# STEADY-STATE MEAN SQUARE PERFORMANCE OF A SPARSIFIED KERNEL LEAST MEAN SQUARE ALGORITHM

*Badong Chen*[1*] *, Zhengda Qin*[1] *, Lei Sun*[2]

[1] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China.
[2] School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China
e-mail: chenbd@mail.xjtu.edu.cn

## ABSTRACT

In this paper, we investigate the convergence performance of a sparsified kernel least mean square (KLMS) algorithm in which the input is added into the dictionary only when the prediction error in amplitude is larger than a preset threshold. Under certain conditions, we derive an approximate value of the steady-state excess mean square error (EMSE). Simulation results confirm the theoretical predictions and provide some interesting findings,showing that the sparsification can not only be used to constrain the network size (hence reduce the computational burden) but also be used to improve the steady-state performance in some cases.

***Index Terms***— KLMS, sparsification, mean square performance.

## 1. INTRODUCTION

Kernel adaptive filtering (KAF) algorithms are powerful online learning methods particularly useful for nonlinear and non-stationary complex system modeling [1–15]. The kernel least mean square (KLMS) [1] is the simplest but very efficient KAF algorithm which naturally creates a growing RBF type network. In order to further simplify the computational complexity and improve the practicability, so far many sparsification (or quantization) methods have been proposed to curb the KLMS network growth [4]. The mean square convergence behavior of the KLMS has been studied in the literatures, where, however for simplicity, no sparsification has been considered [13], or the dictionary (the set of the hidden layer centers) has been assumed to be pre-tuned [14].

In this paper, we continue to study the mean square convergence performance of the KLMS, where a simple online sparsification method is adopted to constrain the network growth. It is an error threshold based sparsification rule under which the input is added into the dictionary only when the prediction error in amplitude is larger than a preset threshold. An approximate value of the steady-state *excess mean square error* (EMSE) [13] has been obtained under certain conditions. Simulation results have been presented to confirm the theoretical predictions.

## 2. KLMS WITH ERROR THRESHOLD BASED SPARSIFICATION

Suppose the goal is to learn a nonlinear mapping $f : \mathbf{U} \rightarrow \mathbf{R}$ that fits the data $\{\mathbf{u}(i), d(i)\}$, $i = 1, 2, \cdots$, where $\mathbf{u}(i) \in \mathbf{U} \subset \mathbf{R}^m$ is the $m$-dimensional input at the instant $i$, and $d(i) \in \mathbf{R}$ is the desired response. With KLMS, this learning problem can be solved by [1]

$$\begin{cases} f_0 = 0 \\ e(i) = d(i) - f_{i-1}(\mathbf{u}(i)) \\ f_i = f_{i-1} + \eta e(i) \kappa\left(\mathbf{u}(i), .\right) \end{cases} \tag{1}$$

where $f_i$ denotes the estimated mapping at the iteration $i$, $e(i) = d(i) - f_{i-1}(\mathbf{u}(i))$ is the prediction error based on the last estimate $f_{i-1}$, $\eta > 0$ is the step-size parameter, and $\kappa : \mathbf{U} \times \mathbf{U} \rightarrow \mathbf{R}$ stands for a reproducing Mercer kernel func-

*This work was supported by 973 Program (2015CB351703) and National Natural Science Foundation of China (61372152, 61673059).

