A FIXED POINT UPDATE FOR KERNEL WIDTH ADAPTATION IN INFORMATION THEORETIC CRITERIA

António R. C. Paiva

Scientific Computing and Imaging Institute
University of Utah
Salt Lake City, UT 84112
Email: arpaiva@sci.utah.edu

José C. Príncipe

Computational NeuroEngineering Laboratory
University of Florida
Gainesville, FL 32611
Email: principe@cnel.ufl.edu

ABSTRACT

This paper presents a fixed point update for adaptation of the kernel width parameter in information theoretic criteria. These criteria are typically non-parametric and require a kernel width parameter to be appropriately set. The kernel width sets the smoothing bandwidth for estimation of the probability distribution of the error and, consequently, affects the performance surface. Hence, adaptation of the kernel width allows for the criterion, and its performance surface, to be adjusted to changes in the signal distribution. It is shown that the proposed fixed point update converges faster and is more stable when compared to a gradient update, and has *no parameters*. Moreover, it can be simplified to achieve the same computational complexity as the stochastic gradient update.

1. INTRODUCTION

Adaptive systems learn from examples by optimizing its parameters according to a prescribed criterion or cost function. The criterion measures the performance of the system for a specific task and must be chosen carefully to ensure that the adaptive system learns towards the optimum for the application [1]. The mean squared error (MSE) is the most widely used criterion, primarily due to its ease of mathematical treatment, and because, in some cases, the optimization is convex which leads to closed form solutions. On the other hand, it is only a second order moment of the error. While this ensures optimality of the optimization for Gaussian distributed error, if the error is non-Gaussian, a different set of system parameters may exist which is better suited for the application [2].

One solution is to develop a criterion that accounts for second and higher-order moments. Information theoretic criteria have this property because they take the whole signal distribution into account [3]. Adaptive systems training using information theoretic criteria have been shown to yield better results in a number of applications where the signals are not Gaussian, such as adaptive system training [2], blind source separation [4], and independent component analysis [5].

However, the advantages of information theoretic criteria have been achieved at a cost: the need to set a kernel width[1] parameter for the sample estimators [3]. This parameter controls the smoothing bandwidth in the estimation of the probability distribution of the signal [6] and, consequently, affects the performance surface [7]. If the error distribution changes, for example, as a result of learning or due to a changing environment, the kernel width should be adapted to ensure that the criterion accurately reflects the signal statistics.

Kernel width selection has been discussed extensively in the statistical literature (see Jones et al. [8] for a survey), typically based on the mean integrated squared error (MISE), or its variants, or based on cross-validation methods. Instead, Singh and Principe [9, 10] recently proposed to minimize the Kullback-Leibler (KL) divergence between the estimated and true signal distribution, a strategy that has been shown to compare favorably to optimization of the MISE [9]. They then propose a gradient update rule for adaptation of the kernel width. Although this strategy avoids having to set the kernel width directly and adapts to changes in signal statistics, it still requires a stepsize to be appropriately chosen. Moreover, as shown in our results, choosing the stepsize is not trivial, and represents a tradeoff between tracking ability and stability of the kernel width estimate. Instead, this paper proposes a fixed point update rule for kernel width adaptation that effectively eliminates the stepsize and yields an information theoretic cost without extra free parameters. The results show that this approach is much more stable and has better tracking ability. Finally, all of this is achieved with the computational complexity as the gradient update rule.

The authors would like to thank Abhishek Singh for helpful discussions.

[1] Also known as, kernel size or kernel bandwidth.

2. CRITERION FOR KERNEL WIDTH ADAPTATION

Information theoretic criteria summarize the signal distribution. In these criteria, the kernel width controls the smoothing introduced by a kernel function used for non-parametric estimation of the probability density function (pdf) from samples, as in Parzen windows [6]. Thus, the ideal kernel width should be such that the corresponding pdf estimate approximates the true distribution.

To solve this problem, Singh and Principe [9] proposed to minimize the KL divergence between the true and estimated pdfs, denoted $f(x)$ and $\hat{f}_\sigma(x)$, as a function of the kernel width. That is, to minimize

$$D_{KL}(f\|\hat{f}_\sigma) = \int f(x)\log\left(\frac{f(x)}{\hat{f}_\sigma(x)}\right)dx, \tag{1}$$

where the subscript σ in $\hat{f}$ explicitly shows the dependence of the estimated pdf on the kernel width σ. Expanding eq. 1, we obtain

$$\begin{aligned} &D_{KL}(f\|\hat{f}_\sigma) \\ &\quad= \int f(x)\log(f(x))dx - \int \log(\hat{f}_\sigma(x))f(x)dx \qquad \text{(2a)} \\ &\quad= \int f(x)\log(f(x))dx - E\left[\log(\hat{f}_\sigma(x))\right]. \qquad \text{(2b)} \end{aligned}$$

