Regularization Networks

9.520 Class 18, 2004

Tomaso Poggio

Plan

- Radial Basis Functions and their extensions
- Additive Models
- Regularization Networks
- Dual Kernels
- Conclusions

About this class

We describe a family of regularization techniques based on radial kernels K and called RBFs. We introduce RBF extensions (somewhat less rigorous) such as Hyper Basis Functions and characterize their relation with other techniques including MLPs and splines.

Radial Basis Functions

Radial Basis Functions, as MLPs, have the universal approximation property.

Theorem: Let K be a Radial Basis Function function and I_i the n-dimensional cube $[0, 1]^n$. Then finite sums of the form

$$f(\mathbf{x}) = \sum_{i=1}^{N} c_i K(\mathbf{x} - \mathbf{x}_i)$$

are dense in $C[I_i]$. In other words, given a function $h \in C[I_i]$ and $\epsilon > 0$, there is a sum, $f(\mathbf{x})$, of the above form, for which:

$$|f(\mathbf{x}) - h(\mathbf{x})| < \epsilon \quad \text{for all} \quad \mathbf{x} \in I_n \ .$$

Notice that RBF correspond to RKHS defined on an infinite domain. Notice also that RKHS do not in general have the same approximation property: RKHS generated by a K with an infinite countable number of strictly positive eigenvalues are dense in L_2 but not necessarily in $C(X)$, though they can be embedded in $C(X)$. See Zhou results.

Density of a RKHS on a bounded domain (the non-RBF case)

We first ask under which condition is a RKHS dense in $L_2(X, \nu)$.

1. when L_K is *strictly positive* the RKHS is infinite dimensional and dense in $L_2(X, \nu)$.

2. in the *degenerate case* the RKHS is finite dimensional and not dense in $L_2(X, \nu)$.

3. in the *conditionally strictly positive case* the RKHS is not dense in $L_2(X, \nu)$ but when completed with a finite number of polynomials of appropriate degree can be made to be dense in $L_2(X, \nu)$.

Density of a RKHS on a bounded domain (cont)

Density of RKHS – defined on a compact domain X – in $C(X)$ (in the sup norm) is a trickier issue that has been answered very recently by Zhou (in preparation). It is however guaranteed for radial kernels K for K continuous and integrable, if density in $L_2(X, \nu)$ holds (with X the infinite domain). These are facts for radial kernels and unrelated to RKHS properties

- span $K(x-y) : y \in R^n$ is dense in $L^2(R^n)$ iff the Fourier transform of K goes not vanish on set of positive Lebesque measure (N. Wiener).

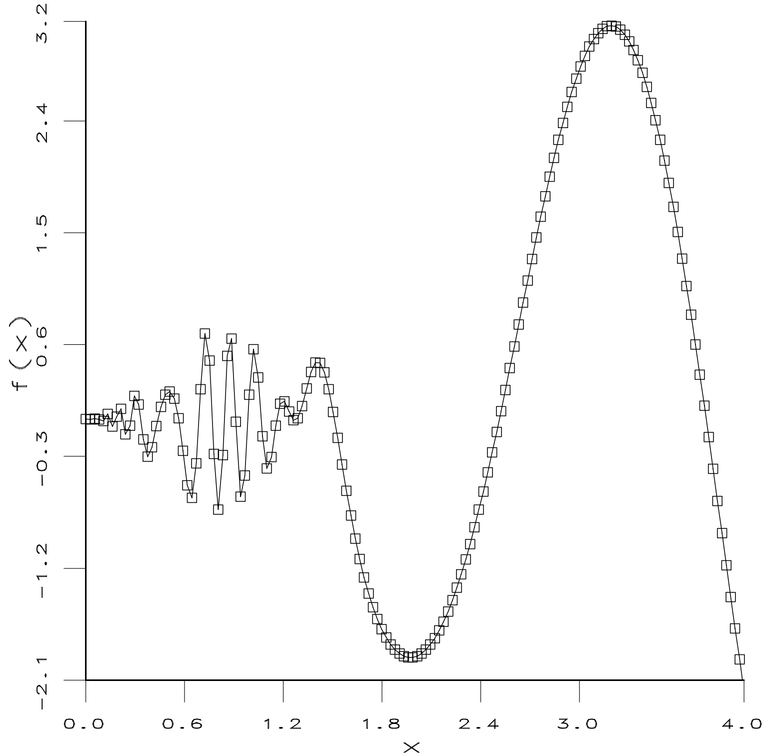
- span $K(x-y) : y \in R^n$ is dense in $C(R^n)$ (topology of uniform convergence) if $K \in C(R^n)$, $K \in L^1(R^n)$.