tion. In this work, without mentioned otherwise, we choose the following Gaussian kernel:

$$\kappa\left(u,u'\right)=\exp\left(-\frac{\left\|u-u'\right\|^2}{2\sigma^2}\right) \tag{2}$$

with $\sigma>0$ being the kernel bandwidth. With Gaussian kernel, the KLMS produces a growing RBF network by allocating a new kernel unit for every new example with input $\mathbf{u}(i)$ as the center and $\eta e(i)$ as the coefficient. In order to constrain the network growth and obtain a compact model, one can use some sparsification method to prune the insignificant centers [4]. In this study, we consider an error threshold based sparsification method, which can be regarded as a special case of the simple *Novel Criterion* (NC) based sparsification [4]. In the error threshold based sparsification, when a new data pair $\{\mathbf{u}(i),d(i)\}$ is presented, the input $\mathbf{u}(i)$ will be added into the dictionary only when the prediction error $e(i)$ is in amplitude larger than a certain preset threshold, say $\varepsilon$ ($\varepsilon\geq 0$). With such a sparsification method, the mapping update equation of the KLMS becomes

$$f_i=f_{i-1}+\eta g\left(e(i)\right)\kappa\left(\mathbf{u}(i),.\right) \tag{3}$$

where $g\left(e(i)\right)$ is an error nonlinearity given by

$$g\left(e(i)\right)=\begin{cases}0 & if\ |e(i)|\leq\varepsilon\\ e(i) & if\ |e(i)|>\varepsilon\end{cases} \tag{4}$$

In general, increasing $\varepsilon$ will decrease the network size but the performance may degrade, because a large value of $\varepsilon$, apparently, results in discarding more centers. However, our simulations show that sometimes the performance will even improve with $\varepsilon$ increasing in a certain range.

## 3. STEADY-STATE EMSE OF THE KLMS WITH ERROR THRESHOLD BASED SPARSIFICATION

Assume that the desired signal is related to the input vector via [13, 15]

$$d(i)=f(\mathbf{u}(i))+v(i) \tag{5}$$

where $f(.)$ is the unknown nonlinear mapping that needs to be estimated, and $v(i)$ denotes an additive noise. With a similar derivation as in [13,16–18], it can be shown that the following relation holds:

$$\begin{aligned}E\left[\left\|\tilde{f}_i\right\|_{H_\kappa}^2\right]=&E\left[\left\|\tilde{f}_{i-1}\right\|_{H_\kappa}^2\right]-2\eta E\left[e_a(i)g(e(i))\right]\\&+\eta^2E\left[g^2(e(i))\right]\end{aligned} \tag{6}$$

where $E$ stands for the expectation over the distribution of training data, $\tilde{f}_i=f-f_i$ is the residual mapping, $\|.\|_{H_\kappa}$ denotes the norm in the reproducing kernel Hilbert space (RKHS) $H_\kappa$ induced by the kernel function $\kappa$, and $e_a(i)=\tilde{f}_{i-1}\left(\mathbf{u}(i)\right)$ is the *a priori* error. If the filter is stable and reaches a steady-state, we have $E\left[\left\|\tilde{f}_i\right\|_{H_\kappa}^2\right]=E\left[\left\|\tilde{f}_{i-1}\right\|_{H_\kappa}^2\right]$ as $i\to\infty$. It follows easily that

$$\lim_{i\to\infty}2E\left[e_a(i)g(e(i))\right]=\lim_{i\to\infty}\eta E\left[g^2(e(i))\right] \tag{7}$$

If $g(.)$ is second order differentiable and at the steady-state, the following assumptions hold [18]:

**A1**: The noise $\{v(i)\}$ is zero-mean, i.i.d., and independent of the input signal $\{\mathbf{u}(i)\}$;

**A2**: The *a priori* $e_a(i)$ is zero-mean and independent of the noise $\{v(i)\}$, and is relatively small such that its third and higher-order moments are negligible.

