Analysis of Short Term Memory Structures for Neural Networks

Jose C. Principe, Hui-H. Hsu, Jyh-M. Kuo

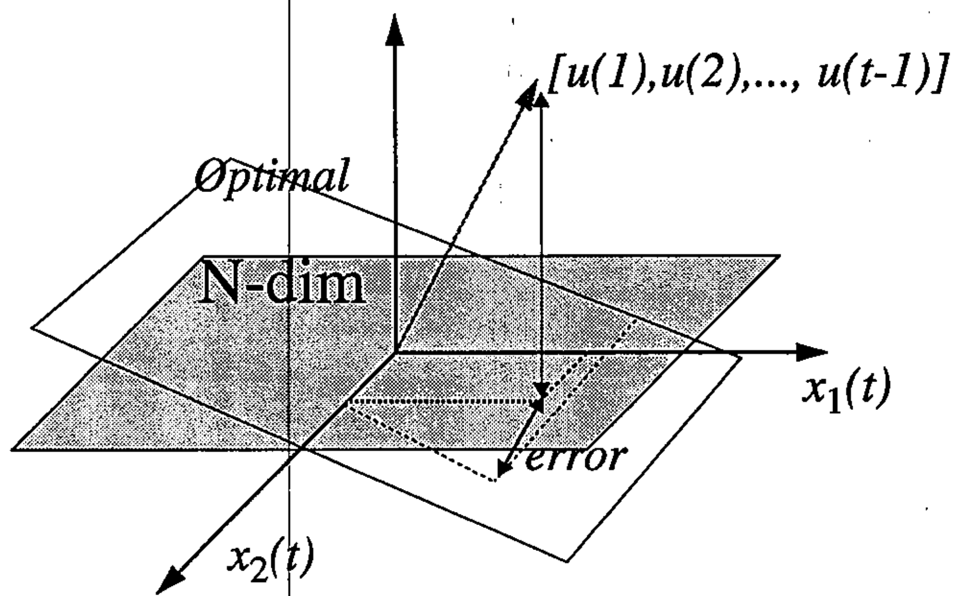
Computational NeuroEngineering Laboratory
Electrical Engineering Dept, University of Florida
Gainesville, FL 32611

Outline

1. Temporal processing; Time-Lagged Recurrent Networks (TLRN)
2. Memory filters, definition and properties.
3. Examples of memory filters.
4. Memory Filters in Nonlinear Prediction
5. Conclusions.

Problem

How do we store a growing input history $[u(1),\ldots, u(t-1)]$ in a N-dimensional state vector $[x_1(t), x_2(t),\ldots, x_N(t)]$?



The usefulness of $x(t)$ depends on how well it spans large temporal vector spaces ($u(t)$) and how well it spans an optimal decision space (*min E*). ($u(t) \text{-->} x(t) \text{-->} E$).

neural net   *perf. crit.*

Memory Architectures

1. memory by feedforward delay

basis vectors $\delta(t-k)$, $k=0,1,2,\ldots$ (# weights ~ memory depth)

2. memory by linear feedback (IIR filters)

basis vectors $a_0^t, ta_1^t, t^2a_2^t, \ldots$, where a_i are functions of the system parameters.

$\Rightarrow$ adaptive vector space!

Hence, recurrent networks allow for optimizing the temporal basis vectors with respect to the performance criterion.

Recurrent Network Design

Formal design:

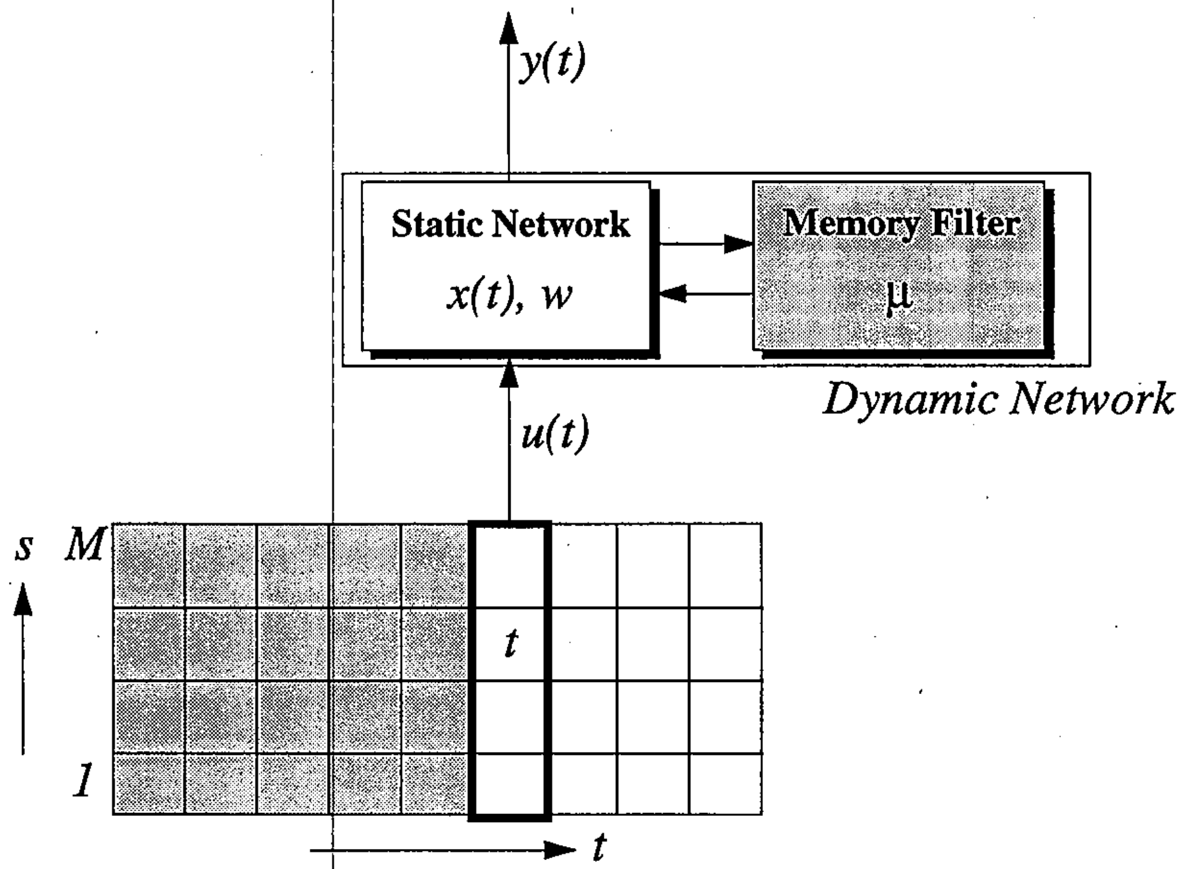
Design by Lyapunov functions (stability methods, Johnson, 1982) leads to stable networks.

Heuristic design:

Restricted architectures such as Jordan net, Elman net, Mozer net, local-recurrent-global-feedforward net, Williams and Zipser, fully recurrent nets, Giles' 2nd-order recurrent nets, and several others.

The sheer amount of different architectures for the same task indicates a rather ad hoc design method.

Unify Short term Memory Mechanisms



Linear Memory Filter

Definitions.

A sequence $g(t)$ is the impulse response of a memory filter if the following two conditions hold:

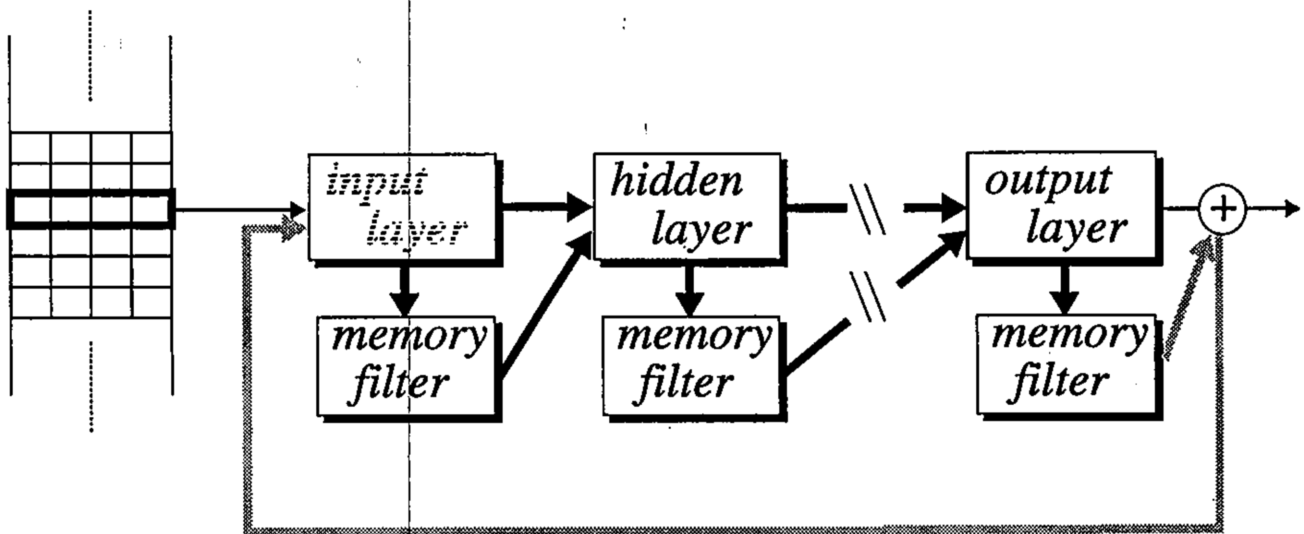
- $g(t)$ is causal, that is, $g(t)=0$ for $t < 0$.
- $g(t)$ is normalized such that $\sum_{t=0}^{\infty} |g(t)| = 1$.