Some good properties of RBF

- Well motivated in the framework of regularization theory;
- The solution is unique and equivalent to solving a linear system;
- Degree of smoothness is tunable (with λ);
- Universal approximation property;
- Large body of applied math literature on the subject;
- Interpretation in terms of *neural networks*(?!);
- Biologically plausible;
- Simple interpretation in terms of *smooth look-up table*;
- Similar to other non-parametric techniques, such as nearest neighbor and kernel regression (see end of this class).

Some not-so-good properties of RBF

- Computationally expensive ($O(\ell^3)$);
- Linear system to be solved for finding the coefficients often badly ill-conditioned;
- The same degree of smoothness is imposed on different regions of the domain (we will see how to deal with this problem in the class on wavelets);



This function has different smoothness properties in different regions of its domain.

A first extension: less centers than data points

We look for an *approximation* to the regularization solution:

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} c_i K(\mathbf{x} - \mathbf{x}_i)$$

$$\Downarrow$$

$$f^*(\mathbf{x}) = \sum_{\alpha=1}^{m} c_\alpha K(\mathbf{x} - \mathbf{t}_\alpha)$$

where $m << \ell$ and the vectors $\mathbf{t}_\alpha$ are called **centers**.

Homework: show that the interpolation problem is still well-posed when $m < \ell$.

(Broomhead and Lowe, 1988; Moody and Darken, 1989; Poggio and Girosi, 1989)

Least Squares Regularization Networks

$$f^*(\mathbf{x}) = \sum_{\alpha=1}^{m} c_\alpha K(\mathbf{x} - \mathbf{t}_\alpha)$$

Suppose the centers $\mathbf{t}_\alpha$ have been fixed.

How do we find the coefficients c_α?

$$\Downarrow$$

Least Squares

Finding the coefficients

Define

$$E(c_1, \ldots, c_m) = \sum_{i=1}^{\ell} (y_i - f^*(\mathbf{x}_i))^2$$

The least squares criterion is

$$\min_{c_\alpha} \; E(c_1, \ldots, c_m)$$

The problem is convex and quadratic in the c_α, and the solution satisfies:

$$\frac{\partial E}{\partial c_\alpha} = 0$$

Finding the centers

Given the centers $\mathbf{t}_\alpha$ we know how to find the c_α.

How do we choose the $\mathbf{t}_\alpha$?

1. a subset of the examples (random);
2. by a clustering algorithm (k-means, for example);
3. by least squares (*moving centers*);
4. a subset of the examples: Support Vector Machines;

Centers as a subset of the examples

Fair technique. The subset is a random subset, which should reflect the distribution of the data.

Not many theoretical results available (but we proved that solution exists since matrix is till pd).

Main problem: how many centers?

Main answer: we don't know. Cross validation techniques seem a reasonable choice.

Finding the centers by clustering

Very common. However it makes sense only if the input data points are clustered.

No theoretical results.

Not clear that it is a good idea, especially for pattern classification cases.

Moving centers

Define

$$E(c_1, \dots, c_m, \mathbf{t}_1, \dots, \mathbf{t}_m) = \sum_{i=1}^{\ell} (y_i - f^*(\mathbf{x}_i))^2$$

The least squares criterion is

$$\min_{c_\alpha, \mathbf{t}_\alpha} \; E(c_1, \dots, c_m, \mathbf{t}_1, \dots, \mathbf{t}_m).$$

The problem is not convex and quadratic anymore: expect multiple local minima.

Moving centers

:-) Very flexible, in principle very powerful (more than SVMs);

:-) Some theoretical understanding;

:-(Very expensive computationally due to the local minima problem;

:-(Centers sometimes move in “weird” ways;

Connection with MLP

Radial Basis Functions with moving centers is a particular case of a function approximation technique of the form:

$$f(\mathbf{x}) = \sum_{i=1}^{N} c_i H(\mathbf{x}, \mathbf{p}_i)$$

where the parameters $\mathbf{p}_i$ can be estimated by least squares techniques.

