Rejected by NIPS'99
Published as TR
Will probably be submitted to IJCNN or ICA-2000

Fast Fixed-Point Algorithms for Bayesian Blind Source Separation

Harri Lappalainen
Neural Networks Research Centre
Helsinki University of Technology, Finland
Harri.Lappalainen@hut.fi

Abstract

In this paper, various fast algorithms for linear blind source separation (BSS) are developed. The fast fixed-point algorithm for independent component analysis [8, 6] is interpreted as an EM-algorithm with a modification for accelerating the convergence rate which can otherwise be slow particularly in a low noise environment. The new point of view opens way for developing several new fast fixed-point algorithms for extracting signals with various properties. A Bayesian version of the ordinary independent component analysis and a version which takes into account both the time-domain behavior and non-Gaussianity of the source signals are studied in more depth. Both extensions have practical importance. The Bayesian version can be used for optimizing model structure and comparing different hypotheses. Natural signals have typically non-Gaussian distributions and time-dependencies. There has, therefore, been a demand for an algorithm which can utilize both types of information.

1 Introduction

The objective of the linear blind source separation (BSS) is to find the unknown source signals which appear in the data as linear but unknown mixtures. The basic data model is

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \boldsymbol{\nu}(t), \tag{1}$$

where $\mathbf{x}(t)$ denotes the vector of observed signals at time t, $\mathbf{s}(t)$ denotes the vector of source signals, A the mixing matrix and $\boldsymbol{\nu}(t)$ the additive noise. Equation 1 can be written also in matrix form as

$$\mathbf{X} = \mathbf{AS} + \mathbf{N}. \tag{2}$$

It is well known that the BSS problem is ill posed if the sources are i.i.d. with Gaussian distributions. It is possible, however, to retrieve sources which have non-Gaussian distributions or have time-dependencies. Using the non-Gaussianity of the source distributions for BSS is called independent component analysis (ICA) [9, 4, 2, 8, 11]. BSS using time-correlations of the source signals does not have an established name but it has been used, for instance, in [15, 13, 3].

The new BSS algorithms developed in this paper are generalizations of the fast fixed-point algorithm for ICA (FastICA) [8, 6]. Section 2 introduces the FastICA and gives it an interpretation which allows new generalizations.

The first generalization is the Bayesian FastICA which will be developed in section 3. The Bayesian approach makes it possible to find the optimal structure for the ICA model. It is, for instance, easy to find out what is the most probable number of source signals.

Section 4 introduces a BSS algorithm which makes use of both non-Gaussianity and time-dependence of the source signals. The resulting algorithm is likely to be very useful since a large number of real world signals are non-Gaussian and have time-dependencies. Finally, some other extensions are discussed in section 5.

2 FastICA as an asymptotic EM-algorithm

The fixed-point algorithm [8, 6] is, to the best of our knowledge, the fastest existing ICA algorithm. We are going to derive it as an EM-algorithm with a modification which will yield a good convergence rate even at the limit of the noise variance $\mathrm{Var}(\boldsymbol{\nu})$ approaching zero. From the given point of view it will then be evident how to apply a similar modification to a large number of other algorithms.

The idea behind the EM-algorithm is to estimate the mixing matrix $\mathbf{A}$ and the sources $\mathbf{s}(t)$ alternately. The algorithm proceeds in alternating steps where first $\mathbf{A}$ is fixed and $\mathbf{s}(t)$ is estimated. Then $\mathbf{s}(t)$ is fixed and a new $\mathbf{A}$ is estimated.

Let us see what happens if we assume there is no additive noise on the data, i.e., the noise term $\mathbf{N}$ is left out from equation 2 and the data model is simply $\mathbf{X} = \mathbf{AS}$. Assuming $\mathbf{A}$ fixed, the estimate of the sources is $\hat{\mathbf{S}} = \mathbf{A}^\dagger\mathbf{X}$. Then the new estimate of the mixing matrix $\mathbf{A}$ will be $\hat{\mathbf{A}} = \mathbf{X}\hat{\mathbf{S}}^\dagger = \mathbf{X}\mathbf{X}^\dagger\mathbf{A} = \mathbf{A}$. This means that the estimate of $\mathbf{A}$ does not change at all in the alternating process.