where $E[\cdot]$ denotes the expectation over x with regards to the true distribution. Since the first term of eq. 2b does not depend on the kernel width, minimizing $D_{KL}(f\|\hat{f}_\sigma)$ with respect to σ is achieved by maximizing the second term. Hence, the optimum kernel width is the σ that maximizes the criterion,

$$J_{KL}(\sigma) = E\left[\log(\hat{f}_\sigma(x))\right]. \tag{3}$$

In practice, $J_{KL}(\sigma)$ must be estimated from samples $\{x_1, x_2, \ldots, x_N\}$. The estimated distribution evaluated at x is given by [6],

$$\hat{f}_\sigma(x) = \frac{1}{N}\sum_{i=1}^{N} K_\sigma(x - x_i), \tag{4}$$

where K_σ is the smoothing kernel function, often taken to be a symmetric probability density, with width σ. In this work, we will consider only the Gaussian kernel,

$$G_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right). \tag{5}$$

Substituting in eq. 3 and approximating the expectation as the mean over the samples, yields the following estimator,

$$\hat{J}_{KL}(\sigma) = \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{1}{N-1}\sum_{j=1, j\neq i} G_\sigma(x - x_j)\right). \tag{6}$$

3. FIXED POINT UPDATE

The optimum kernel width corresponds to the maximum of $J_{KL}(\sigma)$, which can be found by equating to zero its derivative with regards to σ. The derivative of $J_{KL}(\sigma)$ is

$$\frac{\partial J_{KL}(\sigma)}{\partial\sigma} = E\left[\frac{\partial \hat{f}_\sigma(x)/\partial\sigma}{\hat{f}_\sigma(x)}\right] \tag{7}$$

$$= E\left[\frac{\sum_{i=1}^{N}\exp\left(-\frac{(x-x_i)^2}{2\sigma^2}\right)\left(\frac{(x-x_i)^2}{\sigma^3} - \frac{1}{\sigma}\right)}{\sum_{i=1}^{N}\exp\left(-\frac{(x-x_i)^2}{2\sigma^2}\right)}\right]. \tag{8}$$

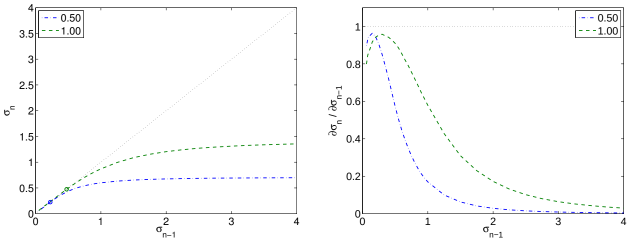
Then, equating to zero yields

$$\begin{aligned} &\frac{\partial J_{KL}(\sigma)}{\partial\sigma} = 0 \\ &\Leftrightarrow \frac{1}{\sigma^2}E\left[\frac{\sum_{i=1}^{N}\exp\left(-\frac{(x-x_i)^2}{2\sigma^2}\right)(x-x_i)^2}{\sum_{i=1}^{N}\exp\left(-\frac{(x-x_i)^2}{2\sigma^2}\right)}\right] = 1 \\ &\Leftrightarrow \sigma = \sqrt{E\left[\frac{\sum_{i=1}^{N}\exp\left(-\frac{(x-x_i)^2}{2\sigma^2}\right)(x-x_i)^2}{\sum_{i=1}^{N}\exp\left(-\frac{(x-x_i)^2}{2\sigma^2}\right)}\right]}. \qquad (9) \end{aligned}$$

Thus, a fixed point update towards the optimum kernel width is

$$\sigma_{n+1} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{j=1,j\neq i}^{N}\exp\left(-\frac{(x_i-x_j)^2}{2\sigma_n^2}\right)(x_i-x_j)^2}{\sum_{j=1,j\neq i}^{N}\exp\left(-\frac{(x_i-x_j)^2}{2\sigma_n^2}\right)}}, \tag{10}$$

where the expectation has been replaced with the sample mean. Note that the fixed point update converges because the mapping has derivative smaller than one and is therefore contractive, as demonstrated in Fig. 1. It is interesting to verify that the proposed fixed point update ensures that the estimated kernel width is non-negative. Note that the



(a) (b)

Fig. 1: Fixed point update (eq. 10) as a function of σ_{n-1}, (a) for signals with the standard deviation shown in the legend, and (b) its corresponding derivative. The '○' marks in (a) indicate the fixed points.

gradient update does not ensure non-negativity and it must be enforced externally.

The main problem with this fixed point update rule is its computational complexity, which is $\mathcal{O}(N^2)$, due to the approximation of the expectation with the sample mean. Typically, however, the kernel width will be adapted together with the system parameters at every update step. Hence, instead of approximating the expectation directly, it can be approximated over time.