Radial Basis Functions corresponds to the choice $N = m$ and $\mathbf{p}_i = \mathbf{t}_i$, and

$$H(\mathbf{x}, \mathbf{p}_i) = K(\|\mathbf{x} - \mathbf{t}_i\|)$$

Extensions of Radial Basis Functions (much beyond what SVMs can do!!

- Different variables can have different scales: $f(x, y) = y^2 \sin(100x)$;

- Different variables could have different units of measure $f = f(\mathbf{x}, \dot{\mathbf{x}}, \ddot{\mathbf{x}})$;

- Not all the variables are independent or relevant: $f(x, y, z, t) = g(x, y, z(x, y))$;

- Only some linear combinations of the variables are relevant: $f(x, y, z) = \sin(x + y + z)$;

Extensions of regularization theory

A priori knowledge:

- the relevant variables are linear combination of the original ones:

$$\mathbf{z} = W\mathbf{x}$$

for some (possibly rectangular) matrix W;

- $f(\mathbf{x}) = g(W\mathbf{x}) = g(\mathbf{z})$ and the function g is smooth;

The regularization functional is now

$$\sum_{i=1}^{\ell} (y_i - g(\mathbf{z}_i))^2 + \lambda \Phi[g]$$

where $\mathbf{z}_i = W\mathbf{x}_i$.

Extensions of regularization theory (continue)

The solution is

$$g(\mathbf{z}) = \sum_{i=1}^{\ell} c_i K(\mathbf{z} - \mathbf{z}_i) \ .$$

Therefore the solution for f is:

$$f(\mathbf{x}) = g(W\mathbf{x}) = \sum_{i=1}^{\ell} c_i K(W\mathbf{x} - W\mathbf{x}_i)$$

Extensions of regularization theory (continue)

If the matrix W were known, the coefficients could be computed as in the radial case:

$$(K + \lambda I)\mathbf{c} = \mathbf{y}$$

where

$$(\mathbf{y})_i = y_i \ , \quad (\mathbf{c})_i = c_i \ , (K)_{ij} = K(W\mathbf{x}_i - W\mathbf{x}_j)$$

and the same argument of the Regularization Networks technique apply, leading to *Generalized Regularization Networks*:

$$f^*(\mathbf{x}) = \sum_{\alpha=1}^{m} c_\alpha K(W\mathbf{x} - W\mathbf{t}_\alpha)$$

Extensions of regularization theory (continue)

Since W is usually not known, it could be found by *least squares*. Define

$$E(c_1, ..., c_m, W) = \sum_{i=1}^{\ell} (y_i - f^*(\mathbf{x}_i))^2$$

Then we can solve:

$$\min_{c_\alpha, W} \; E(c_1, ..., c_m, W)$$

The problem is not convex and quadratic anymore: expect multiple local minima.

From RBF to HyperBF

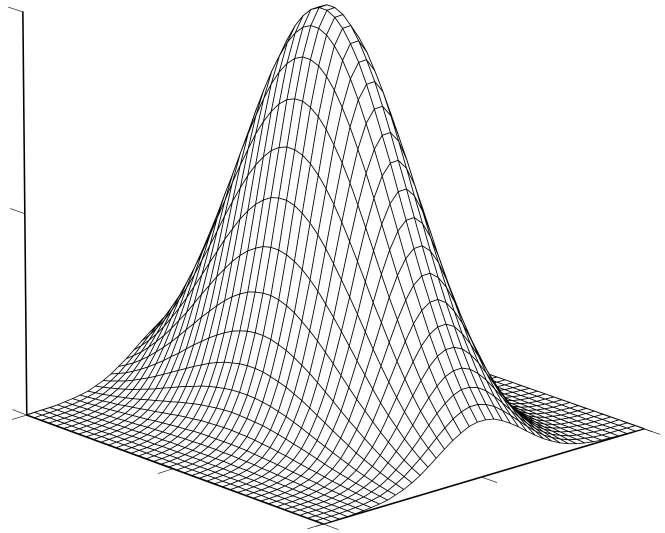
When the basis function K is radial the Generalized Regularization Networks becomes

$$f(\mathbf{x}) = \sum_{\alpha=1}^{m} c_\alpha K(\|\mathbf{x} - \mathbf{t}_\alpha\|_w)$$