In order for the EM-algorithm to work, the data has to have noise and the estimation of the $\mathbf{s}(t)$ has to use information which will filter away some of the noise present in the sources. In the case of ICA, for instance, the estimation of the sources should make use of the assumption of non-Gaussianity of the source signals. It can be said, in general, that the more noise is filtered away at each step, the faster the algorithm will converge.

Let us denote the estimate of the sources in the case of no additive noise by $\hat{\mathbf{S}}^{\mathbf{0}} = \mathbf{A}^\dagger\mathbf{X}$. If we assume the noise is Gaussian with the variance $\mathrm{Var}(\boldsymbol{\nu}) = \epsilon$, where ϵ is small, then the estimated mean of the sources $\hat{\mathbf{S}}$ will be close to $\mathbf{S}^{\mathbf{0}}$. To see this, let us consider the posterior distribution of the source $s_i(t)$ on a logarithmic scale.

$$\ln p_i(s_i(t)|\mathbf{A}, \mathbf{x}(t)) \approx \ln p_i(s_i(t)) - \frac{(s_i(t) - \hat{s}_i^0(t))^2}{2C_i\epsilon} \tag{3}$$

Here $p_i(s)$ is the distribution of the ith source signal and C_i is a constant depending on $\mathbf{A}$.

When the noise is small and Gaussian, the posterior distribution of each source is also close to Gaussian with a correspondingly small variance. The mean $\hat{s}_i(t)$ of the posterior can be solved by differentiating equation 3 with respect to $s_i(t)$ and solving for zero. This will result in $\hat{s}_i(t) \approx \hat{s}_i^0(t) + C_i\epsilon\frac{\partial}{\partial s_i}\ln p_i(\hat{s}_i^0(t))$.

Let us denote $f_i(s) = C_i\frac{\partial}{\partial s}\ln p_i(s)$ and denote by $\mathbf{F}(\mathbf{S})$ the matrix whose entries are $f_i(s_i(t))$. Using this notation, the estimated posterior mean of $\mathbf{S}$ is approximately equal to $\hat{\mathbf{S}} \approx \hat{\mathbf{S}}^{\mathbf{0}} + \epsilon\mathbf{F}(\hat{\mathbf{S}}^{\mathbf{0}})$. The new estimate for $\mathbf{A}$ will then be $\hat{\mathbf{A}} = \mathbf{X}\hat{\mathbf{S}}^\dagger$ and it is fairly easy to show that $\hat{\mathbf{S}}^\dagger$ approximately equals $\hat{\mathbf{S}}^{0\dagger} + \epsilon\mathbf{F}(\hat{\mathbf{S}}^{\mathbf{0}})^T(\hat{\mathbf{S}}^{\mathbf{0}}\hat{\mathbf{S}}^{0T})^{-1}$ and therefore $\hat{\mathbf{A}} \approx \mathbf{A} + \epsilon\Delta\mathbf{A}$, where $\Delta\mathbf{A} = \mathbf{X}\mathbf{F}(\hat{\mathbf{S}}^{\mathbf{0}})^T(\hat{\mathbf{S}}^{\mathbf{0}}\hat{\mathbf{S}}^{0T})^{-1}$. It is evident that the changes made to the mixing matrix $\mathbf{A}$ at each step of the EM-algorithm will become smaller as the noise variance ϵ approaches zero.

Recall now that the estimation step $\hat{\mathbf{S}} \approx \hat{\mathbf{S}}^{\mathbf{0}} + \epsilon\mathbf{F}(\hat{\mathbf{S}}^{\mathbf{0}})$ can be viewed as noise removal. According to this view, only a small part of the noise will be filtered away when ϵ is small. It is, however, possible to check how much of the noise is removed by seeing what would happen if the observations $\mathbf{X}$ were replaced by pure Gaussian noise having the same covariance matrix as $\mathbf{X}$. There should then be no meaningful signal to detect and, ideally, the filtering should remove the signal completely.

