Estimating the Information Potential with the Fast Gauss Transform

Seungju Han, Sudhir Rao, and Jose Principe

CNEL, Department of Electrical and Computer Engineering, University of Florida, Gainesville, USA
{han, sudhir, principe}@cnel.ufl.edu
http://www.cnel.ufl.edu

Abstract. In this paper, we propose a fast and accurate approximation to the information potential of Information Theoretic Learning (ITL) using the Fast Gauss Transform (FGT). We exemplify here the case of the Minimum Error Entropy criterion to train adaptive systems. The FGT reduces the complexity of the estimation from $O(N^2)$ to $O(pkN)$ where p is the order of the Hermite approximation and k the number of clusters utilized in FGT. Further, we show that FGT converges to the actual entropy value rapidly with increasing order p unlike the Stochastic Information Gradient, the present $O(pN)$ approximation to reduce the computational complexity in ITL. We test the performance of these FGT methods on System Identification with encouraging results.

1 Introduction

Information Theoretic Learning (ITL) is a methodology to non-parametrically estimate entropy and divergence directly from data, with direct applications to adaptive systems training [1]. The centerpiece of the theory is a new estimator for Renyi's quadratic entropy that avoids the explicit estimation of the probability density function. The argument of the logarithm of Renyi's entropy is called the Information Potential (IP), and since the logarithm is a monotonic function, it is sufficient to use the IP in training [2]. ITL has been used in ICA [3], blind equalization [4], clustering [5], and projections that preserve discriminability [6]. One of the difficulties of ITL is that the calculation of the IP is $O(N^2)$, which may become prohibitive for large data sets. A stochastic approximation of the IP called the Stochastic Information Gradient (SIG) [7] decreases the complexity to $O(N)$, but slows down training due to the noise in the estimate. This paper presents an effort to make the estimation faster and more accurate using the Fast Gauss Transform (FGT). The FGT is one of a class of very interesting and important new families of fast evaluation algorithms that have been developed over the past dozen years to enable rapid calculation of approximations at arbitrary accuracy to matrix-vector products of the form $\boldsymbol{A}d$ where $a_{ij} = \Phi(|x_i - x_j|)$ and Φ is a particular special function. These sums first arose in astrophysical observations where the function Φ was the gravitational field. The basic idea is to cluster the sources and target points using appropriate data structures, and to replace the sums

J. Rosca et al. (Eds.): ICA 2006, LNCS 3889, pp. 82 – 89, 2006.

with smaller summations that are equivalent to a given level of precision. We will use here the FGT algorithm proposed by Greengard and Strain [8] and the *farthest-point clustering* proposed by Gonzalez [9] for evaluating Gaussian sums.

The paper will be organized as follows. First we will briefly describe one of the simplest ITL algorithms that minimize the error entropy between a desired response and the adaptive filter output. Next, we present the FGT algorithm and its interaction with the MEE criterion, followed by some simulation results and conclusions.

2 Minimum Error Entropy (MEE)

Suppose that the adaptive system is an FIR structure with a weight vector $\mathbf{w}$. The error samples are $e_k = d_k - \mathbf{w}_k^T \mathbf{u}_k$, where d_k is the desired response, and $\mathbf{u}_k$ is the input vector. The error PDF is estimated using Parzen windows as

$$\hat{f}_e(e) = \frac{1}{N}\sum_{i=1}^{N} k_\sigma(e - e_i) \tag{1}$$

where $k_\sigma(\cdot)$ is kernel function with a kernel size σ. So, Renyi's quadratic entropy estimator for a set of discrete data samples becomes:

$$H_{R2}(e) = -\log \int f^2(e)\, de = -\log V(e) \tag{2}$$

$$V(e) = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} k_{\sigma\sqrt{2}}(e_j - e_i). \tag{3}$$

Minimizing the entropy in (2) is equivalent to maximizing the information potential since the log is a monotonic function. Thus, the weight update of MEE is

