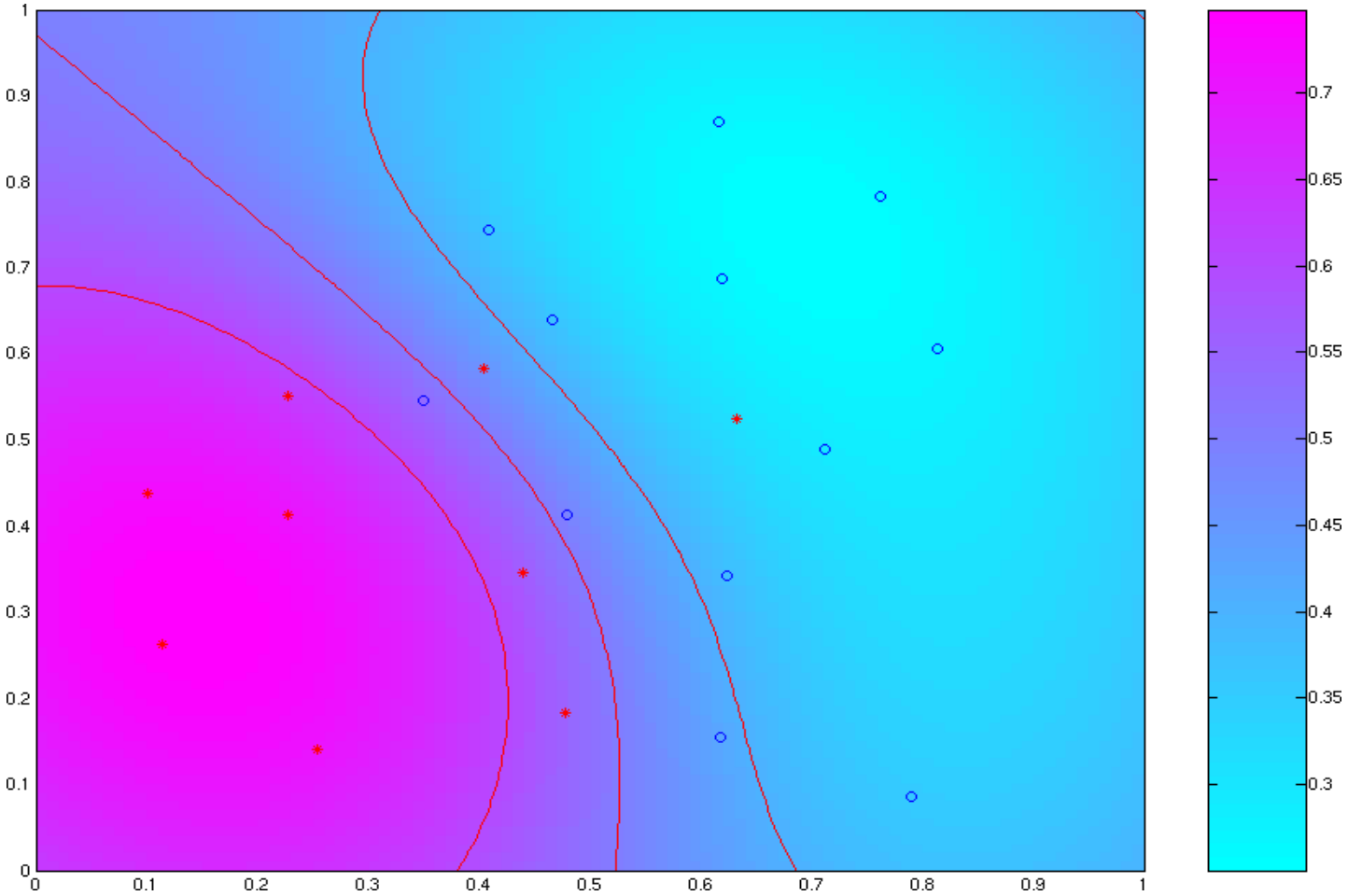
Optimization Problem

Big mess . . . but convex.

A Toy Example



Noisy Data



Summary

Clifford Hammersley Theorem and Graphical Models

- Decomposition results
- Key connection
- Normal distribution

Conditional Distributions

- Log partition function
- Expectations and derivatives
- Inner product formulation and kernels
- Gaussian Processes

Applications

- Generalized kernel trick
- Conditioning gives existing estimation methods back