# A CLOSED FORM SOLUTION FOR A NONLINEAR WIENER FILTER

*Puskal P. Pokharel, Jian-Wu Xu, Deniz Erdogmus[1], Jose C. Principe*

Computational NeuroEngineering Laboratory, ECE Department,
University of Florida, Gainesville, FL 32611

[1]CSEE Department, OGI,
Oregon Health & Science University, Portland, OR 97006

## ABSTRACT

In this paper a nonlinear extension to the Wiener filter is presented. A direct approach has been devised of replacing the autocorrelation function with a novel function called correntropy, derived from ideas on kernel-based learning theory and information theoretic learning. The linear Wiener filter, widely used because of its simplicity and optimality for linear systems and Gaussian distribution, is no longer effective when dealing with nonlinear time series data. The proposed method incorporates higher order moments in the general form of autocorrelation and improves upon the linear filter. Moreover, the computation cost is still lower than some kernel based methods and has a closed form solution to the problem unlike neural network based methods.

## 1. INTRODUCTION

The Wiener filter is one of the true achievements of 20th century optimal system's design. It extended the also well known solution of regression to linear functional spaces, i.e. the space of functions of time (Hilbert Space). However, the way Wiener filters are applied normally in digital computers is in linear vector spaces ($\mathbb{R}_L$) because of the finite impulse response (FIR) filter.

Due to the power of the solution and the relatively easy implementation, Wiener filters have been extensively utilized in all the areas of electrical engineering. Despite this wide spread use, Wiener filters are solutions in linear vector spaces. Therefore, many attempts have been made to create nonlinear solutions to the Wiener filter mostly based on Volterra series [1], but unfortunately the solutions are very complex with many coefficients. There are also two types of nonlinear models that have been commonly used: The Hammerstein and the Wiener models. They are composed of a static nonlinearity and a linear system, where the linear system is adapted using the Wiener solution. However, the choice of the nonlinearity is critical for good performance, because linear solution is obtained in the transformed space. The recent advances of nonlinear signal processing have used nonlinear filters, commonly known as dynamic neural networks [2] that have been extensively used in the basic same applications of Wiener filters when the system under study is nonlinear. However, there are no analytical solutions to obtain the parameters of neural networks. They are normally trained using the back propagation algorithm or its modifications. In some other cases, a nonlinear transformation of the input is first implemented and a regression is computed at the output. A good example of this is the radial basis function (RBF) network [3] and the kernel methods [4,5]. The disadvantage of these alternate techniques of projection is the tremendous amount of computation required, which make them impractical for most real world cases.

The present paper addressed this problem of practical implementation of optimal nonlinear mappings. We show how to extend the analytic solution in linear vector spaces proposed by Wiener to a nonlinear manifold that is obtained by a reproducing kernel Hilbert space. Therefore, we still can compute an analytic optimal solution for a broad class of nonlinear systems, with control of the size of the linear vector space.

## 2. RKHS BASED ON CORRENTROPY

This paper improves upon the concept of the Wiener filter by introducing a nonlinear signal processing framework based on a new similarity function.

Correntropy, as proposed in [6], is a function that generalizes the autocorrelation function to nonlinear spaces. The correntropy of the random process $x(t)$ at instances $t_1$ and $t_2$ is defined as

$$V(t_1,t_2) = E[K(x_{t_1}, x_{t_2})], \qquad (1)$$

where E[.] is the expected value operator, and $K$ a kernel function that obeys the Mercer's conditions [5].One widely used kernel function, also used in this paper, is the Gaussian kernel given by

$$K(x,y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \tag{2}$$

Correntropy has very nice properties that make it useful for nonlinear signal processing [6]. First and foremost, it is a positive function, which means that it also defines a RKHS, but unlike the RKHS defined by the covariance function of random processes it contains higher order statistical information. It can be seen that for $V$ to be a function of a single parameter all the even-order (joint) moments must be invariant to a time shift. That is, the correntropy can be written as $V_{xx}(t,t-\tau) \equiv V_{xx}(\tau)$. This is a stronger condition than wide sense stationarity, which involves only second order moments. We will assume this condition in the rest of the paper when using $V_{xx}(\tau)$.

