Annealed Expectation-Maximization by Entropy Projection

Matthias Seeger
Institute for Adaptive and Neural Computation
University of Edinburgh
5 Forrest Hill, Edinburgh EH1 2QL
seeger@dai.ed.ac.uk

Abstract

We present a new technique of annealing the *EM algorithm* to allow for its tractable application to fitting models which include graph structures like assignments. The method, which can be generally used to sparsify dependence models, is applied to solve the assignment problem for the *shared-resources Gaussian mixture model* (e.g. [4], [5],[9]), and is compared to (and contrasted to) the widely used technique of *deterministic annealing* (e.g. [8],[2]).

1 Motivation

A motivation for this work was to solve an assignment problem in the *shared-resources Gaussian mixture model (SRGMM)* independently proposed in [4] (*semi-tied covariance matrices*) and in [9], and used successfully in Hidden Markov large vocabulary continuous speech recognition systems (e.g. [4], [5],[9]). The SRGMM is a special case of a *basic assignment learning model (BALM)*: Given sets of K *components* $\boldsymbol{C}_k$ and L *resources* $\boldsymbol{A}_l$, each component k chooses one of the resources, say $l(k)$, these assignments are part of the model parameter vector $\boldsymbol{\theta}$, as is a *prior* $\boldsymbol{\pi}$ over $\{1,\ldots,K\}$. When sampling from a BALM, we first choose the component $k \sim \boldsymbol{\pi}$. Then, $\boldsymbol{C}_k$ *combines* with its resource $\boldsymbol{A}_{l(k)}$ to produce a sample point $\bar{\boldsymbol{x}}_v$, this is done independently of all other components and resources. It is easy to show how more complicated (e.g. hierarchical) models can be build from BALM's. In the SRGMM, the resources are *rotation matrices* $\boldsymbol{A}_l$ (or matrices of determinant 1), the components are "incomplete" Gaussians $(\boldsymbol{\mu}_k, \boldsymbol{D}_k)$, where $\boldsymbol{D}_k$ is diagonal and positive, and a combination results in a Gaussian with mean $\boldsymbol{\mu}_k$ and inverse covariance $\boldsymbol{A}_l^T \boldsymbol{D}_k \boldsymbol{A}_l$. The autors cited above motivate this particular decomposition and show that the SRGMM scales very favorably between modeling power and robustness on scarce data. For example, augmenting the large set of Gaussians employed in a state-of-the-art speech recognizer (thousands of mixtures of typically 8 to 32 Gaussians) by a handful of resources $\boldsymbol{A}_l$ led to significant improvements (in test error) over methods like *principal component analysis* or *linear discriminant analysis* which only globally transform the data space ([4], [5],[9])[1]

[1]The SRGMM can be viewed as providing a small set of transformations (rotations) of the data space to *support* components which are only capable of modeling mean and

Once the assignments $k \mapsto l(k)$ are fixed, learning the remaining parameters via EM is fairly straightforward (see [9], [4]), but the problem of how to learn the assignments automatically remained unsolved. Our approach here is to *randomize* the assignments and to learn the conditional probabilities $P(l|k)$ from the observed data using the *EM algorithm* (see section 2). However, since our goal is to estimate a BALM, we have to finally convert these distributions into assignments. We show in the following sections how this problem can be solved in a principled way, by controlling the inherent amount of randomness in the $P(l|k)$ without stepping out of the domain of EM. The approach is very general and not restricted to BALM's.

2 Annealing the EM algorithm

The *EM algorithm* [3] is a powerful statistical tool and central to many modern learning algorithms. It can be motivated very intuitively in geometrical terms (see [1],[6]). Let $\mathcal{X}$ be a set and S a manifold[2] of distributions $P(\boldsymbol{x})$ over $\mathcal{X}$. Let $\boldsymbol{x} = (\boldsymbol{x}_v, \boldsymbol{x}_h)$ where $\boldsymbol{x}_v$ is observable, $\boldsymbol{x}_h$ is hidden. Define the *model submanifold* M as manifold embedded in S and parameterized by $\boldsymbol{\theta}$. The EM algorithm is an iterative procedure to, given a sample $\bar{\boldsymbol{x}}_v$ from $\boldsymbol{x}_v$, find a (local) maximum $\hat{P} = \hat{P}(\hat{\boldsymbol{\theta}}) \in M$ of the *marginal likelihood function* $P \mapsto P(\bar{\boldsymbol{x}}_v)$. EM is equivalent to the following iterative dual minimization procedure (e.g. [1]). The *relative entropy* (or *Kullback-Leibler divergence*)