then, from (7) and in a similar way as in [18], one can derive an approximate value of the steady-state *excess mean square error* (EMSE) by taking the Taylor series expansion:

$$\begin{aligned}S=&\lim_{i\to\infty}E\left[e_a^2(i)\right]\\\approx&\frac{\eta Tr\left(R_X\right)E\left[g^2\left(v\right)\right]}{2E\left[g'\left(v\right)\right]-\eta Tr(R_X)E\left[g\left(v\right)g''\left(v\right)+\left|g'\left(v\right)\right|^2\right]}\\\overset{(a)}{=}&\frac{\eta E\left[g^2\left(v\right)\right]}{2E\left[g'\left(v\right)\right]-\eta E\left[g\left(v\right)g''\left(v\right)+\left|g'\left(v\right)\right|^2\right]}\end{aligned} \tag{8}$$

where $g'(.)$ and $g''(.)$ denote the first and second order derivatives of the function $g(.)$, $R_X=E\left[\varphi\left(\mathbf{u}\left(i\right)\right)\varphi(\mathbf{u}\left(i\right))^T\right]$, with $\varphi(.)$ being the nonlinear mapping that transforms the input into a high-dimensional feature space, $Tr(.)$ is the trace operator, and (a) comes from

$$\begin{aligned}Tr\left(R_X\right)=&Tr\left(E\left[\varphi\left(\mathbf{u}\left(i\right)\right)\varphi(\mathbf{u}\left(i\right))^T\right]\right)\\=&Tr\left(E\left[\varphi(\mathbf{u}\left(i\right))^T\varphi\left(\mathbf{u}\left(i\right)\right)\right]\right)\\\overset{(b)}{=}&Tr\left(E\left[\kappa\left(\mathbf{u}\left(i\right),\mathbf{u}\left(i\right)\right)\right]\right)=1\end{aligned} \tag{9}$$

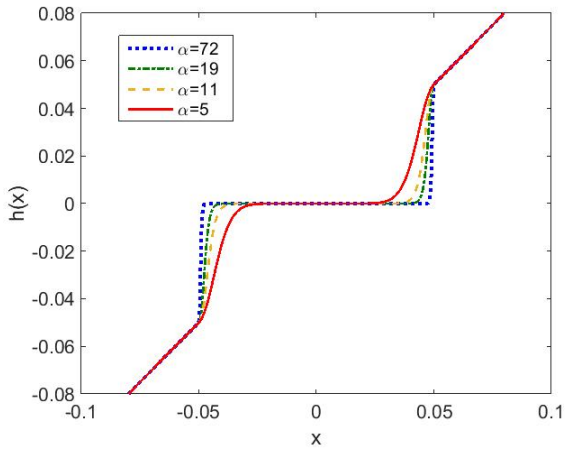


**Fig. 1:** Curves of the function $h(.)$ with different $\alpha$ ($\varepsilon$=0.05)

where (b) follows from the well-known kernel trick.

However, the function $g(.)$ in (4) is non-smooth and obviously, is not second order differentiable. To evaluate the steady-state EMSE by (8), we propose in this work to approximate the non-smooth function $g(.)$ using the following second order differentiable function:

$$h(x)=\begin{cases}\lambda\left[1-\exp\left(-\beta x^{\alpha}\right)\right] & if \quad 0\le x\le\varepsilon\\ -\lambda\left[1-\exp\left(-\beta|x|^{\alpha}\right)\right] & if -\varepsilon\le x<0\\ x & if \quad |x|>\varepsilon\end{cases} \tag{10}$$

where $\alpha$, $\beta$, $\lambda$ are appropriate positive numbers satisfying $h(\varepsilon)=\varepsilon$ and $h'(\varepsilon)=1$ in order to ensure the continuity of the function and its first-order derivative. Thus, we have

$$\begin{cases}\lambda\left(1-\exp\left(-\beta x^{2\alpha}\right)\right)=x\\ 2\lambda\beta\alpha\exp\left(-\beta x^{2\alpha}\right)x^{2\alpha-1}=1\end{cases} \tag{11}$$

Given the value of $\lambda$, $\alpha$ and $\beta$ can be determined by

$$\begin{cases}\alpha=-\varepsilon\left[\lambda\left(1-\frac{\varepsilon}{\lambda}\right)\log\left(1-\frac{\varepsilon}{\lambda}\right)\right]^{-1}\\ \beta=-\frac{1}{\varepsilon^{\alpha}}\log\left(1-\frac{\varepsilon}{\lambda}\right)\end{cases} \tag{12}$$

Fig. 1 shows the curves of the function $h(.)$ with different values of $\alpha$ (where $\varepsilon$=0.05). As one can see, the function $h(.)$ approaches closer and closer to the function $g(.)$ as $\alpha$ gets larger.