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \mu \nabla V(e) \tag{4}$$

where for a Gaussian kernel the gradient is,

$$\nabla V(e) = \frac{1}{2\sigma^2 N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} G_{\sigma\sqrt{2}}(e_j - e_i)\{e_j - e_i\}\{\mathbf{u}_j - \mathbf{u}_i\}. \tag{5}$$

For online training methods, the information potential can be estimated using the Stochastic Information Gradient (SIG) as shown in (6), where the sum is over the most recent *L* samples at time *k*. Thus for a filter order of length *M*, the complexity of MEE is equal to O(*ML*) per weight update,

$$V(e) \approx \frac{1}{L}\sum_{i=k-L}^{k-1} k_{\sigma\sqrt{2}}(e_k - e_i) \tag{6}$$

where $e_i = d_i - \mathbf{w}_k^T \mathbf{u}_i$, for $k - L \le i \le k$.

3 MEE Using the Fast Gauss Transform

For efficient computation of information potential, we use the principle of Fast Gauss Transform. Direct evaluation of the information potential (3) requires O(N^2). We apply the FGT idea by using the following expansions for the Gaussian in one dimension (the method can be easily extended to multiple dimensions):

$$\exp\left(-\frac{(e_j-e_i)^2}{4\sigma^2}\right)=\sum_{n=0}^{p-1}\frac{1}{n!}\left(\frac{e_i-s}{2\sigma}\right)^n h_n\left(\frac{e_j-s}{2\sigma}\right)+\varepsilon(p) \tag{7}$$

where the Hermite function $h_n(x)$ is defined by

$$h_n(x)=(-1)^n\frac{d^n}{dx^n}\left(\exp\left(-x^2\right)\right). \tag{8}$$

In practice a single expansion about one center is not always valid or accurate over the entire domain. A space subdivision scheme is applied in the FGT and the Gaussian functions are expanded at multiple centers. To efficiently subdivide the space, we use a very simple greedy algorithm, called *farthest-point clustering* that computes a data partition with a maximum radius at most twice the optimum. The direct implementation of farthest-point clustering has running time O(*kN*), which *k* is the number of clusters. Thus, the information potential $V(e)$ is given as

$$V(e)\approx\frac{1}{2\sigma N^2\sqrt{\pi}}\sum_{j=1}^{N}\sum_{B}\sum_{n=0}^{p-1}\frac{1}{n!}h_n\left(\frac{e_j-s_B}{2\sigma}\right)C_n(B) \tag{9}$$

where B is a cluster with center s_B and $C_n(B)$ is defined by

$$C_n(B)=\sum_{e_i\in B}\left(\frac{e_i-s_B}{2\sigma}\right)^n. \tag{10}$$

From the above equation, we can see that the total number of operations required is O(*pkN*) per data dimension. The truncation order *p* depends on the desired accuracy alone, and is independent of *N*.

The gradient of the information potential with respect to the weights is given as

$$\nabla V(e)=\frac{1}{2\sigma N^2\sqrt{\pi}}\sum_{j=1}^{N}\sum_{B}\sum_{n=0}^{p-1}\frac{1}{n!}\left[h_{n+1}\left(\frac{e_j-s_B}{2\sigma}\right)\left[\frac{\mathbf{u}_j}{2\sigma}\right]\cdot C_n(B)+h_n\left(\frac{e_j-s_B}{2\sigma}\right)\cdot\nabla C_n(B)\right] \tag{11}$$

where $\nabla C_n(B)$ is defined by

$$\nabla C_n(B)=\sum_{e_i\in B}n\left(\frac{e_i-s_B}{2\sigma}\right)^{n-1}\left[-\frac{\mathbf{u}_i}{2\sigma}\right]. \tag{12}$$

4 Simulations

4.1 Entropy Estimation Using Fast Gauss Transform

We start by analyzing the accuracy of the FGT in the calculation of the IP for the Gaussian and Uniform distributions, using the original definition (3), the SIG (6) and the FGT approximation (9) for two sample sizes (100 and 1,000 samples). For a comparison between SIG and FGT we use $p = L$ in all our simulations. We fix the radius of the farthest point clustering algorithm at $r = \sigma$. This radius is related to the number of clusters, i.e., as the radius increases, the number of clusters (hence the computation time) decreases, but the approximation accuracy may suffer. Results are depicted in Fig. 1 and 2.