$$D(Q \,\|\, P) = E_Q \left[\log \frac{Q(\boldsymbol{x})}{P(\boldsymbol{x})} \right] \tag{1}$$

measures the divergence between distributions and can be motivated in the present context from various angles (e.g. [1]). Amari defines the following projections based on D. Given $\tilde{Q}, \tilde{P} \in S$ and $\Gamma \subset S$, the *m-projection* of $\tilde{Q}$ to Γ amounts to finding $\hat{P} \in \Gamma$ which minimizes $D(\tilde{Q}\|\cdot)$. The *e-projection* of $\tilde{P}$ to Γ is $\hat{Q} = \operatorname{argmin}_{Q \in \Gamma} D(Q\|\tilde{P})$. If Γ is convex, both projections are uniquely defined. Define the *data submanifold* to contain all $P(\boldsymbol{x}) \in S$ such that the marginal $P(\boldsymbol{x}_v)$ is equal to the (marginal) empirical distribution $\delta(\boldsymbol{x}_v, \bar{\boldsymbol{x}}_v)$ of $\bar{\boldsymbol{x}}_v$. EM starts from some $(\hat{Q}, \hat{P})$, $\hat{Q} \in A, \hat{P} \in M$, then iterates alternating *E steps* in which $\hat{Q}$ becomes the *e-projection* of $\hat{P}$ to A, and *M steps* in which $\hat{P}$ becomes the *m-projection* of $\hat{Q}$ to M. For example, if $\hat{P} = \hat{P}(\hat{\boldsymbol{\theta}})$ is the current model in the E step, the e-projection of $\hat{P}$ to A results in $\hat{Q} = P(\boldsymbol{x}_h|\bar{\boldsymbol{x}}_v, \hat{\boldsymbol{\theta}})\delta(\boldsymbol{x}_v, \bar{\boldsymbol{x}}_v)$, and then $-D(\hat{Q}\|P(\boldsymbol{\theta})) - H(\hat{Q}) = E_{\hat{Q}}[\log P(\boldsymbol{x}; \boldsymbol{\theta})] = E_{P(\mathbf{x}_h|\bar{\mathbf{x}}_v, \hat{\theta})}[\log P(\boldsymbol{x}_h, \bar{\boldsymbol{x}}_v; \boldsymbol{\theta})]$, which is the EM criterion in its usually presented form.

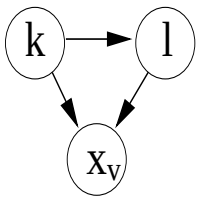
EM often behaves poorly on models with structure variables (e.g. graph structures). Due to intractably difficult search spaces, the projections in the M steps can only be done partially, and the final solution often corresponds to poor local maximum of the log likelihood. A standard technique for such situations is *simulated annealing* [7], and one can easily embed (see e.g. [10]) a very general notion of *annealed EM algorithms* in the EM framework given above, special cases include *deterministic annealing* (see subsection 5.1) as well as the work presented here. The basic idea is to run a *sequence* of EM algorithms on the data, each having its own model and data submanifold. After convergence of one algorithm, we use the solution to initialize the next one. The "art" is to choose the A and M sequences to achieve a somewhat continuous transition between early stages where hardly no local maxima

scaling of data.

[2] We shall not use geometrical properties of S here.

are present and where it is rather easy to explore large regions of the model space in the M steps, to late stages where model and data manifolds are close to the ones we are aiming for. The successive $\hat{P}$ are, if annealing is done carefully, better and better suited as initial distributions to guarantee that the final hard EM run will find a good maximum.

3 M step annealing and entropy projection



Recall BALM's from section 1. *Randomization* of the assignments $l(k)$ amounts to introducing the new hidden variable l, i.e. now $\boldsymbol{x}_h = (k, l)$. The new model submanifold M^* is based on the one for BALM's (say, M), but also includes distributions $P^{(k)} = (P_l^{(k)})_l$ to model the $P(l|k)$. Sampling amounts to drawing $k \sim \boldsymbol{\pi}$, $l \sim P^{(k)}$, then combining $\boldsymbol{C}_k$ and $\boldsymbol{A}_l$ to produce $\bar{\boldsymbol{x}}_v$. Given $\nu_k \in (0, \log L)$, $k = 1, \ldots, K$, we can define a submanifold M_ν, $\boldsymbol{\nu} = (\nu_k)_k$, within M^* by restricting the *entropy* of $P^{(k)}$ to be ν_k, $k = 1, \ldots, K$. If all the ν_k are quite large, the choices of resources by the components are fairly uncertain and the likelihood function over models in M_ν is well-behaved. If all ν_k are close to zero, the distributions $P^{(k)}$ are virtually assignments, and M_ν is therefore almost identical to the model submanifold of the BALM. All in all, we can use a sequence of M_ν model submanifolds with varying ν_k parameters to construct an annealing scheme, this is referred to as *M step annealing*.

