A Fixed Point Update for Kernel Width Adaptation in Information Theoretic Criteria

António R. C. Paiva and José C. Príncipe

Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, UT
Computational NeuroEngineering Lab, University of Florida, Gainesville, FL
arpaiva@sci.utah.edu, principe@cnel.ufl.edu

Abstract

- Information theoretic criteria generalize the traditional mean squared error (MSE) criterion, and have been shown to yield better results in a number of applications.
- However, for information theoretic criteria to achieve these improvements, a kernel width parameter needs to be appropriately set.
- This paper presents a fixed point update for adaptation of the kernel width parameter and introduces *no additional parameters*.
- Adaptation of the kernel width allows for the information theoretic criterion, and its performance surface, to be adjusted to changes in the signal distribution.

Criterion for kernel width adaptation

- In information theoretic criteria, the kernel width controls the smoothing used for non-parametric density estimation, as in Parzen windows.
- Hence, Singh and Principe [1, 2] proposed to minimize the Kullback-Leibler divergence between the true and estimated pdf, denoted $f(x)$ and $f_\sigma(x)$. That is,

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

- Since the first term does not depend on σ, the optimum kernel width maximizes,

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

- Plugging in Parzen's pdf estimation on N samples with the Gaussian smoothing kernel, it yields the estimator,

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

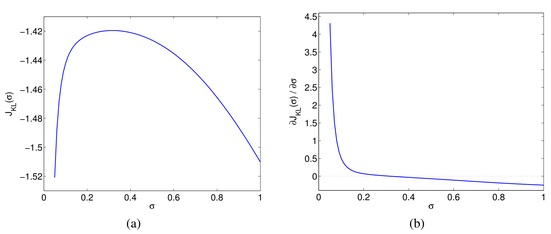

Figure 1: (a) Kernel width adaptation criterion $J_{KL}(\sigma)$ and (b) its derivative, as functions of σ for a Gaussian signal with unit variance.

Fixed point update of the kernel width

- The maximum of $J_{KL}(\sigma)$ can be found by equating its derivative to zero.
- The derivative of $J_{KL}(\sigma)$ is

$$\begin{aligned} \frac{\partial J_{KL}(\sigma)}{\partial\sigma} &= E\left[\frac{\partial\hat{f}_\sigma(x)/\partial\sigma}{\hat{f}_\sigma(x)}\right] \\ &= E\left[\frac{\sum_{i=1}^{N}\exp\left(-\frac{(x-x_i)^2}{2\sigma^2}\right)\left(\frac{(x-x_i)^2}{\sigma^3}-\frac{1}{\sigma}\right)}{\sum_{i=1}^{N}\exp\left(-\frac{(x-x_i)^2}{2\sigma^2}\right)}\right]. \end{aligned}$$

Equating to zero, it yields

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

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

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

by replacing the expectation with the sample mean. However, this update rule has computational complexity $\mathcal{O}(N^2)$.

- For online adaptation, we propose to approximate the expectation over time instead. Accordingly, the estimated kernel width for the nth update step is

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

where,

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

The computational complexity of this update rule is only $\mathcal{O}(N)$.

- Figure 2 suggests that the mapping converges because it has derivative smaller than one and, thus, is contractive.

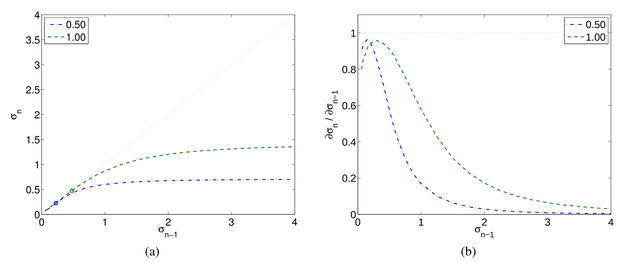

Figure 2: Fixed point update as a function of σ_{n-1}, (a) for signals with the standard deviation shown in the legend, and (b) its corresponding derivative.

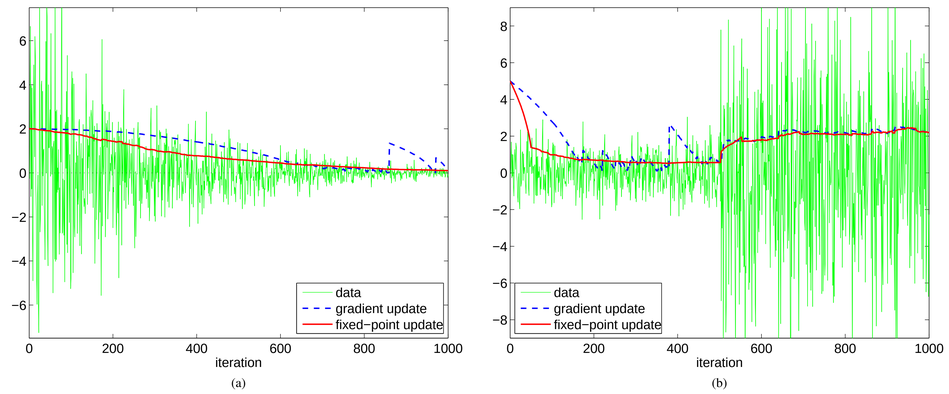

Figure 3: Comparison of the estimated kernel width using the gradient and fixed point update of non-stationary signals. Two cases are shown: (a) the power of the signal decays exponentially, and (b) the signal has an abrupt change in power.

Results

- The fixed point update was compared to the stochastic gradient update, as proposed in [1, 2], for the estimation of the kernel width on non-stationary signals.
- The examples recreate typical situations where the error signal of an adaptive system changes over time. This can be a normal consequence of the training (Fig. 3(a)), or due to changes in the environment which cause a mismatch with the learned system, such as those often encountered in mobile communication systems (Fig. 3(b)).
- Figure 3 shows that the fixed point update rule converges faster to the optimum kernel width and yields a more stable estimate.
- 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, which yields very different gradient magnitudes depending on the value of the current kernel width with regards to the optimum, as shown in Figure 1(a).

Conclusion

- This paper presents a fixed point update for adaptation of the kernel width parameter which can be simplified to the same computation complexity, $\mathcal{O}(N)$, as the gradient update.
- Our simulation results show that the proposed update rule converges significantly faster than the gradient update previously proposed, and yields much more stable kernel width estimates.
- In addition, the proposed fixed point update eliminates the need to specify the learning rate for kernel width adaptation without introducing any additional parameters.

References

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

[2] 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.