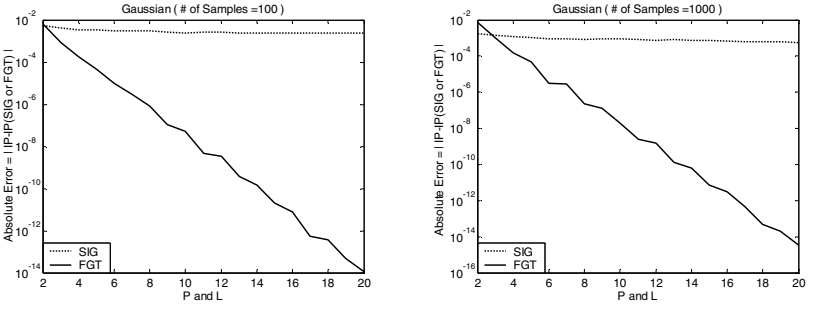

Fig. 1. Plot of the absolute error for SIG and FGT with respect to the IP estimated using Parzen window for a Gaussian distribution with 100 and 1000 samples

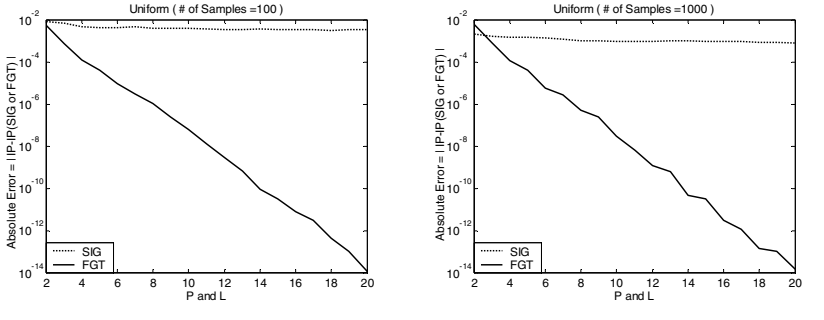

Fig. 2. Plot of the absolute error for SIG and FGT with respect to the IP estimated using Parzen window for a uniform distribution with 100 and 1000 samples

As can be observed in Fig.1 and 2, the absolute error between the IP and the FGT estimation decreases with the order p of the Hermite expansion to very small values, while that of the SIG fluctuates around 0.005 (100 samples) and 0.001 (1000 samples). We can conclude that from a strictly absolute error point of view, a FGT with order p >3 outperforms the SIG method for all cases.

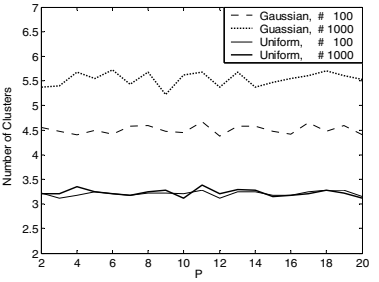

Fig. 3. Plot of the average number of clusters in FGT when estimating the IP for the Gaussian and uniform distribution with 100 and 1000 samples (40 times Monte Carlo)

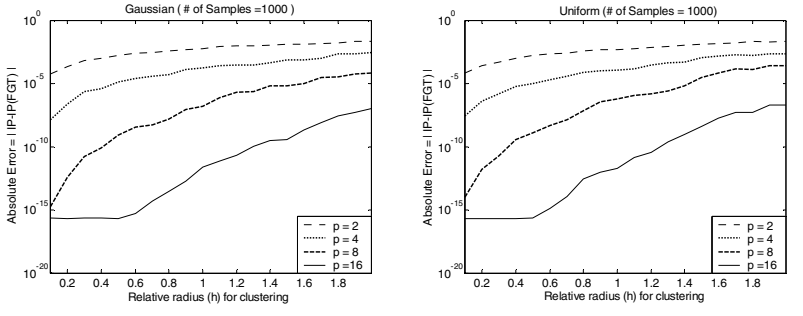

Fig. 4. Plot of the absolute error for a given p (=2, 4, 8 and 16) as the radius of the farthest-point clustering algorithm ($r = h \times \sigma$) for Gaussian and uniform distribution with 1000 samples