For discrete-time strictly stationary stochastic processes we can estimate the correntropy function as

$$\hat{V}_{xx}[m] = \frac{1}{N-m+1}\sum_{l=m}^{N} K(x_l - x_{l-m}) \tag{3}$$

This new approach quantifies the average angular separation in the kernel feature space of the random process at a given temporal lag.

*Theorem 1*: For any symmetric positive definite kernel (i.e., Mercer Kernel) $K(x_{t_1}, x_{t_2})$ defined on $\mathbb{R}\times\mathbb{R}$, the correntropy defined as $V(t_1,t_2) = E[K(x_{t_1}, x_{t_2})]$ is a reproducing kernel.

*Proof*: It can be proven that $V(t_1,t_2)$ is positive definite and symmetrical [6]. The Moore-Aronszajn theorem [7] then implies that it is a reproducing kernel.

*Theorem* 2: $V(t_1,t_2)$ is the autocorrelation function of some random process.

*Proof*: It can be proved that R is the covariance function of a random process if and only if, R is a symmetric non-negative kernel [8]. The theorem then follows.

*Theorem* 3: Let $x(n)$ be a stationary stochastic process. Then there exists a mapping $f$ on $\{x(n)\}$ such that $V(i,j) = E\{f(x(i))\cdot f(x(j))\}$ for a class of joint probability density functions (PDFs) of $(x(i),x(j))$.

*Proof*: Let $P_{ij}(x,y)$ be the joint PDF of $(x(i),x(j))$ such that,

$$P_{ij}(x,y) = \sum_{i=1}^{\infty}\alpha_i\,\theta_i(x)\theta_i(y)\,, \tag{4}$$

where $\theta_i(\cdot),\alpha_i$ are respectively the eigen functions and the eigen values of $P_{ij}(x,y)$ and let $\xi_i(\cdot),\lambda_i$ be respectively the eigen functions and eigen values of the kernel function. Then,

$$\begin{aligned} &E\{f(x)f(y)\}\\ &= \iint P_{ij}(x,y)f(x)f(y)dxdy\\ &= \iint \sum_i \alpha_i\theta_i(x)\theta_i(y)f(x)f(y)dxdy\\ &= \sum_i \alpha_i \int\theta_i(x)f(x)dx\int\theta_i(y)f(y)dy\\ &= \sum_i \alpha_i\beta_i^2 \end{aligned} \tag{5}$$

Where, $\beta_i = \int\theta_i(x)f(x)dx$.

Now,

$$\begin{aligned} &E\{K(x,y)\}\\ &= \iint P_{ij}(x,y)K(x,y)dxdy\\ &= \iint \sum_j\sum_i \alpha_i\theta_i(x)\theta_i(y)\lambda_j\xi_j(x)\xi_j(y)dxdy\\ &= \sum_j\sum_i \alpha_i\lambda_j\gamma_{ij} \end{aligned} \tag{6}$$

Observing (4) and (5) we can construct $f$ such that,

$$\beta_i = \sqrt{\sum_j \lambda_j\gamma_{ij}}\,.$$

Then the following exits satisfying the theorem,

$$f(x) = \sum_i \beta_i\theta_i(x) \tag{7}$$

## 3. FILTER BASED ON CORRENTROPY

Since there exists a mapping $f$ which makes the correntropy of $x(n)$ the autocorrelation of $f(x(n))$, let's use this function to map the input data which is then linearly filtered. This would allow the autocorrelation matrix so required to be replaced by the correntropy matrix. In practice we would not explicitly use $f$ as will be seen later. But this idea helps to efficiently get a non-linear version of the Wiener filter. So let $f(x(n))$ be the input to the Wiener structure and L be the length of the filter. Then we can form a composite vector using L lags of $f(x(n))$ denoted by,

$$F(n) = \begin{bmatrix} f(x(n)) & f(x(n-1)) & \cdots & f(x(n-L-1)) \end{bmatrix}^T \tag{8}$$