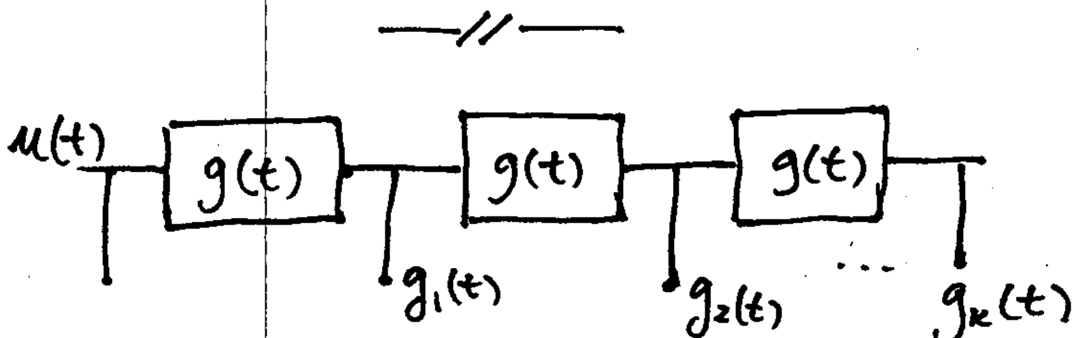
Memory depth: The center of mass of the last tap.

Memory Resolution: The number of taps per unit time.

A memory filter is BIBO stable. ($\sum_{t=0}^{\infty} g(t) < \infty$)