Specifically, we propose to approximate the expectation by evaluating its argument for the current sample (with the sums running over the previous $N-1$ samples) and average over the previous N values. This requires that we keep track of the previous N evaluations of the argument of the expectation. Summarizing, the estimated kernel width for the nth update step is,

$$\sigma_n = \sqrt{\frac{1}{N}\sum_{i=n-N+1}^{n} \tilde{\sigma}_i^2}, \qquad (11)$$

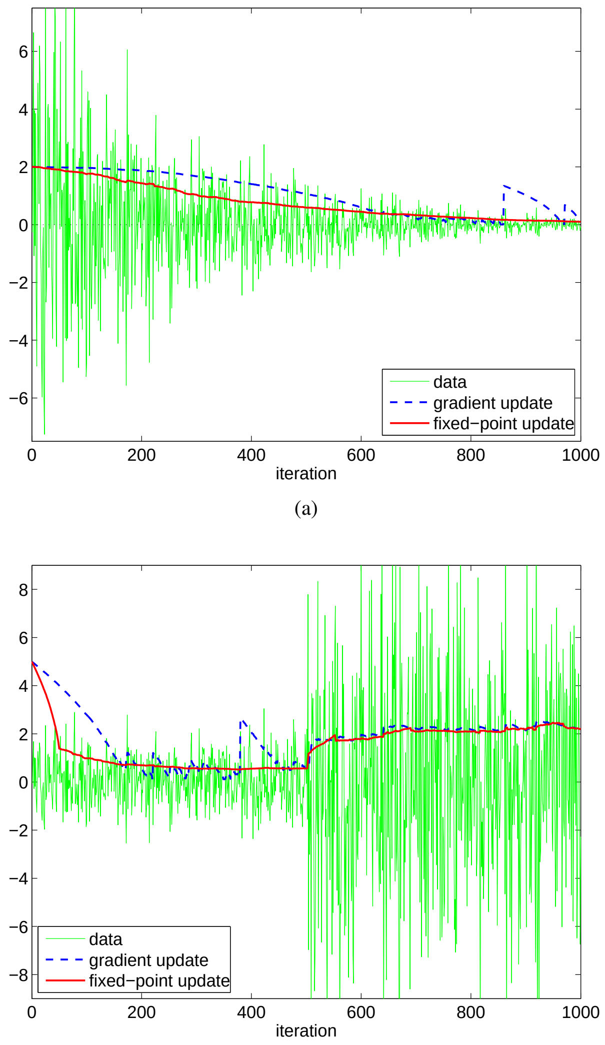
where,

$$\tilde{\sigma}_n^2 = \frac{\sum_{i=n-N+1}^{n-1} \exp\left(-\frac{(x_n-x_i)^2}{2\sigma_{n-1}^2}\right)(x_n-x_i)^2}{\sum_{j=n-N+1}^{n-1} \exp\left(-\frac{(x_n-x_i)^2}{2\sigma_{n-1}^2}\right)}. \qquad (12)$$

In essence, this approach is equivalent to using an averaging filter to smooth noisy estimates of the square of the kernel width, $\tilde{\sigma}_i^2$. Note that the memory depth of the averaging was set equal to the number of samples for density estimation N, since this parameter already controls the temporal resolution of the gradient on the system parameters. Moreover, in this way one avoids dependence on additional parameters. Obviously, this parameter can be fine tuned, but a user does not need to and setting it is very intuitive. All our results use the default value.

The computation complexity of the fixed point update described in eqs. 11 and 12 is only $\mathcal{O}(N)$. This is just like the stochastic gradient update proposed in [9, 10], and avoids the need to set a stepsize. The only caveat being the need to store N intermediate values of $\tilde{\sigma}_i^2$. Nevertheless, since in practice N is a relatively small number, typically around 100 or less, this is a minor issue.

Integrating the kernel width update rule in the learning process of an adaptive system follows naturally. At every step, one starts by updating the estimate of the kernel width given the latest sample. Then, the system's parameters are updated according to the information theoretic criteria update strategy using the estimated kernel size.



(a)

(b)

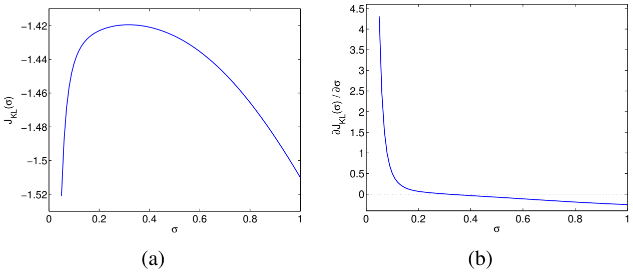
Fig. 2: Comparison of the estimated kernel width using the gradient and fixed point update of non-stationary signals. Two cases are shown: (a) the power of the signal decays exponentially, and (b) the signal has an abrupt change in power.