With the above second order differentiable function $h(.)$, and under the assumptions **A1** and **A2**, the steady-state EMSE of the KLMS with error threshold based sparsification can thus be, approximately, evaluated by

$$S\approx\frac{\eta E\left[h^{2}\left(v\right)\right]}{2E\left[h'\left(v\right)\right]-\eta E\left[h\left(v\right)h''\left(v\right)+\left|h'\left(v\right)\right|^{2}\right]} \tag{13}$$

where $h(.)$ is given by (10), and the functions $h'(.)$ and $h''(.)$ are

$$h'\left(x\right)=\begin{cases}\alpha\beta\left(\lambda-h\left(x\right)\right)x^{\alpha} & if \quad 0\le x\le\varepsilon\\ \alpha\beta\left(\lambda+h\left(x\right)\right)|x|^{\alpha} & if -\varepsilon\le x<0\\ 1 & if \quad |x|>\varepsilon\end{cases} \tag{14}$$

$$h''\left(x\right)=\begin{cases}h'\left(x\right)\left(\alpha-1-\alpha\beta x^{\alpha}\right)x^{-1} & if \quad 0\le x\le\varepsilon\\ h'\left(x\right)\left(\alpha-1-\alpha\beta|x|^{\alpha}\right)x^{-1} & if -\varepsilon\le x<0\\ 0 & if \quad |x|>\varepsilon\end{cases} \tag{15}$$

*Remark*: Note that the steady-state EMSE in (13) is valid only when the *a priori* error $e_a(i)$ is small at the steady-state such that its third and higher-order moments are negligible. When the threshold $\varepsilon$ is too large, the derived value will deviate from the actual performance seriously, since a larger threshold results in a larger *a priori* error in general. In practical applications, the threshold $\varepsilon$ is usually set at a small value to ensure good performance.

## 4. SIMULATION RESULTS

Consider the identification of the following nonlinear system

$$\begin{aligned}d(i)&=f\left(\mathbf{u}(i)\right)+v(i)\\&=\sin\left(u(i)\right)+0.5u(i-1)-0.1u^{2}(i-2)+v(i)\end{aligned} \tag{16}$$

where $\mathbf{u}(i)=[u(i-2),u(i-1),u(i)]^{T}$ with input sequence $\{\mathbf{u}(i)\}$ being a white Gaussian process with unit variance. In the simulations, the step-size and Gaussian kernel bandwidth are set to $\eta=0.5$, $\sigma$=1.0.

Fig. 2 shows the theoretical and simulated steady-state EMSEs in zero-mean Gaussian noises with different noise variances (0.0025,0.01) and error thresholds $\varepsilon$, and Fig. 3 illustrates the results in zero-mean Uniform noises with different noise variances (0.01,0.04) and error thresholds $\varepsilon$. The theoretical EMSEs are computed by using (13) with $\alpha=62$, and the simulated EMSEs are evaluated as averages over 50 independent Monte Carlo runs and in each run, 100000 iterations are performed to ensure the algorithm to reach the steady-state. The steady-state EMSE in every Monte Carlo run is obtained as the average over the last 1000 iterations.