Architecture of TLRN networks



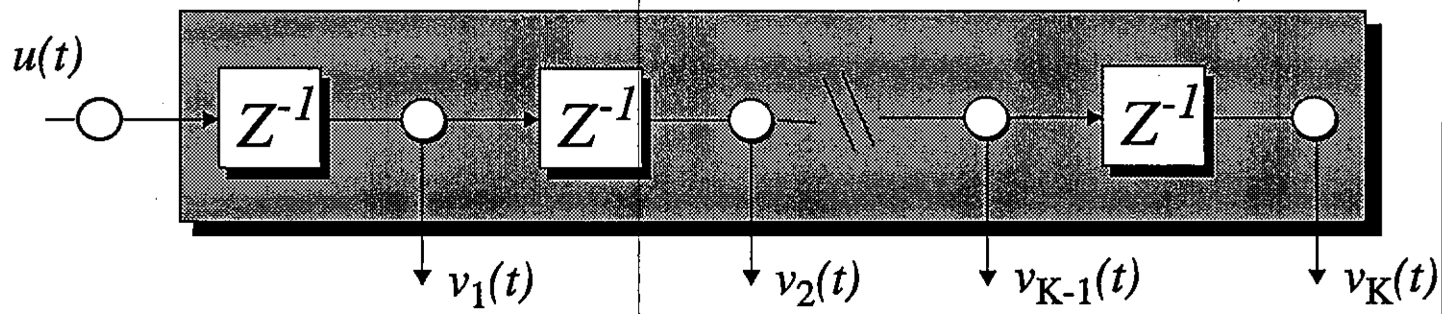


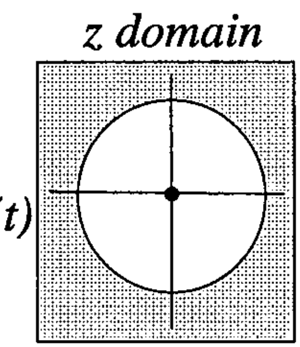
TLRN= Time-Lagged, Recurrent Networks

$$g_k(t) = g(t) * g_{k-1}(t)$$

$*$ → CONVOLUTION

The Tapped Delay Line





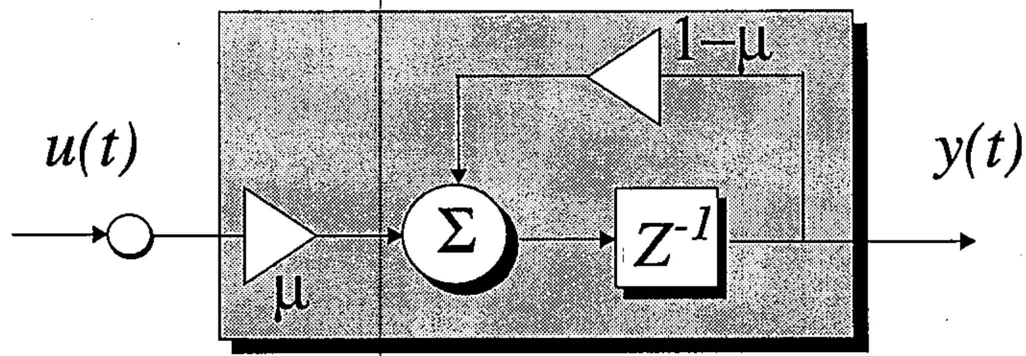
Delay Operator $G(z) = z^{-1}$.

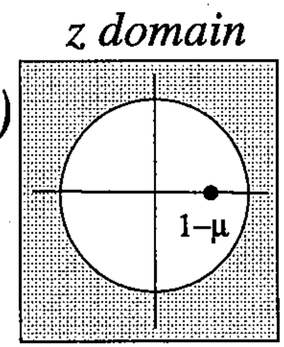
Memory Depth $D_k \equiv \sum_{t=0}^{\infty} t g_k(t) = K$,

Resolution $R_k \equiv \frac{K}{D_k} = 1$.

Notes. General applications; high resolution, # weights proportional to depth!

The Leaky Integrator





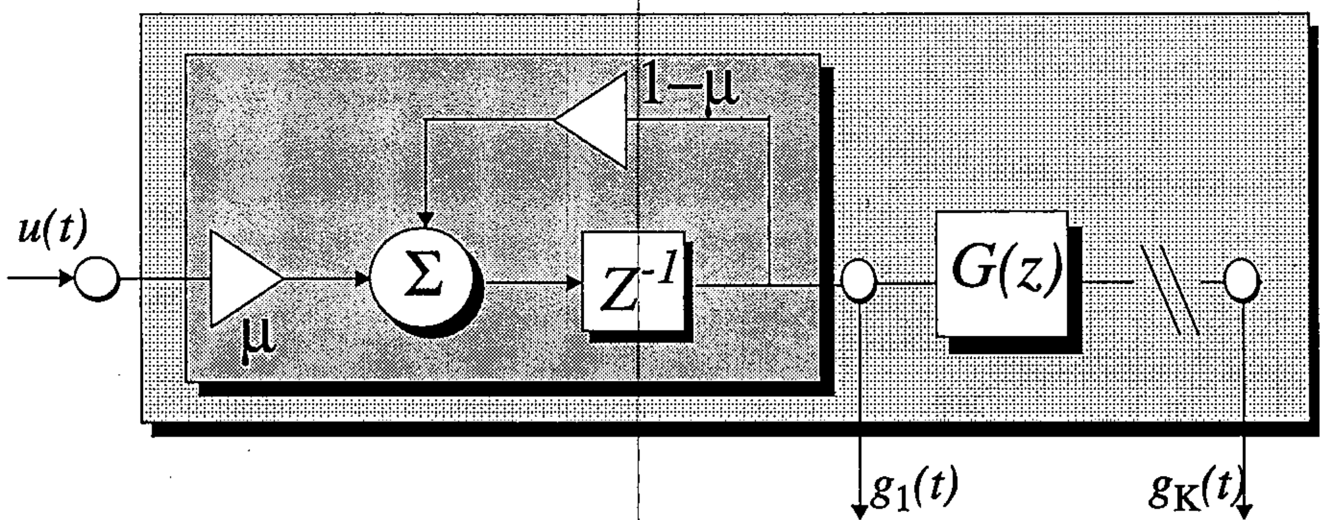
Delay Operator $G(z) = \dfrac{\mu}{z - (1-\mu)}$.