Let $\mathbf{X}_\nu$ denote the Gaussian noise which has the same covariance matrix as $\mathbf{X}$, i.e., $\mathbf{X}_\nu\mathbf{X}_\nu^T = \mathbf{X}\mathbf{X}^T$. Let $\hat{\mathbf{S}}_\nu^{\mathbf{0}}$ denote the corresponding estimate of the noise free case: $\hat{\mathbf{S}}_\nu^{\mathbf{0}} = \mathbf{A}^\dagger\mathbf{X}_\nu$. Since $\mathbf{X}$ and $\mathbf{X}_\nu$ have the same covariances, also $\hat{\mathbf{S}}^{\mathbf{0}}$ and $\hat{\mathbf{S}}_\nu^{\mathbf{0}}$ have: $\hat{\mathbf{S}}_\nu^{\mathbf{0}}\hat{\mathbf{S}}_\nu^{0T} = \hat{\mathbf{S}}^{\mathbf{0}}\hat{\mathbf{S}}^{0T}$. Finally, let $\hat{\mathbf{A}}_\nu = \mathbf{A} + \epsilon\Delta\mathbf{A}_\nu$, where $\Delta\mathbf{A}_\nu = \mathbf{X}_\nu\mathbf{F}(\hat{\mathbf{S}}_\nu^{\mathbf{0}})^T(\hat{\mathbf{S}}^{\mathbf{0}}\hat{\mathbf{S}}^{0T})^{-1}$. If the noise were removed completely the matrix $\hat{\mathbf{A}}_\nu$ would vanish. Therefore, $\hat{\mathbf{A}}_\nu$ can be used for measuring the amount of noise left by the filtering.

In the FastICA algorithm, the convergence is enhanced by subtracting $\hat{\mathbf{A}}_\nu$ from $\hat{\mathbf{A}}$ and then renormalizing the column vectors of the resulting matrix. This yields a fixed-point algorithm whose convergence points are the same as for the original EM-algorithm [6], but whose convergence is much faster, particularly when the noise variance ϵ is small. To see this, notice that $\hat{\mathbf{A}} - \hat{\mathbf{A}}_\nu = \epsilon(\Delta\mathbf{A} - \Delta\mathbf{A}_\nu)$. Due to the normalization, ϵ will be cancelled. In addition, any multiplicative and additive constants and additive linear terms of $f_i(\cdot)$ will cancel out.

The noise removal in the original FastICA algorithm utilizes only the knowledge about non-Gaussianity of the source signals. Also, the function $f(\cdot)$ is fixed, which corresponds to implicitly assuming that the source signals have distributions of the form

$$p_i(s) = e^{\alpha + \beta s + \gamma s^2 + \zeta F_i(s)}, \tag{4}$$

where $F_i'(s) = f_i(s)$. This is because any multiplicative and additive constants and additive linear terms of $f_i(\cdot)$ will cancel out. The distribution appears to have four degrees of freedom, but only one is actually available: α will be fixed by requiring $p_i(s)$ to integrate to unity and β and γ will typically be fixed because, due the preprocessing of the data and normalization of the mixing matrix $\mathbf{A}$ in the FastICA algorithm, the source signals have unit covariance matrix and zero mean.

The unit covariance matrix of the sources simplifies the terms $\Delta\mathbf{A}$ and $\Delta\mathbf{A}_\nu$. The term $(\hat{\mathbf{S}}^{\mathbf{0}}\hat{\mathbf{S}}^{0T})^{-1}$ reduces into division by the number M of the observation vectors and the term $\Delta\mathbf{A}_\nu$ into $\mathbf{X}_\nu\mathbf{F}(\hat{\mathbf{S}}_\nu^{\mathbf{0}})^T/M = \mathbf{A}\hat{\mathbf{S}}_\nu^{\mathbf{0}}\mathbf{F}(\hat{\mathbf{S}}_\nu^{\mathbf{0}})^T/M = C_f\mathbf{A}$, where C_f is the expectation of $s_\nu f(s_\nu)$ for a normally distributed variable s_ν. Since $f(\cdot)$ is fixed, C_f is constant and needs to be computed only once.