**Table 1:** Theoretical and simulated EMSEs with different error threshold $\varepsilon$ in different noises

| | Gaussian( $\sigma_v^2 = 0.1^2$ $\eta = 0.5$) | | | Uniform( $\sigma_v^2 = 0.1^2$ $\eta = 0.5$) | |
|---|---|---|---|---|---|
| $\varepsilon$ | Simulation | Theory | $\varepsilon$ | Simulation | Theory |
| 0.01 | $0.00338 \pm 3.36 \times 10^{-8}$ | 0.00333 | 0.00 | $0.00756 \pm 1.26 \times 10^{-8}$ | 0.00750 |
| 0.06 | $0.00340 \pm 5.10 \times 10^{-8}$ | 0.00340 | 0.05 | $0.00747 \pm 1.24 \times 10^{-8}$ | 0.00750 |
| 0.11 | $0.00335 \pm 5.30 \times 10^{-8}$ | 0.00378 | 0.11 | $0.00715 \pm 1.02 \times 10^{-8}$ | 0.00700 |
| 0.16 | $0.00341 \pm 5.50 \times 10^{-8}$ | 0.00478 | 0.17 | $0.00607 \pm 1.23 \times 10^{-8}$ | 0.00550 |

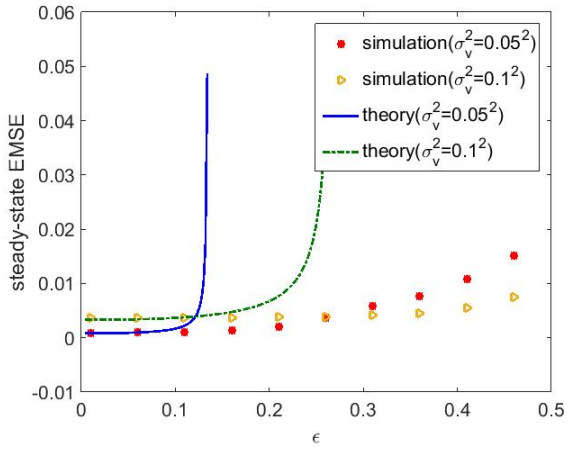


**Fig. 2:** Theoretical and simulated steady-state EMSEs with different noise variances and error thresholds $\varepsilon$ (Gaussian noise case)

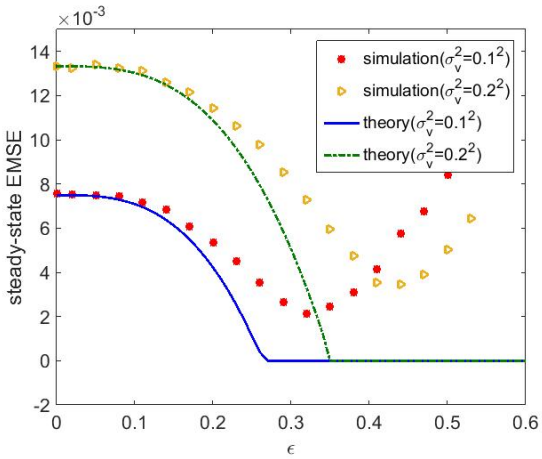


**Fig. 3:** Theoretical and simulated steady-state EMSEs with different noise variances and error thresholds $\varepsilon$ (Uniform noise case)

From Fig. 2 and Fig. 3, one can observe: 1) when the error threshold $\varepsilon$ is relative small (say $\varepsilon < 0.1$), the theoretical values match the simulated results very well; 2) while as the error threshold becomes larger and larger, the theoretical EMSEs will deviate from the simulated performance seriously, and this is due to the fact that a larger error threshold usually results in a larger *a priori* error; 3) when the error threshold is large, a larger noise variance may even result in a smaller EMSE in simulation; 4) for the Uniform noise case, when the error threshold $\varepsilon$ is small, the performance will even improve with $\varepsilon$ increasing. Our simulation results suggest that the error threshold $\varepsilon$ can not only be used to constrain the network size but also be used to reduce the noise effects in some cases. Table 1 presents the detailed values of the theoretical and simulated EMSEs with different error threshold $\varepsilon$ in different noises.