Fig. 3 shows the relation between FGT estimation and the number of clusters. According to data size, the number of clusters does not vary for the uniform distribution, while for the Gaussian distribution the number of cluster is larger as the number of data samples increases.

We also fix the number of points to N=1000 and vary the radius r for clustering from 0.1σ to 2σ and plot the absolute error for a given p (= 2, 4, 8 and 16) in Fig. 4. The results show that the error of the FGT is reduced as the radius decreases, as expected such that the user can control the approximation error to IP.

However, for our ITL application, the accuracy of the IP is not the primary objective. Indeed, in ITL we would like to train adaptive systems using gradient information, so the smoothness of the cost function is perhaps more important.

4.2 System Identification

We next consider the system identification of a moving-average model with a 9th order transfer function given by

$$H(z) = 0.1 + 0.2z^{-1} + 0.3z^{-2} + 0.4z^{-3} + 0.5z^{-4} + 0.4z^{-5} + 0.3z^{-6} + 0.2z^{-7} + 0.1z^{-8} \tag{13}$$

using the minimization of the error entropy [10]. Although the true advantage of MEE is for nonlinear system identification with nonlinear filters, here the goal is to compare adaptation accuracy and speed so we elected to use a linear plant and a FIR adaptive filter with the same plant order (zero achievable error). A standard method of comparing the performance in system identification problems is to plot the weight error norm since this is directly related to misadjustment. In each case the power of the weight noise was plotted versus the number of epochs performed. In this simulation, the inputs to both the plant and the adaptive filter are also white Gaussian or uniform noise. We choose a proper kernel size by using Silverman's rule ($\sigma = 0.707$) the radius of the farthest point clustering algorithm $r=\sigma$.

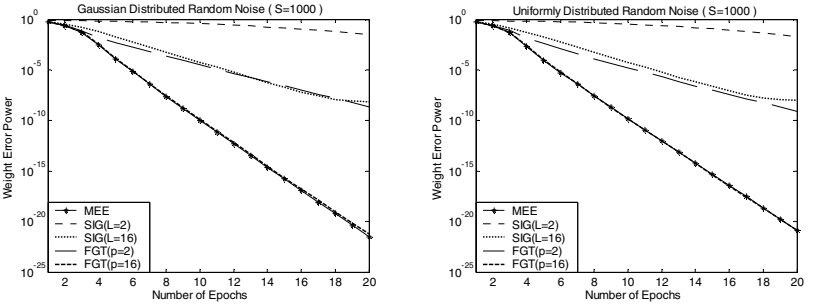

Fig. 5. Comparison of different methods for system identification with Gaussian and uniform noise, using (S) = 1000 samples

As can be observed in Fig. 5, all the versions of IP produce converging filters. However, the speed of convergence and the actual value of the final error are different. The FGT method performs better in training the adaptive system as compared to SIG. A SIG with 16 samples approaches the FGT with $p=2$, and the FGT with $p=16$ is virtually identical to the true IP. The case of the uniform input noise does not change the conclusions.

Fig. 6. shows the plot of the number of clusters during adaptation. Since the error is decreasing at each epoch, the number of clusters gets progressively smaller. In this case, where the achievable error is zero, the number reduces to one cluster after 5 epochs.