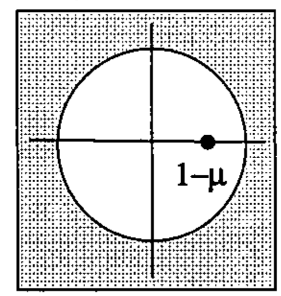
memory Depth $D = \sum_{t=0}^{\infty} t g(t) = \dfrac{1}{\mu}$.

Resolution $R = 1/D = \mu$.

<u>Notes</u>. Also called context units, memory neurons. Apply to problems where *deep memory with low resolution* is needed. Stable for $0 < \mu < 2$.

The Gamma Memory Filter





$$\text{delay operator } G(z) = \frac{\mu}{z - (1-\mu)} \quad .$$

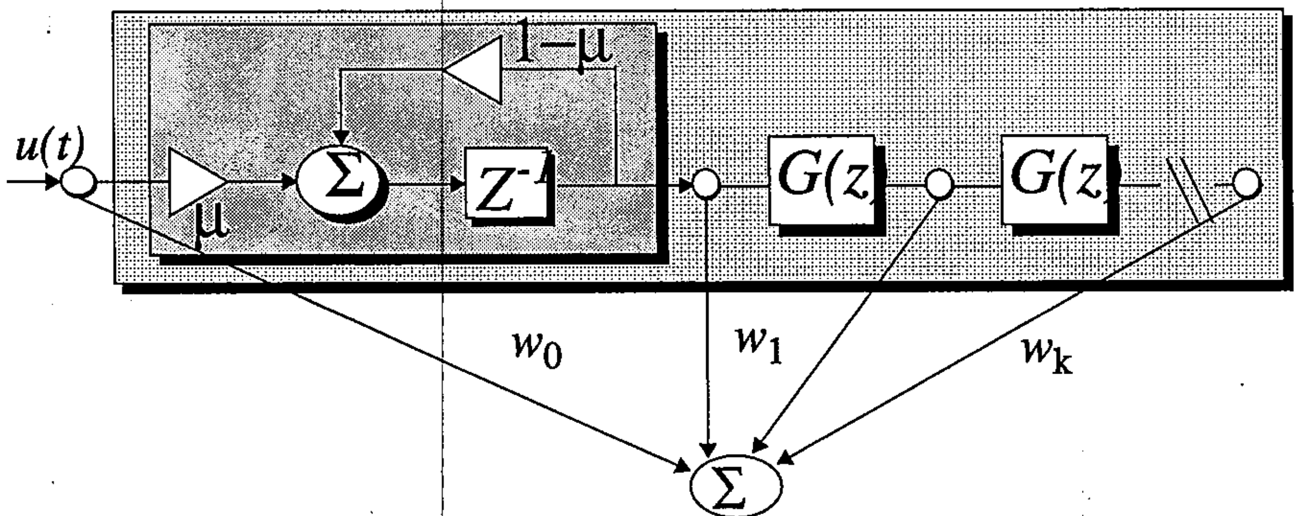
$$\text{depth} \quad D_k = \frac{K}{\mu} \ ,$$

$$\text{resolution} \quad R = K/\left(\frac{K}{1-\mu}\right) = \mu \quad .$$

Notes. Gamma filter generalizes tapped delay line and leaky integrator into a single structure (order, μ).

ADALINE (μ) or Gamma Filter

The gamma filter just extends the adaptive linear network with a variable pole, adapted to the input signal statistics.



Learning equations:

$$\Delta w_k(n) = \eta_1 e(n)\, x_k(n) \qquad k = 0, \ldots, L$$

$$\Delta\mu(n) = \eta_2 \sum_{k=0}^{L} e(n)\, w_k \alpha_k(n)$$

where η is step size, e(n) the error and $\alpha_k(n) \equiv \frac{\partial}{\partial\mu} x_k(n)$

Calculation of the gradient

- Gamma filter equations ($x_0(n)=x(n)$)

$$y(n) = \sum_{k=0}^{K} w_k x_k(n)$$

$$x_k(n) = (1-\mu)x_k(n-1) + \mu x_{k-1}(n-1)$$

- Gradients are

$$\Delta w_k = -\frac{\partial J}{\partial w_k} = \eta_1 \sum_{n=0}^{T} e(n) x_k(n)$$