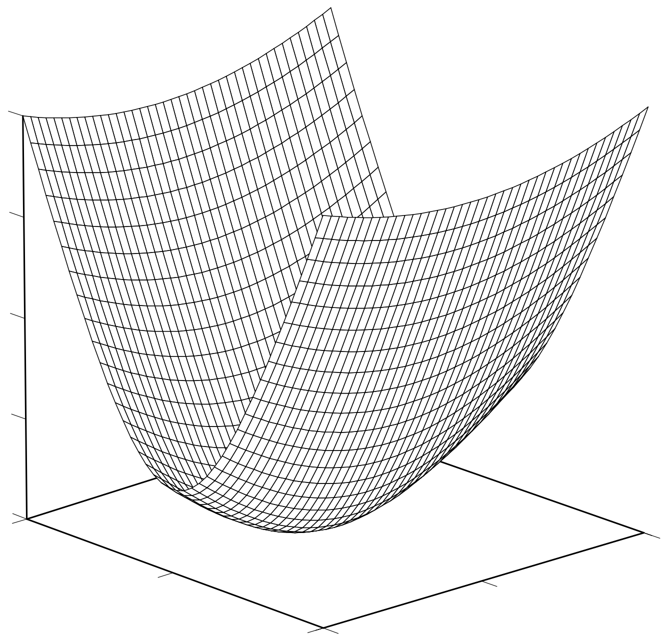
that is a *non radial basis function* technique (we define $|\mathbf{x}|^2 = \mathbf{x}^\top W^\top W \mathbf{x}$).

Least Squares

1. $\min_{c_\alpha} \; E(c_1, \ldots, c_m)$
2. $\min_{c_\alpha, \mathbf{t}_\alpha} \; E(c_1, \ldots, c_m, \mathbf{t}_1 \ldots, \mathbf{t}_m)$
3. $\min_{c_\alpha, W} \; E(c_1, ..., c_m, W)$
4. $\min_{c_\alpha, \mathbf{t}_\alpha, W} E(c_1, \ldots, c_m, \mathbf{t}_1, \ldots, \mathbf{t}_m, W)$



A nonradial Gaussian function



A nonradial multiquadric function

Additive models

In statistics an additive model has the form

$$f(\mathbf{x}) = \sum_{\mu=1}^{d} f_\mu(x^\mu)$$

where

$$f_\mu(x^\mu) = \sum_{i=1}^{\ell} c_i^\mu G(x^\mu - x_i^\mu)$$

In other words

$$f(\mathbf{x}) = \sum_{\mu=1}^{d} \sum_{i=1}^{\ell} c_i^\mu G(x^\mu - x_i^\mu)$$

Additive stabilizers

To obtain an approximation of the form

$$f(\mathbf{x}) = \sum_{\mu=1}^{d} f_\mu(x^\mu)$$

we choose a stabilizer corresponding to an additive basis function

$$K(\mathbf{x}) = \sum_{\mu=1}^{d} \theta_\mu K(x^\mu)$$

This scheme leads to an approximation scheme of the additive form with

$$f_\mu(x^\mu) = \theta_\mu \sum_{i=1}^{\ell} c_i K(x^\mu - x_i^\mu)$$

Notice that the additive components are not independent since there is only one set of c_i – which makes sense since I have only l data points to determine the c_i.

Extensions of Additive Models

We start from the non-independent additive component formulation obtained from additive stabilizers

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} c_i \sum_{\mu=1}^{d} \theta_\mu K(x^\mu - x_i^\mu)$$

We assume now that the parameters θ_μ are free, eg different for each i. We now have to fit

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} \sum_{\mu=1}^{d} c_i^\mu K(x^\mu - x_i^\mu)$$

with $\ell \times d$ independent c_i^μ. In order to avoid overfitting we reduce the number of centers ($m << l$):

$$f(\mathbf{x}) = \sum_{\mu=1}^{d} \sum_{\alpha=1}^{m} c_\alpha^\mu K(x^\mu - t_\alpha^\mu)$$

Extensions of Additive Models

If we now allow for an arbitrary linear transformation of the inputs:

$$\mathbf{x} \rightarrow W\mathbf{x}$$

where W is a $d' \times d$ matrix, we obtain:

$$f(\mathbf{x}) = \sum_{\mu=1}^{d'} \sum_{\alpha=1}^{m} c_\alpha^\mu K(\mathbf{x}^\top \mathbf{w}_\mu - t_\alpha^\mu)$$

where $\mathbf{w}_\mu$ is the μ-th row of the matrix W.