From the present point of view, it is evident that the principle of enhancing the convergence in FastICA is directly applicable to a large number of algorithms which use the alternating estimation principle. The only things needed for deriving new fixed-point BSS algorithms are the function for finding an estimate $\hat{\mathbf{S}}$ of the source signals given all the other values (the observations $\mathbf{X}$, the mixing matrix $\mathbf{A}$ and possibly some other parameters) and the function for estimating the mixing matrix (and possibly the other parameters) given the sources and the observations. The resulting modifications will have much faster convergence rates than the original algorithms, particularly when the noise is low.

It should to be pointed out that although the above derivation was not the original motivation for developing the FastICA [8, 6], the algorithm was interpreted as an approximation for the EM-algorithm in [6]. The interpretation was different from ours, however: the noise was assumed to be finite (not approaching zero) and the estimate of the sources was taken to be the maximum likelihood estimate.

3 Bayesian FastICA

We shall next study a Bayesian extension to the basic FastICA algorithm. The difference to FastICA is that the noise is not assumed to be infinitesimal and therefore the posterior distributions of the sources and the mixing matrix cannot be assumed to have delta distributions. The most important benefit of the Bayesian approach is that models with different structures can be compared with each other. It is usually not known, for instance, how many source signals are present in the data, what are the functional forms of their distributions, what kind of time-dependencies the sources have, etc.

The ability to compare different models relies on measuring different models with a common currency, the probability. It has to be emphasized that probability density is not commensurable since it depends on the particular parametrization of the model. The probability density functions can have arbitrary large but narrow peaks which occupy only little probability mass (integrated probability density). These narrow peaks are, in fact, characteristic of overfitted models.

As the function for estimating the sources we have used the one presented in [1] and the function for estimating the mixing matrix and other parameters is presented in [12], all the necessary ingredients for the fixed-point modification of the Bayesian ICA algorithm already exist and no new formulas need to be derived. Both papers used mixture of Gaussians to model the distribution of the sources and ensemble learning, also known as variational learning, for approximating the posterior distributions. In [1], only the distribution of the sources was fully approximated, i.e., EM-algorithm was used, but the method of estimating the source distributions was more efficient than in [12]. We therefore combine the approaches of the two papers and add the modification for fast convergence.

As was explained in the previous section, the convergence of the algorithm can be accelerated by subtracting from the new estimate $\hat{\mathbf{A}}$ for the mixing matrix the estimate $\hat{\mathbf{A}}_\nu$ which is be obtained when the observations $\mathbf{X}$ are replaced by Gaussian noise having the same covariance as the original observations. More precisely, the posterior means of the mixing matrices are subtracted from each other. The column vectors of the corrected estimate $\hat{\mathbf{A}} - \hat{\mathbf{A}}_\nu$ are then normalized in order to compensate for the subtraction.

The ICA algorithms tend to make the covariance matrix of the sources diagonal and therefore it is possible to make similar simplification as is done in the FastICA where the term $\Delta\hat{\mathbf{A}}_\nu$ reduces into $C_f\mathbf{A}$. Since the estimate for the distribution of the source signals is adapted at each step, the functions f_i keeps changing and it is impossible to precompute the constants C_{f_i} as in the basic FastICA, however. In the simulations the constants were estimated numerically by drawing 10,000 samples from the normal distribution.

The data used in the simulations are the same magnetoencephalographic (MEG) recordings as were used in [17, 16]: two minutes of data was recorded on 122 channels and digitized at 148 Hz. The frequency band of the data is 0.5 – 45 Hz. In addition to signals originating from the electical activity of the brain, there are signals originating from eye movements and blinks, muscle activity, cardiac rhythm, digital watch, etc. As these artifacts are fairly strong and quite independent from