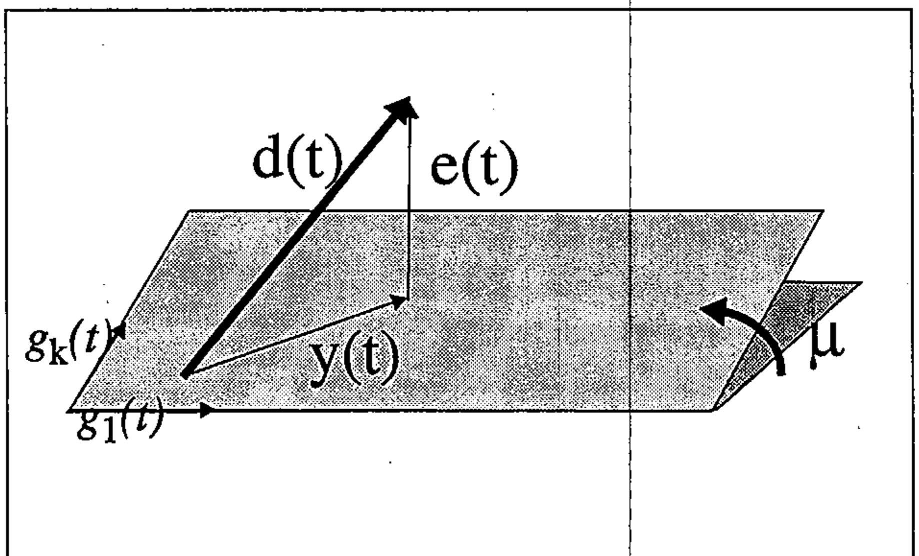
$$\Delta \mu = -\frac{\partial J}{\partial \mu} = \eta_2 \sum_{n=0}^{T} e(n) \sum_{k=0}^{K} w_k \alpha_k(n)$$

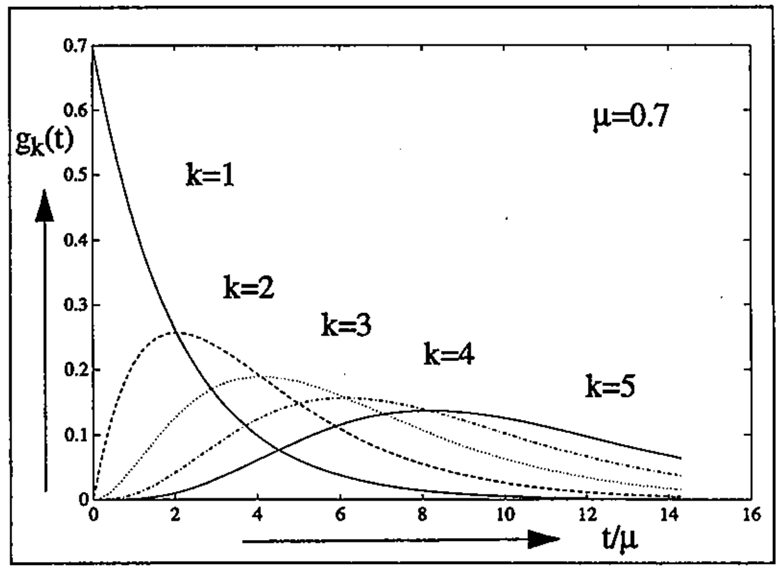
- And the gradient variable is computed as ($\alpha_0(n)=0$)

$$\alpha_k(n) = (1-\mu)\alpha_k(n-1) + \mu\alpha_{k-1}(n-1) + x_{k-1}(n-1) - x_k(n-1)$$

Structure of the gamma space

In continuous time the gamma space is a rigid hyperplane when μ varies. Thus, when the mse is minimized, μ works as an extra degree of freedom that changes the angle between the desired signal and the hyperplane.