Extensions of Additive Models

The expression

$$f(\mathbf{x}) = \sum_{\mu=1}^{d'} \sum_{\alpha=1}^{m} c_\alpha^\mu K(\mathbf{x}^\top \mathbf{w}_\mu - t_\alpha^\mu)$$

can be written as

$$f(\mathbf{x}) = \sum_{\mu=1}^{d'} h_\mu(\mathbf{x}^\top \mathbf{w}_\mu)$$

where

$$h_\mu(y) = \sum_{\alpha=1}^{m} c_\alpha^\mu K(y - t_\alpha^\mu)$$

This form of approximation is called **ridge approximation**

Gaussian MLP network

From the extension of additive models we can therefore justify an approximation technique of the form

$$f(\mathbf{x}) = \sum_{\mu=1}^{d'} \sum_{\alpha=1}^{m} c_\alpha^\mu G(\mathbf{x}^\top \mathbf{w}_\mu - t_\alpha^\mu)$$

Particular case: $m = 1$ (one center per dimension). Then we derive the following technique:

$$f(\mathbf{x}) = \sum_{\mu=1}^{d'} c^\mu G(\mathbf{x}^\top \mathbf{w}_\mu - t_\mu)$$

which is a Multilayer Perceptron with a Radial Basis Functions G instead of the sigmoid function. One can argue rather formally that for normalized inputs the weight vectors of MLPs are equivalent to the centers of RBFs.

Notice that the sigmoid function cannot be derived – directly and formally – from regularization but...