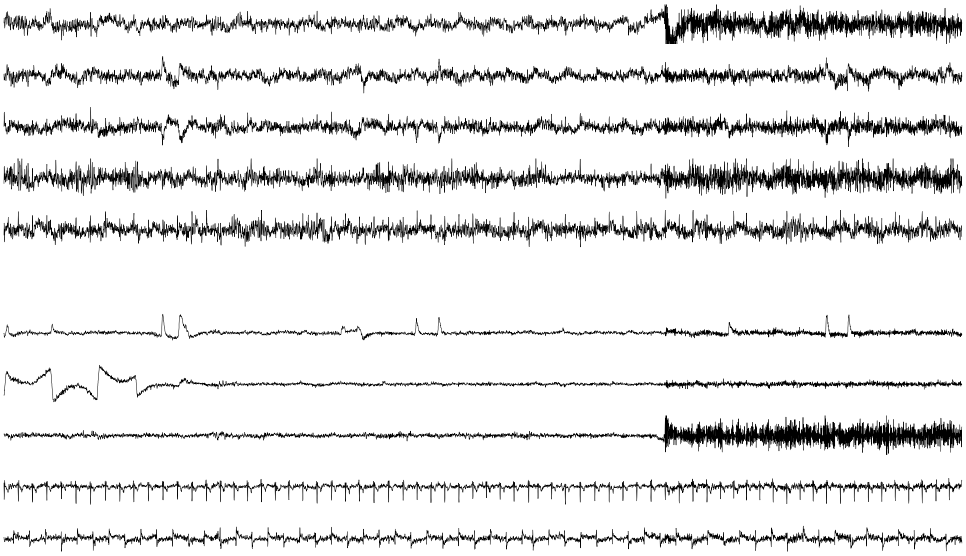

Figure 1: Five channels of the original MEG signals (above) and five of the non-Gaussian sources estimated with the Bayesian FastICA (below). The scales are different for the original signals and the estimated sources. One minute of the recordings is shown.

the brain activity, independent component analysis can be used for identifying and removing them, thus increasing the signal to noise ratio of the brain signals.

Several different model structures were tested and the most probable was the one which had 16 non-Gaussian sources, whose distribution was modeled by a mixture of Gaussians, and 17 Gaussian sources. Figure 1 shows five of the original observed signals and five of the sources whose distribution was estimated to be non-Gaussian. The estimated sources correspond, from top to bottom, to eye blinks, eye movements, muscle activity, cardiac rhythm and digital watch.

The quality of the estimated sources is comparable to the ones reported in [17]. The importance of the method lies in the ability to choose the best model structure, which in this case means choosing the numbers of sources and Gaussians in the mixtures of the source distributions. Also, the results can be directly compared with completely different types of models, such as vector quantization, as long as the Bayesian approach is applied to the other models also. The term corresponding to removing $\mathbf{A}_\nu$ typically reduces the converge time by a factor of ten; the fixed-point iteration will typically have converged in thirty iterations.

4 Sources with time-dependencies

We shall next study a further generalization to the Bayesian ICA which takes into account also the time-dependencies of the source signals in addition to the non-Gaussian distributions. The simplest way to realize such a model is to add linear one-step-forward predictions for the source distributions. This means that instead of assuming that $s_i(t)$ has the distribution $p_i(s_i(t))$ we assume that the difference $d_i(t) = s_i(t) - b_i s_i(t-1)$ has the (non-Gaussian) distribution $p_i(d_i(t))$. Here the b_i are the linear prediction coefficients. If the coefficients b_i were zero, one would obtain the Bayesian ICA of the previous section.

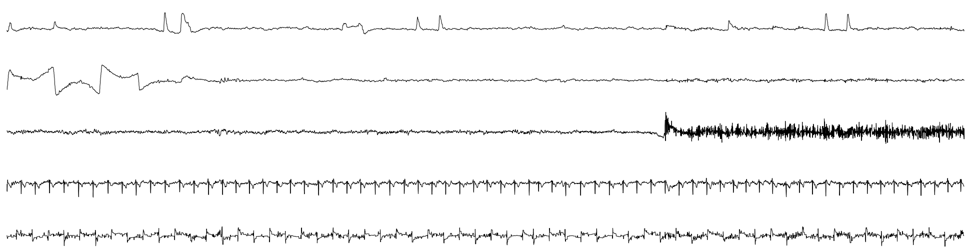

Figure 2: Five non-Gaussian sources were extracted with the Kalman smoothing ICA. The sources correspond to the ones in the previous figure.

The methods presented in the previous section for estimating the mixing matrix and other parameters can be used without any change. The function for estimating the sources needs to be modified to take into account the dependency between successive source signals, however. If the probability distribution $p_i(d_i(t))$ were Gaussian, the posterior distributions of the source signals could be solved analytically, the resulting method being called Kalman smoothing [10, 14].