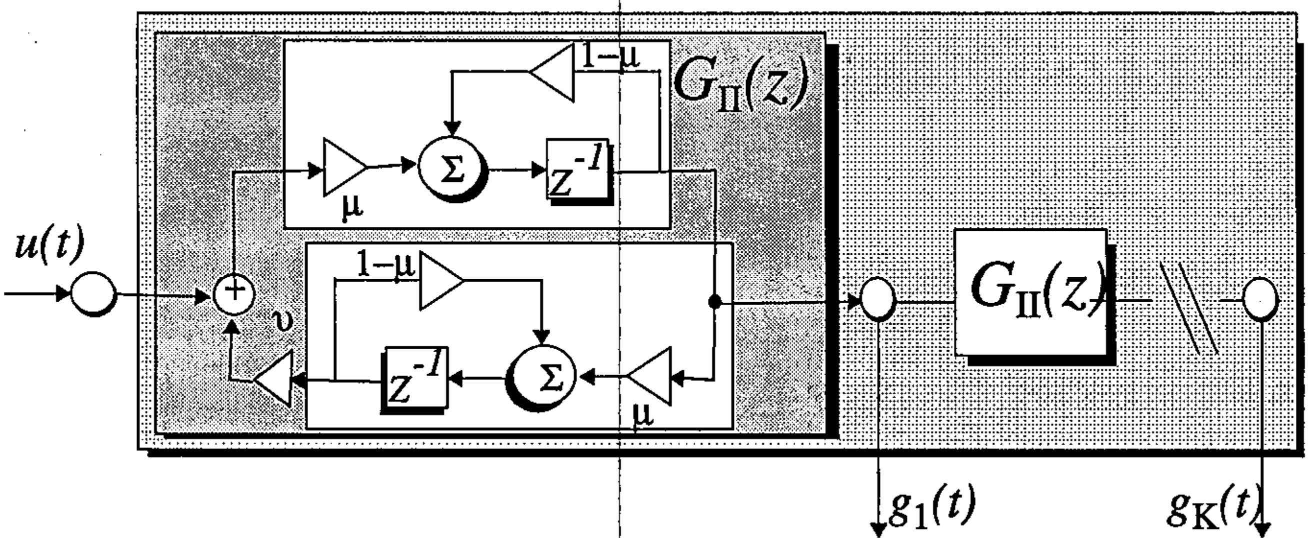


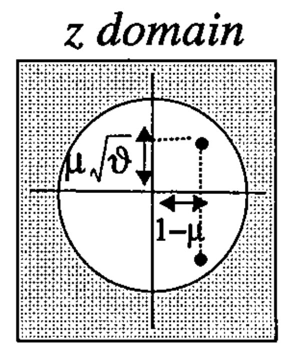
Problem is that the adaptation of μ is non-convex.

The gamma kernel is complete in L_2.

Other TLRN -The Gamma II

Gamma Filter can be extended to complex poles.



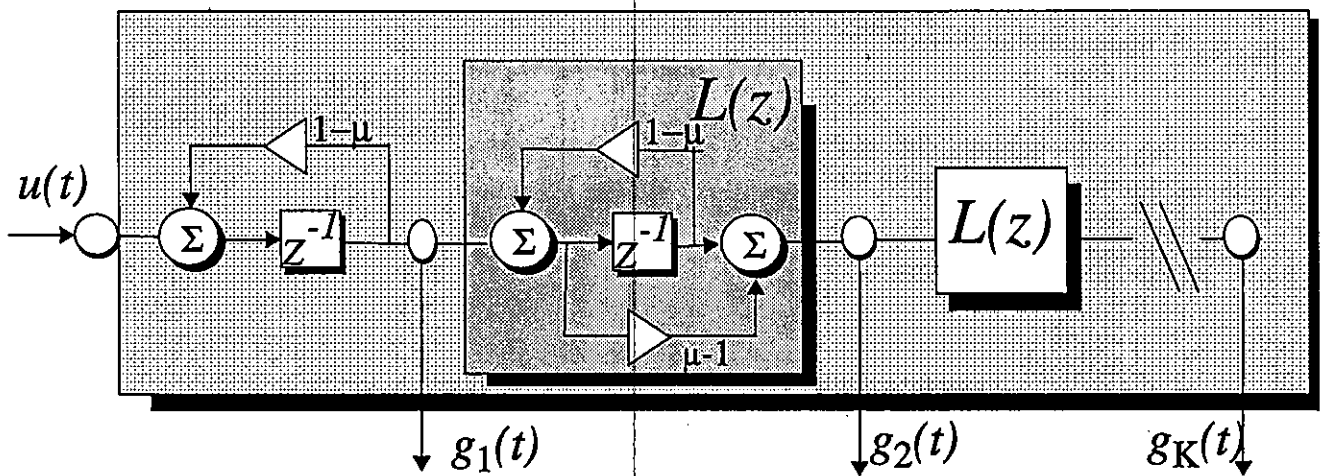


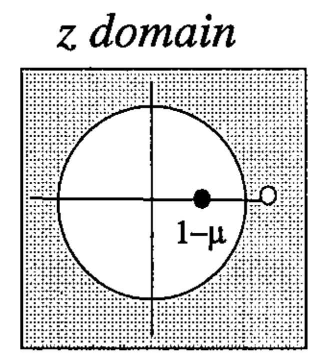
This structure is parametrized by μ and υ. It implements a general frequency dependent delay.