Also we shall have (L+1) filter weights given by the vector,

$$\Omega = \begin{bmatrix} \omega_0 & \omega_1 & \cdots & \omega_{L-1} \end{bmatrix}^T \tag{9}$$

With this formulation the output is given by,

$$y(n) = \Omega^T F(n) = \sum_{i=0}^{L-1}\omega_i\, f(x(n-i)) \tag{10}$$

Hence we can formulate the optimization problem as follows: Minimize the mean square error, $E\{y(n)-d(n)\}^2$ with respect to $\Omega$.

We have, $E\{y(n)-d(n)\}^2 = E\{\Omega^T F(n) - d(n)\}^2$

The optimization is given by the solution of the following:

$$\frac{\partial(E\{F(n)^T\Omega - d(n)\}^2)}{\partial\Omega} = 0 \qquad (11)$$

$$\Rightarrow E\{F(n)F(n)^T\}\Omega = E\{d(n)F(n)\}$$

Let us evaluate the term on the left hand side.

$$E\{F(n)F(n)^T\}$$

$$= E\begin{bmatrix} f(x(n))f(x(n)) & \cdots & f(x(n))f(x(n-L-1)) \\ \vdots & \vdots & \vdots \\ f(x(n-L-1))f(x(n)) & \cdots & f(x(n-L-1))f(x(n-L-1)) \end{bmatrix} \qquad (12)$$

Choosing $f(\cdot)$ given by (7) implies,

$$E\{f(x(i))f(x(j))\} = E\{K(x(i), y(i))\} \qquad (13)$$

Substituting (13) in (12) and from (11),

$$V\Omega = E\{d(n)f(n)\} \qquad (14)$$

where, $V$ is the correntropy matrix whose ijth (i,j=1,2,…,L) element is $E\{K(x(n-i+1), x(n-j+1))\}$. Further assuming ergodicity we can approximate the $E\{\}$ by the time average. So we have, $\Omega = V^{-1}\frac{1}{N}\sum_{k=1}^{N} d(k)F(k)$ (15)

and the filter output becomes,

$$\begin{aligned} y(n) &= F^T(n)\Omega \\ &= F^T(n)V^{-1}\frac{1}{N}\sum_{k=1}^{N} d(k)F(k) \\ &= \frac{1}{N}\sum_{k=1}^{N}\sum_{j=0}^{L-1}\sum_{i=0}^{L-1} f(n-i)a_{ij}f(k-j)d(k) \qquad (16) \\ &= \frac{1}{N}\sum_{k=1}^{N} d(k)\sum_{j=0}^{L-1}\sum_{i=0}^{L-1} a_{ij}\{f(n-i)f(k-j)\} \\ &\cong \frac{1}{N}\sum_{k=1}^{N}\left\{d(k)\sum_{j=0}^{L-1}\sum_{i=0}^{L-1} a_{ij}K(x(n-i), x(k-j))\right\} \end{aligned}$$

Where $a_{ij}$ is the ijth element of $V^{-1}$. The final expression is obtained by approximating $\{f(n-i)f(k-j)\}$ by $K(x(n-i), x(k-j))$, which holds good on an average because of (13). Hence, we do not need to find the transformation *f*(.) since it is never utilized in the calculations, as expected by the "kernel trick". The final output is obtained by matching the scale of $y(n)$ to that of the desired signal. There is a mismatch in scale most likely because of the above approximation and because the mean of $f(x(n))$ is likely not zero.