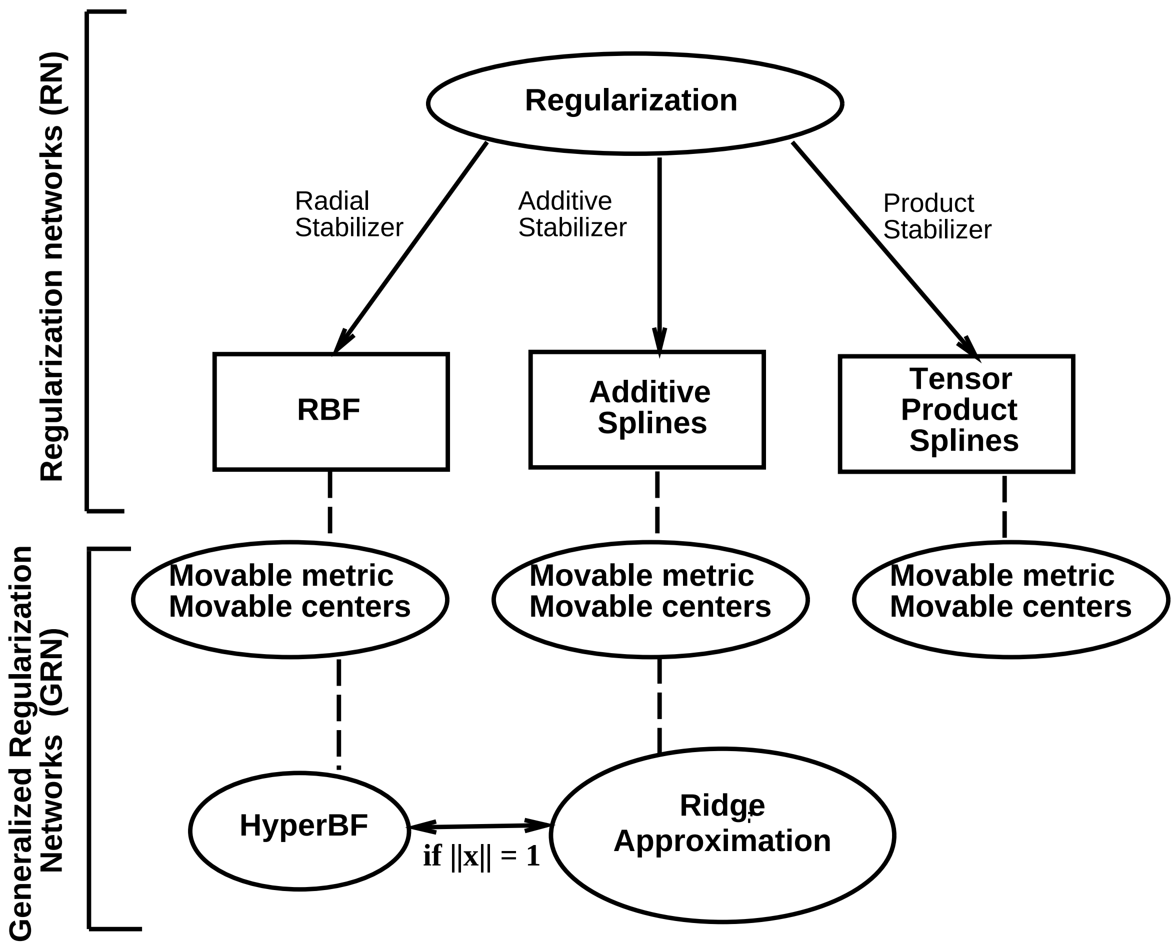
Sigmoids and Regularization

Suppose to have learned the representation

$$f(\mathbf{x}) = \sum_{\mu=1}^{d'} c^{\mu} K'(\mathbf{x}^{\top}\mathbf{w}_{\mu} - t_{\mu})$$

where $K'(x) = |x|$. Notice that a finite linear combination of translates of a sigmoidal, piece-wise linear basis function can be written as a linear combination of translates of $|x|$. There is a very close relationship between 1-D radial and sigmoidal functions.

Regularization Networks



Regularization networks and Kernel regression

- Kernel regression: no complex global model of the world is assumed. Many simple local models instead (a case of *kernel methods*)

$$f(\mathbf{x}) = \frac{\sum_{i=1}^{\ell} w_i(\mathbf{x}) y_i}{\sum_{i=1}^{\ell} w_i(\mathbf{x})}$$

- Regularization networks: fairly complex global model of the world (a case of *dictionary methods*)

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} c_i K(\mathbf{x} - \mathbf{x}_i)$$

Are these two techniques related? Can you say something about the apparent dichotomy of “local” vs. “global”?

Least square Regularization networks

A model of the form

$$f(\mathbf{x}) = \sum_{\alpha=1}^{m} c_\alpha K(\mathbf{x} - \mathbf{t}_\alpha)$$

is assumed and the parameters c_α and $\mathbf{t}_\alpha$ are found by

$$\min_{c_\alpha, \mathbf{t}_\alpha} E[\{c_\alpha\}, \{\mathbf{t}_\alpha\}]$$

where

$$E[\{c_\alpha\}, \{\mathbf{t}_\alpha\}] = \sum_{i=1}^{\ell} (y_i - f(\mathbf{x}_i))^2$$

Least square Regularization networks

The coefficients c_α and the centers $\mathbf{t}_\alpha$ have to satisfy the conditions:

$$\frac{\partial E}{\partial c_\alpha} = 0 \ , \quad \frac{\partial E}{\partial \mathbf{t}_\alpha} = 0 \quad \alpha = 1, \ldots, m$$

The equation for the coefficients gives:

$$c_\alpha = \sum_{i=1}^{\ell} H_{\alpha i} y_i$$

where

$$H = (K^T K)^{-1} K^T \ , \quad K_{i\alpha} = K(\mathbf{x}_i - \mathbf{t}_\alpha)$$

Dual representation

Substituting the expression for the coefficients in the regularization network we obtain

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} y_i \sum_{\alpha=1}^{m} H_{i\alpha}^T K(\mathbf{x} - \mathbf{t}_\alpha)$$