As the distributions are non-Gaussian, we will have to resort to an iterative approach. When computing the distribution of the source $s_i(t)$, the values $\mathbf{x}(t)$, $\hat{s}_i(t-1)$ and $\hat{s}_i(t+1)$ from the previous iteration are used for computing the new estimate of the posterior distribution of $s_i(t)$. Otherwise the method is the same as in the previous section.

Five out of 24 non-Gaussian source signals of the most probable model are shown in figure 2. There were, in addition, 21 Gaussian sources. The signal to noise ratio for the estimates of the source signals in figure 2 is clearly significantly better than for the sources in figure 1. As the sources clearly have time-dependencies, it is not surprising that the probability of the model far exceeded the best model which assumed that the signals at different time instances are independent.

5 Other generalizations of BSS

The linear one-step-forward prediction which was used for modeling the time-dependencies could easily be extended. One could, for example, apply nonlinear prediction using radial basis function networks as in [5].

An important direction of generalization is the modeling of correlated variances of source signals. In the real world, there are often signals which are activated together, although they appear otherwise independent. In the MEG data, for instance, the signal produced by the muscle activity is more than one-dimensional. This is reflected by the fact that ICA will pick up several source signals whose envelopes are very similar but which are otherwise independent.

A simple fixed-point algorithm for finding subspaces of source signals whose variances are correlated will be presented in [7]. In that paper, the grouping of signals whose variances are correlated was not done adaptively. It should be fairly easy to model the variances of the sources with linear mappings from higher order sources. Using the simple one-step-forward prediction on the higher order sources it would then be easy to take into account also the fact that often the variance of a signal is correlated in time: a signal which is active is more likely to be active in the near future also.

6 Acknowledgements

The author would like to thank Mr. Ricardo Vigário for providing the MEG data and Dr. Aapo Hyvärinen for pointing out many of the references.

References

[1] H. Attias. Independent factor analysis. *Neural Computation*, 11(4):803–851, 1999.

[2] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[3] A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE Trans. on S.P.*, 45(2):434–44, 1997.

[4] P. Comon. Independent component analysis – a new concept? *Signal Processing*, 36:287–314, 1994.

[5] Z. Ghahramani and S. T. Roweis. Learning nonlinear dynamical systems using an EM algorithm. To appear in *Advances in Neural Information Processing Systems 11*, 1999. MIT Press.

[6] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10(3):626–634, 1999.

[7] A. Hyvärinen and P. Hoyer. Independent subspace analysis shows emergence of phase and shift invariant features from natural images. In *Proc. Int. Joint Conf. on Neural Networks*, Washington, D.C., 1999.

[8] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[9] C. Jutten and J. Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[10] R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction. *Journal of Basic Engineering (ASME)*, 83D:95–108, 1961.

[11] J. Karhunen, E. Oja, L. Wang, R. Vigario, and J. Joutsensalo. A class of neural networks for independent component analysis. *IEEE Trans. on Neural Networks*, 8(3):486–504, 1997.

[12] H. Lappalainen. Ensemble learning for independent component analysis. In *Proceedings of the ICA'99*, pages 7–12, Aussois, France, 1999.

[13] L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.*, 72:3634–3636, 1994.

[14] H. E. Rauch. Solutions to the linear smoothing problem. *IEEE Transactions on Automatic Control*, 8:371:372, 1963.

[15] L. Tong, R.-W. Liu, V.C. Soon, and Y.-F. Huang. Indeterminacy and identifiability of blind identification. *IEEE Trans. on Circuits and Systems*, 38, 1991.

[16] R. Vigário. WWW address for the MEG data:
`http://www.cis.hut.fi/~rvigario/NIPS97_data.html`.

[17] R. Vigário, V. Jousmäki, M. Hämäläinen, R. Hari, and E. Oja. Independent component analysis for identification of artifacts in magnetoencephalographic recordings. In *Advances in Neural Information Processing Systems 10 (Proc. NIPS'97)*, pages 229–235, Cambridge, MA, 1998. MIT Press.