Running EM on M^* is standard textbook material (also [9]). The equations for the SRGMM with fixed assignments are given in [9],[4], and the step to M^* requires only straightforward modifications. For the variant on M_ν, only the update of the $P^{(k)}$ distributions in the M step must be altered in a nontrivial way, and these updates can be done independently of each other. It is easy to show that the update of $P^{(k)}$ works as follows: Compute $\hat{Q}_l^{(k)} = \hat{Q}(l|k, \bar{\boldsymbol{x}}_v)$ as posterior from the joint distribution $\hat{Q}$. Then, choose the new $P^{(k)}$ such as to

$$\text{Minimize } D(\hat{Q}^{(k)} \| P), \quad \text{subj. to } H(P) = \nu_k. \tag{2}$$

We refer to this procedure as *entropy projection*.

3.1 Realization of entropy projection

This subsection deals with how to realize entropy projection (2) in practice. For details and proofs we must refer to [10]. Replace (in this subsection) $\hat{Q}^{(k)}$ by Q, ν_k by ν. We distinguish between $H(Q) < \nu$ and $H(Q) > \nu$. In the former case, called *overrelaxation case*, we can replace the constraint by $H(P) \geq \nu$. This strictly convex problem has a unique solution which can easily be found by standard techniques. In the case $H(Q) > \nu$, called *underrelaxation case*, we can replace the constraint by $H(P) \leq \nu$, i.e. the complement of the feasible region is convex[3]. We introduce a *Lagrange multiplier* λ for the entropy constraint and have the *Lagrangian* $L(P, \lambda) = D(Q\|P) + \lambda H(P)$, $\lambda \geq 0$.

One can show that the set of (P, λ) s.t. $\nabla_P L(P, \lambda) = \mathbf{0}$ consists largely of the union of non-intersecting continuously differentiable solution paths $\lambda \mapsto P_\lambda$ (of restricted length) along which $\lambda \mapsto H_\lambda = H(P_\lambda)$ decreases monotonically. Our approach to find a pair with $H_\lambda \approx \nu$ is to solve a sequence of *inner optimizations* for selected

[3]Sloppy speaking, we face the problem of leaving a convex region on the shortest path possible.

values of λ, i.e. to compute P_λ as minimizer of $L(\cdot, \lambda)$, and we try to initialize the optimizer with a $P_{\lambda'}$ computed earlier, where λ' is close to λ. In the outer loop, the successive λ are chosen, in the spirit of a *line search routine*, using extra- and interpolation of the $\lambda \mapsto H_\lambda$ function. In certain very rare cases we encountered *hysteresis effects* (different non-crossing solution paths share common λ regions, see e.g. [11]), due to symmetries in Q which are eventually broken by pushing λ over certain limits. These can be dealt with using a fairly complicated extension of the basic line search routine[4].

4 Experiments

We present preliminary results on a highly prototypical toy example for the SRGMM. The final version of the paper will include results on real-world data. The dataset and its true generative model are shown in figure 4, upper left. The average log likelihood of the data under the true model is -5.9741. In general, initialization can be done by employing simple fast methods like *K-means* (or the *deterministic annealing* version, see subsection 5.1), the $P^{(k)}$ are set to the uniform distribution. The resources are sampled uniformly from all rotations. Rather than employing entropy projection, we can run unrestricted EM (on M^*, see section 3) and hope for the $P^{(k)}$ to become peaked. We compared this method to *monotonic M step annealing*[5]. We did not use entropy projection for *overrelaxation* (see subsection 3.1), but experiments involving this are in preparation. In any case, during the first EM iteration we used no restriction on the m-projection and kept the resources fixed[6]. Initialization of the means with the true ones lead to convergence of the SRGMM to the true model (the solutions had EM crit.[7] close to -6.00) for both algorithms, the resources learned the two different orientations (axis-aligned, diagonal) perfectly, however, convergence of the annealed algorithm required considerably less time. In a second set of tests, we sampled the initial means randomly from the dataset, see figure 4, upper left. Here, unrestricted EM (see figure 4, lower left) exhibits the problem that some components (here: 3,7) try to model too large data regions by using *both* resources. The fit is not very good (EM crit. -6.51), but can be improved by subsequently cooling the solution down using entropy projection (EM. crit. -6.4).