## 4. EXPERIMENTS AND RESULTS

We shall use the normalized mean square error as a means of comparing the performance of the novel correntropy filter (CF) and the linear Wiener filter (WF). The normalized mean square error (MSE) is nothing but the MSE calculated after normalizing the desired signal and the output each to unit variance. Here we shall show the results of one step prediction of the Mackey-Glass (MG30) time series. The simulations implement equation (16) for the CF and the equation below for the linear Wiener filter. $y(n) = X(n)^T(R^{-1}P)$, where $R$ is the autocorrelation matrix of the input and $P$ is the cross correlation vector [9]. One crucial parameter to choose is the kernel size. The kernel size, $\sigma$ is chosen to be 0.15 for these experiments. It has been observed that the kernel size should be around 15% of the standard deviation of the input data. The plots include comparisons of MSE values for different filter lengths (L in equation (16)) and size of the training data (N). The correntropy filter achieves the best result for the filter length of L=6 (fig. 1), which is also the optimal length according to Takens embedding theorem [10]. Fig. 2 shows how the increase in data size affects the performance. Fig. 3 shows how the two filters perform for different prediction steps (the prediction step for the previous two figures is one). In any case, the CF always performs better and when the optimal filter length of 6 is chosen the MSE using the CF is less than 50 % of that using the linear Wiener filter.

## 5. CONCLUSION

This paper presents an investigation on a new type of Wiener filter based on the recently introduced correntropy function. Correntropy is a positive function and as such establishes a RKHS with an inner product that contains information about the higher order moments of the stochastic process. This paper also shows a means of directly using the correntropy as the autocorrelation function of the projected data. This approach can be used to extend other linear supervised (and possibly non-supervised) learning schemes to nonlinear algorithms very effectively. Obviously, the computation cost for this new approach is larger than that for the linear Wiener filter. But our approach is still less expensive than other kernel based methods like kernel regression. This is due to the use of the correntropy matrix as the similarity measure, whose dimension is the same as the filter order, instead of the matrix of projected points required by all kernel methods. We also expect that the correntropy filter solution may be worse than the kernel methods solution. For instance the MSE comparison with a RBF network is shown in fig. 4. Fifty training samples were used (for both methods) with one RBF centered at each data vector formed by using embedding six lags of data. But the advantage is that the filter order is decoupled from the Gram matrix size. The correntopy filter has a structure very similar to the Hammerstein model, but notice that here the nonlinearity is implicitly derived from the pdf of the data. The definition of this new RKHS can therefore offer a lot of practical advantages for nonlinear signal processing, in particular all the kernel methods mushrooming in the machine learning literature.

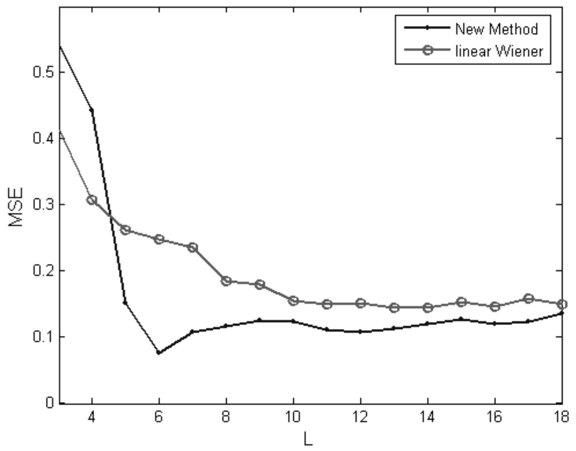

Figure 1. MSE for various filter lengths using N=100.

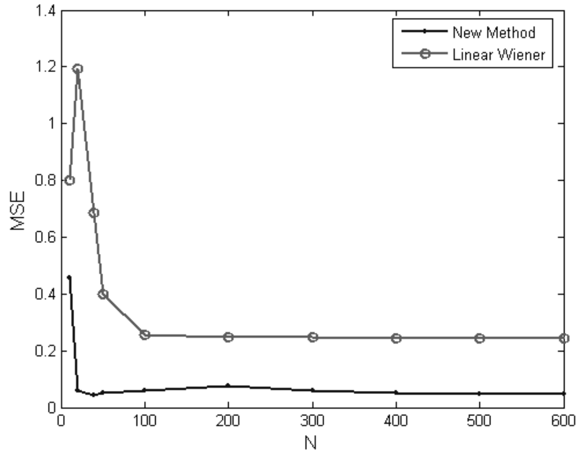

Figure 2. MSE for various training data size with L=6.