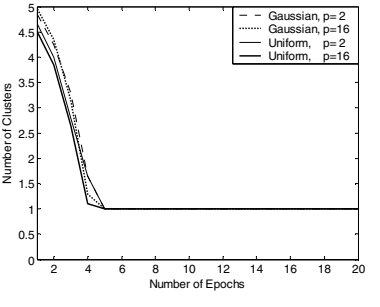

Fig. 6. Plot of the average number of clusters during adaptation in system identification

5 Conclusions

Information Theoretic Learning and in particular the Minimum Error Entropy criterion has been recently proposed as a more principled approach for training adaptive systems. But, a major bottleneck in this method is the high computational complexity of $O(N^2)$ per epoch, thus limiting its use for many practical applications in signal processing, communications and machine learning. The method of the Fast Gaussian Transform helps alleviate this problem by accurate and efficient computation of entropy using the Hermite series expansion in $O(pN)$ operations. Furthermore, since this series converges rapidly, a small order p gives a very good approximation of the IP and can therefore provide accurate and fast converging optimal filters. Indeed we have shown that the FGT has a performance virtually identical to the exact information potential for p=16. The FGT seems therefore to be preferable to the SIG algorithm we have been using.

We still need to quantify the performance of FGT for training MIMO (multiple input multiple output) systems such in ICA or discriminative projections. In these cases ITL algorithms will be applied to multidimensional signals and the computation becomes prohibitive. A straight application of the algorithm presented in this paper will raise p to the number of dimensions in the complexity calculation. However, recent results show that it is possible to avoid the multiplicative factor in complexity brought by the dimensionality of the space of interactions [11]. If further testing corroborates these initial results, the class of FGT algorithms may very well take away the computational drawback of ITL versus the MSE criterion to adapt nonlinear models both in Adaptive Systems and Pattern Recognition applications.

Acknowledgement. This work was partially supported by NSF grant ECS-0300340.

References

1. Principe, J.C., Xu, D., Fisher, J.: Information Theoretic Learning. In: Haykin, S., Unsupervised Adaptive Filtering. Wiley, New York, Vol.I, (2000) 265-319
2. Principe, J.C., Xu, D.: Information-Theoretic Learning Using Renyi's Quadratic Entropy. In Proceedings of the 1st Int. Workshop on Independent Component Analysis and Signal Separation, Aussois, France, (January 11-15 1999) 407-412
3. Hild, K.E., Erdogmus, D., Principe, J.C.: Blind Source Separation using Renyi's Mutual Information. IEEE Signal Processing Letters, Vol. 8, No. 6, (2001) 174-176
4. Lazaro, M., Santamaria, I., Erdogmus, D., Hild, K.E., Pantaleon, C., Principe, J.C.: Stochastic Blind Equalization Based on PDF Fitting Using Parzen Estimator. IEEE Transactions on Signal Processing, Vol. 53, No. 2, (Feb 2005) 696-704
5. Jenssen, R., Eltoft, T. Principe, J.C.: Information Theoretic Spectral Clustering. In Proc. Int. Joint Conference on Neural Networks, Budapest, Hungary, (Jul 2004) 111-116
6. Torkkola, K.: Learning discriminative feature transforms to low dimensions in low dimensions. In advances in neural information processing systems 14, Vancouver, BC, Canada, (Dec 3-8 2001a) MIT Press
7. Erdogmus, D., Principe, J.C., Hild, K.E.: Online entropy manipulation: stochastic Information Gradient. IEEE Signal Processing Letters, Vol. 10, No. 8, (Aug 2003) 242-245

8. Greengard, L., Strain, J.: The fast Gauss transform. SIAM J. Sci. Statist. Comput. 12 (1991) 79-94
9. Gonzalez. T.: Clustering to minimize the maximum intercluster distance. Theoretical Computer Science, 38 (1985) 293-306
10. Erdogmus, D., Principe, J.C.: An Entropy Minimization algorithm for Supervised Training of Nonlinear Systems. IEEE trans. of Signal Processing, Vol. 50, No. 7, (Jul 2002) 1780-1786
11. Yang, C., Duraiswami, R., Gumerov, N., Davis, L.: Improved fast gauss transform and efficient kernel density estimation. In 9th Int. Conference on Computer Vision, Vol. 1, (2003) 464-471