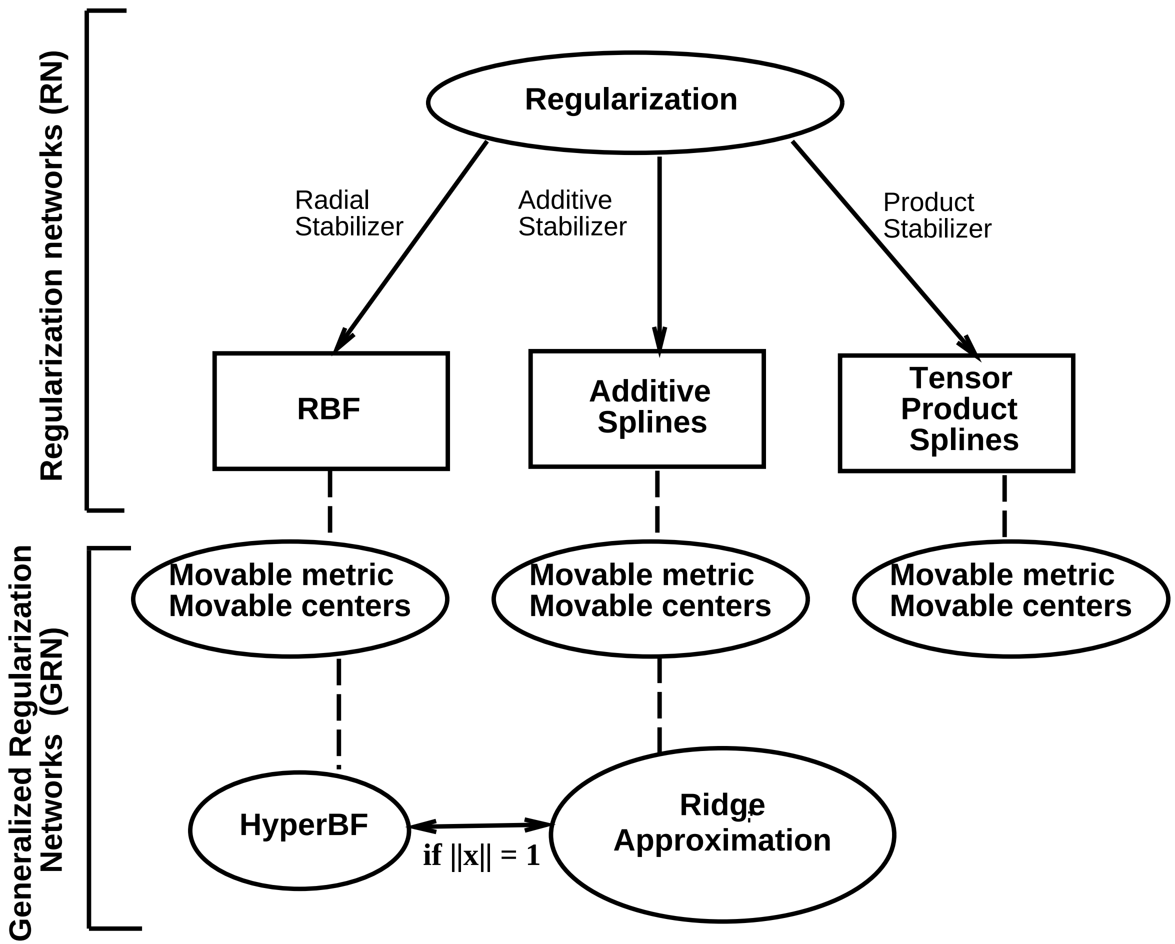
$$\boxed{f(\mathbf{x}) = \sum_{i=1}^{\ell} y_i b_i(\mathbf{x})}$$

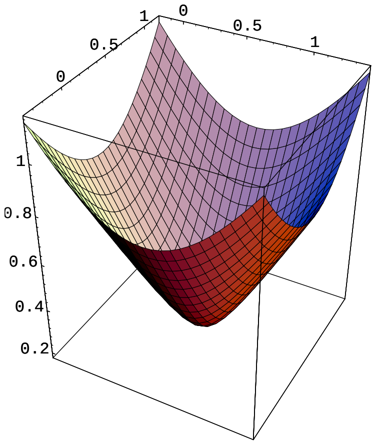
where we have defined

$$b_i(\mathbf{x}) = \sum_{\alpha=1}^{m} H_{i\alpha}^T K(\mathbf{x} - \mathbf{t}_\alpha)$$

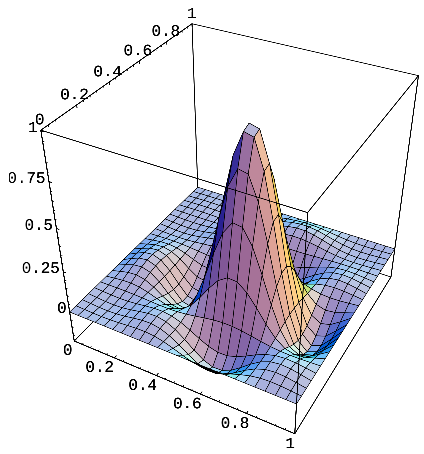
The basis functions $b_i(\mathbf{x})$ are called "dual kernels".