Monotonic M step annealing (see figure 4, upper right) produced a better fit (EM crit. -6.37) and required substantially less iterations for convergence than uncon-

[4] So far we ran more than 50000 entropy projections on randomly drawn (Q, ν) pairs, in about 50 of them (all having high entropy) hysteresis effects disturbed the line search, and in all cases our extended line search found a solution, typically exhibiting an overhead factor of 2 to 3. In general, the problem seems to be rather well-natured: In all these cases, our very aggressive and efficient line search produced the same result as an extremely conservative method pushing λ forward in small constant steps. This is not too surprising since both the criterion and the single constraint are smooth and convex resp. concave.

[5] I.e. all the ν_k are monotonically decreasing in time. Here, we used $\nu_k = 0.4, 0.1, 0.05, 0.01$ for all k, doing 5 EM iterations resp., then ran to convergence with $\nu_k = 0.01$. Note: $\log L = 0.6931$.

[6] This is important since the initial model is completely symmetric w.r.t. components choosing their resources, as long as the diagonal covariance parts $\boldsymbol{D}_k = \boldsymbol{I}$ (spheres are rotationally invariant), and entropy projection starting from uniform distributions $Q^{(k)}$ results in random breaking of the symmetry, i.e. in random assignments. For the same reason, updating the $\boldsymbol{A}_l$ in the first iteration leads to them converging to a single one which destroys the diversity of the model.

[7] The *EM criterion* is $(1/n)(-D(\hat{Q}\|\hat{P}) - H(\hat{Q}))$ (see end of section 2) and a *lower bound* on the avg. log likelihood of the model.

strained EM, but now the two large components 4,7 hinder resource 2 to rotate into a more useful configuration. Since we do not employ *split and merge* techniques in the moment, component 7 is stuck, but note the suboptimal assignment of component 4 to resource 2 (made earlier, when the component mean was situated further below). However, if we start from this solution and let EM run unconstrained for a rather long time (45 iterations), the solution (see figure 4, lower right) is improved (EM crit. -6.25): the assignment of component 4 is changed, resulting in a perfect fit and allowing resource 2 to rotate into a more useful configuration to fit components 2,3. Such correctures of the assignment structure can be enforced more efficiently, using entropy projection for *overrelaxation*, and this suggests *nonmonotonic M step annealing* in which rather cooled down models are heated up again carefully to allow badly fit components to "reconsider" their resource choice.