$$\text{Delay operator } G_{II}(z) = \frac{\mu\,[z-(1-\mu)]}{[z-(1-\mu)]^2+\vartheta\mu^2}$$

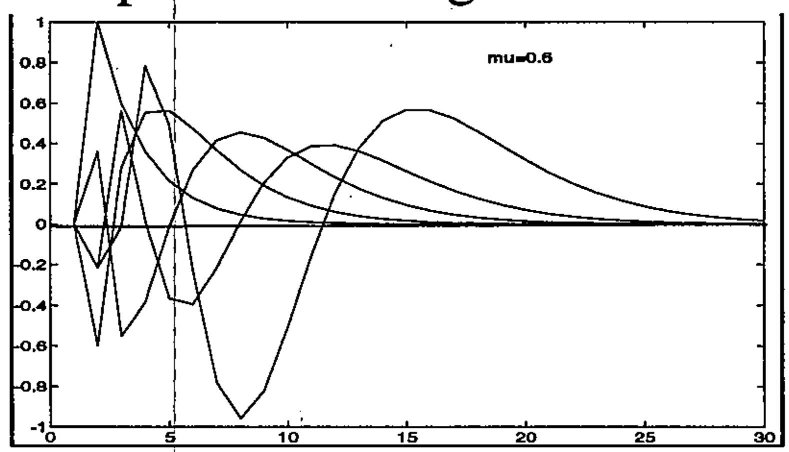
Other TLRN -Laguerre

The Laguerre filters are an orthogonal span of the Gamma space.





Laguerre should adapt faster than gamma for values of $\mu \sim 0, 2$.

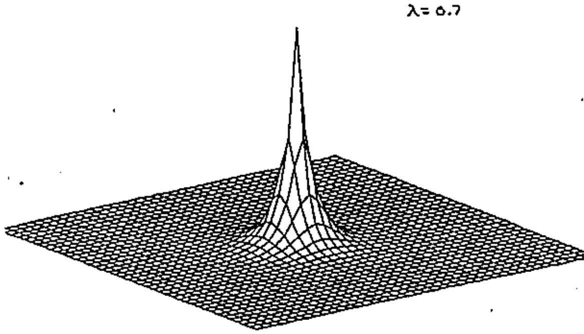


Multi-Dimensional Gamma

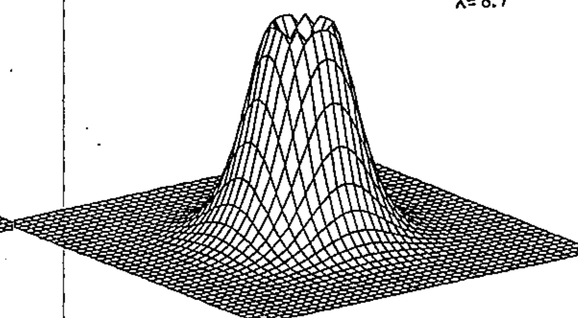
Can be considered an extension of radial basis functions.

$$g_j(\|\vec{x}(n) - c_i\|) = \frac{\mu^j}{(j-1)!}\left[(\vec{x}(n) - c_i)^2\right]^{0.5(j-1)} e^{-\mu\left[(\vec{x}(n) - c_i)^2\right]^{1/2}}$$

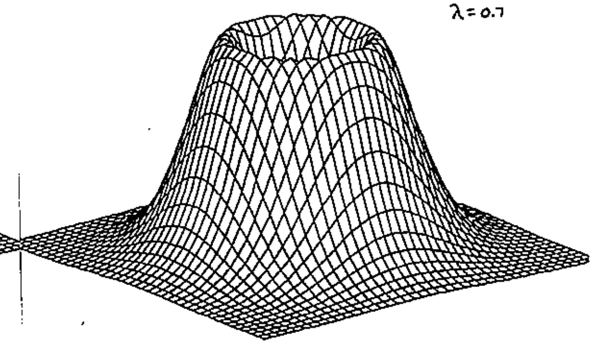
2-D Radially Symmetric Gamma Kernel - 1st Order (Normalized) $\lambda = 0.7$



2-D Radially Symmetric Gamma Kernel - 3rd Order (Normalized) $\lambda = 0.7$



2-D Radially Symmetric Gamma Kernel - 6th Order (Normalized) $\lambda = 0.7$



They are a compromise between local and global approximators.