4. SIMULATION RESULTS

The fixed point update is now compared to the stochastic gradient update[2] for the estimation of the kernel width on non-stationary signals. We show that the fixed point update rule converges faster to the optimum kernel width and yields a more stable estimate.

The examples shown here mimic typical situations where the error signal of an adaptive system changes over time; i.e., the signal is non-stationary. This can be a normal consequence of the training (Fig. 2(a)), or due to changes in the

[2] The kernel width was adapted by doing stochastic gradient ascend according to eq. 8, after dropping the expectation [9, 10].



(a)

(b)

Fig. 3: (a) Kernel width adaptation criterion $J_{KL}(\sigma)$ and (b) its derivative, as functions of σ for a Gaussian signal with unit variance.

environment which cause a mismatch with the learned system, such as those often encountered in mobile communication systems (Fig. 2(b)). Either scenario is modeled here with a Gaussian signal with time-varying power.

As shown in Fig. 2, the fixed point update converges rapidly to the optimum kernel width and is quite stable. In contrast, the adaptation of the kernel width using the gradient update is considerably slower and exhibits occasionally sudden jumps. This is due to the skewness in the criterion, as illustrated in Fig. 3(a), which can yield very different gradient magnitudes depending on the value of the current kernel width with regards to the optimum, as shown in Fig. 3(b). Consequently, when the kernel width is larger than the optimum the gradient is small and the convergence slow, but for small kernel widths the gradient is much larger, yielding the observed jumps. This problem is particularly noticeable for small kernel widths because the optimum kernel width is closer to the region with high gradient magnitudes. Clearly, reducing the stepsize for the gradient update could potentially solve the problem but then the convergence would become even slower. It should be remarked that, as suggested in [9], a small regularization constant of $\varepsilon = 0.01$ was added to the denominator of the stochastic gradient of $J_{KL}(\sigma)$. This already helps to mitigate the observed jumps by preventing large gradient values due to the denominator becoming very small. This is encountered if the current sample is very different from the previous, which may occur, for example, if the signal power increases abruptly. Interestingly, we found empirically that the fixed point update operates better without this regularization term.

5. CONCLUSION

This paper presents a fixed point update for adaptation of the kernel width parameter. Our simulation results show that the proposed update rule converges significantly faster than the gradient update previously presented by Singh and Principe [9, 10], and yields much more stable kernel width estimates. In addition to these improvements, the proposed fixed point update has the advantage that it eliminates the need to specify the learning rate for kernel width adaptation without introducing any additional parameters. This is unlike the gradient update rule which in essence substitutes the problem of setting the kernel width by that of setting a stepsize. Even though the later is preferable, the fixed point update effectively saves the user the need to specify one parameter compared to the typically use of information theoretic criteria. Finally, all of these advantages are attained with the same computation complexity, $\mathcal{O}(N)$, as the gradient update.

6. REFERENCES

[1] José C. Principe, Neil R. Euliano, and W. Curt Lefebvre, *Neural and Adaptive Systems: fundamentals through simulations*, John Wiley & Sons, 2000.

[2] Deniz Erdogmus and José C. Príncipe, "An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems," *IEEE Trans. Signal Processing*, vol. 50, no. 7, pp. 1780–1786, July 2002.

[3] Jose C. Principe, *Information Theoretic Learning*, Springer, 2010.

[4] Anthony J. Bell and Terrence J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comp.*, vol. 7, no. 6, pp. 1129–1159, Nov. 1995.

[5] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja, *Independent Component Analysis*, Wiley, 2001.

[6] Emanuel Parzen, "On the estimation of a probability density function and the mode," *Ann. Math. Stat.*, vol. 33, no. 2, pp. 1065–1076, Sept. 1962.

[7] Deniz Erdogmus and José C. Príncipe, "Convergence properties and data efficiency of the minimum error entropy criterion in adaline training," *IEEE Trans. Signal Processing*, vol. 51, no. 7, pp. 1966–1978, July 2003.

[8] M. C. Jones, J. S. Marron, and S. J. Sheather, "A brief survey of bandwidth selection for density estimation," *J. Am. Stat. Assoc.*, vol. 91, no. 433, pp. 401–407, Mar. 1996.

[9] Abhishek Singh and Jose C. Principe, "Information theoretic learning with adaptive kernels," *Signal Proc.*, 2010, Accepted.

[10] Abhishek Singh and Jose C. Principe, "Kernel width adaptation in information theoretic cost functions," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Dallas, TX, USA, Mar. 2010.