## 5. CONCLUSION

Up to now, there is still no study on the mean square convergence performance of a sparsified KLMS algorithm. In this work, we investigated this problem and under certain conditions, we derived an approximate value of the steady-state excess mean square error (EMSE) of the KLMS with an error threshold based sparsification. Simulation results verified the theoretical predictions and gave some interesting discoveries. In particular, it has been shown that the sparsification procedure can even improve the steady-state performance, which is beyond our expectation.

## 6. REFERENCES


[1] Weifeng Liu, Puskal P Pokharel, and Jose C Principe, "The kernel least-mean-square algorithm," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 543–554, 2008.

[2] Weifeng Liu and José C Príncipe, "Kernel affine projection algorithms," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, no. 1, pp. 1–12, 2008.

[3] Yaakov Engel, Shie Mannor, and Ron Meir, "The kernel recursive least-squares algorithm," *IEEE Transactions on signal processing*, vol. 52, no. 8, pp. 2275–2285, 2004.

[4] Weifeng Liu, Jose C Principe, and Simon Haykin, *Kernel adaptive filtering: a comprehensive introduction*, John Wiley & Sons, 2011.

[5] Cédric Richard, José Carlos M Bermudez, and Paul Honeine, "Online prediction of time series data with kernels," *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 1058–1067, 2009.

[6] Masahiro Yukawa, "Multikernel adaptive filtering," *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4672–4682, 2012.

[7] Chafic Saide, Régis Lengelle, Paul Honeine, Cédric Richard, and Roger Achkar, "Nonlinear adaptive filtering using kernel-based algorithms with dictionary adaptation," *International Journal of Adaptive Control and Signal Processing*, vol. 29, no. 11, pp. 1391–1410, 2015.

[8] Konstantinos Slavakis, Sergios Theodoridis, and Isao Yamada, "Online kernel-based classification using adaptive projection algorithms," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 2781–2796, 2008.

[9] Francesco Orabona, Joseph Keshet, and Barbara Caputo, "Bounded kernel-based online learning," *Journal of Machine Learning Research*, vol. 10, no. Nov, pp. 2643–2666, 2009.

[10] Thomas K Paul and Tokunbo Ogunfunmi, "A kernel adaptive algorithm for quaternion-valued inputs," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 10, pp. 2422–2439, 2015.

[11] Badong Chen, Songlin Zhao, Pingping Zhu, and José C Príncipe, "Quantized kernel least mean square algorithm," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 1, pp. 22–32, 2012.

[12] Weifeng Liu, Il Park, and Jose C Principe, "An information theoretic approach of designing sparse kernel adaptive filters," *IEEE Transactions on Neural Networks*, vol. 20, no. 12, pp. 1950–1961, 2009.

[13] Badong Chen, Songlin Zhao, Pingping Zhu, and José C Príncipe, "Mean square convergence analysis for kernel least mean square algorithm," *Signal Processing*, vol. 92, no. 11, pp. 2624–2632, 2012.

[14] Jie Chen, Wei Gao, Cédric Richard, and Jose-Carlos M Bermudez, "Convergence analysis of kernel lms algorithm with pre-tuned dictionary," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7243–7247.

[15] Badong Chen, Junli Liang, Nanning Zheng, and José C Príncipe, "Kernel least mean square with adaptive kernel size," *Neurocomputing*, vol. 191, pp. 95–106, 2016.

[16] Songlin Zhao, Badong Chen, and Jose C Principe, "Kernel adaptive filtering with maximum correntropy criterion," in *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE, 2011, pp. 2012–2017.

[17] Nabil R Yousef and Ali H Sayed, "A unified approach to the steady-state and tracking analyses of adaptive filters," *IEEE Transactions on Signal Processing*, vol. 49, no. 2, pp. 314–324, 2001.

[18] Badong Chen, Lei Xing, Junli Liang, Nanning Zheng, and Jose C Principe, "Steady-state mean-square error analysis for adaptive filtering under the maximum correntropy criterion," *IEEE signal processing letters*, vol. 21, no. 7, pp. 880–884, 2014.