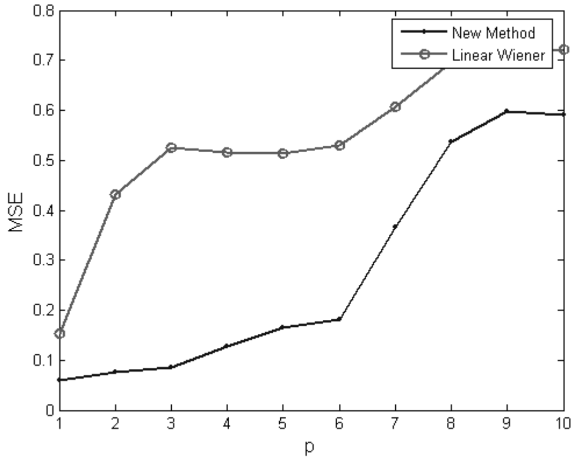

Figure 3. MSE for different prediction steps with L=6, N=100.

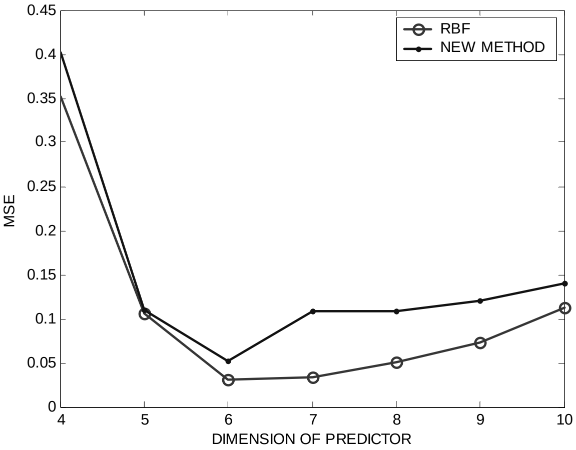

Figure 4. MSE comparison for training data size of 50 samples.

## 6. ACKNOWLEDGMENT

This work was supported in part by NSF grant ECS-0300340.

## 7. REFERENCES

[1] R.J.P. Defigueiredo, and Y. Hu, “On Nonlinear Filtering of Non-Gaussian Processes through Volterra Series,” *Volterra Equations and Applications*, Gordon and Breach Science Publishers, Amsterdam, pp. 197-202, 2000.

[2] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2$^{nd}$ Edition, Prentice Hall, 1998.

[3] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[4] K-R. Müller, A.J. Smola, Gunner Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik, “Using Support Vector Machines for Time Series Prediction”, *Advances in Kernel Methods*, MIT Press, Cambridge, 1999.

[5] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag, New York, 1995.

[6] Santamaría I., P. P. Pokharel, J. C. Principe, “Generalized Correlation Function: Definition, Properties and Application to Blind Equalization,” accepted for *IEEE Trans. Signal Processing.*

[7] N. Aronszajn, “ Theory of Reproducing Kernels,” *Trans. Amer. Math. Soc.*, vol. 68, pp. 337-404, 1950.

[8] E. Parzen, *Statistical Methods on Time Series by Hilbert Space Methods*, Technical Report no. 23, Applied Mathematics and Statistics Laboratory, Stanford University, 1959.

[9] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, New Jersey, 2002.

[10] F. Takens, “Detecting Strange Attractors in Turbulence”, *Dynamical systems and Turbulence*, vol. 898 of *Lecture Notes in Mathematics*, Springer Verlag, Berlin, pp. 366-381, 1981.