Equivalent kernels for multiquadric basis functions

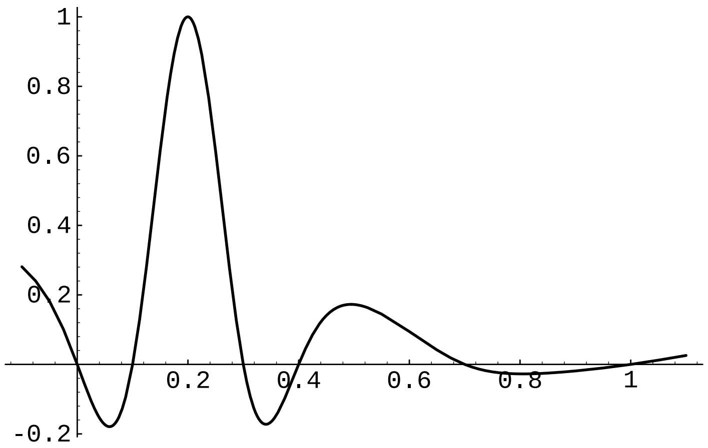




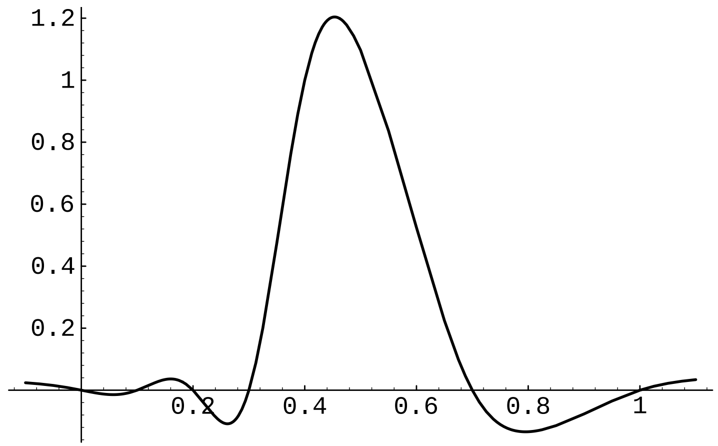
0
0.5
1
0.2
0.4
0.6
0.8
1



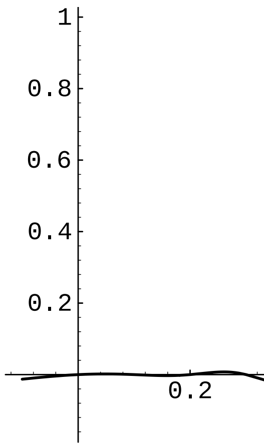
0
0.2
0.4
0.6
0.8
1
0.25
0.5
0.75



-0.2
0.2
0.4
0.6
0.8
1



0.2
0.4
0.6
0.8
1
1.2



0.2
0.4
0.6
0.8
1

Dual formulation of Regularization networks and Kernel regression

$$\boxed{f(\mathbf{x}) = \sum_{i=1}^{\ell} y_i b_i(\mathbf{x})}$$ *Regularization networks*

$$\Updownarrow$$

$$\boxed{f(\mathbf{x}) = \frac{\sum_{i=1}^{\ell} w_i(\mathbf{x}) y_i}{\sum_{i=1}^{\ell} w_i(\mathbf{x})}}$$ *Kernel regression*

In both cases the value of f at point $\mathbf{x}$ is a weighted average of the values at the data points.

Project: is this true for SVMs? Can it be generalized?

Conclusions

- We have extended – with some hand waving – classical, quadratic Regularization Networks including RBF into a number of schemes that are inspired by regularization though do not strictly follow from it.

- The extensions described seem to work well in practice. Main problem – for schemes involving moving centers and or learning the metric – is efficient optimization.