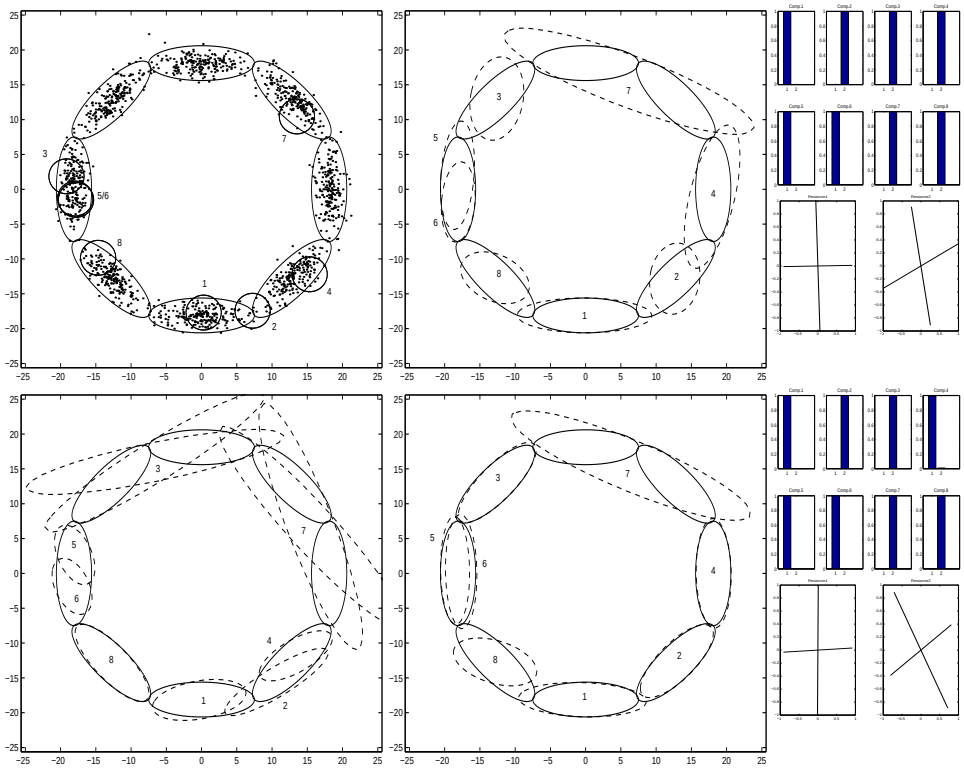


Figure 1: Upper left: Dataset (200 points from each Gaussian), true model and initial SRGMM. Lower left: Unrestricted EM (on M^*). Upper right: Solution produced by "monotonic cooling" (note suboptimal assignment of comp. 4 to res. 2). Lower right: Derived from model "upper right" by running unrestricted EM for many iterations. The assignment of comp. 4 has changed, res. 2 has been rotated to fit comp. 2,3 better.
The small panels belonging to "upper right" and "lower right" show the comp.-to-res. distributions $P^{(k)}$ and the resources $\boldsymbol{A}_l$ (their effect on Euclidean coord. axes).

We have to mention the severe limitedness of these toy experiments. For example, entropy projection is trivial for $L = 2$.[8] Furthermore, the power of SRGMM's is only

[8]As mentioned in subsection 3.1 we ran extensive simulations on the nontrivial case

revealed if $L \ll K$ (see [9],[4]), otherwise the resources can specialize to support a small set of components, resulting in good fits even for bad choices of the assignment. Furthermore, they are triggered to work on spaces of rather large dimension, where robustness caused by data sparseness is an important issue. Experiments in such situations are in preparation.

Nevertheless, some conclusions can be drawn. If the goal is to fit a BALM, careful M step annealing can be much more efficient than running unconstrained EM in hope for sparseness, and with many components, it is very likely that some of them will employ multiple resources in the latter case. Furthermore, it seems that nonmonotonic annealing, i.e. switching between overrelaxation phases to heat the model up, and underrelaxation phases to enforce deterministic assignments, can improve on simple monotonic cooling. Designing automatic schemes for this purpose, possibly in combination with techniques of *E step annealing* (e.g. *deterministic annealing*, see subsection 5.1), is a challenge for future work. When compared to full models from M^* (see section 3), BALM's can be evaluated faster by a factor of L and trained more efficiently (using M step annealing). Learning the assignments in a BALM (as opposed to using a-priori fixed assignments) has its price, in that memory and time requirements[9] increase roughly by a factor of L, but hybrid models could be considered if this is prohibitive.

5 Discussion

We have presented the technique of entropy projection as central building block within a very general notion of annealed EM (e.g. [10]). This notion generalizes, besides the present work on BALM's, a set of established methods (e.g. subsection 5.1), and future work will explore combinations between these. Preliminary toy experiments have been presented in section 4, but experiments on real-world problems are in preparation. The shared-resources Gaussian mixture model (see section 1), together with entropy projection to solve the inherent assignment problem, will be compared to other "sparse" mixture models like mixtures of factor analyzers or of PCA. Several ideas developed for these models, such as *split and merge* techniques, could be applied to SRGMM's. A straightforward extension is to share resources among components of *many* mixture models (see [9],[4]). The SRGMM could be combined in structures like hierarchies, to create a host of new, interesting models. Other types of components or other means of combining components and resources to build covariances can be considered. Finally, entropy projection could be used as a general technique to sparsify and/or learn structures in models fitted by EM (in constrast to heuristic pruning methods), for example Boltzmann machines, (tree-structured) belief networks or transition tables of Hidden Markov models.

There also remain issues of efficient implementation. For example, much time can be saved in the E steps of fitting SGMM's if small components in the $P^{(k)}$ distributions are pushed down to zero, resulting in a more sparse connectivity between components and resources (but note comments on the usefulness of overrelaxation in section 4). This also might alleviate the drawback of rather heavy memory requirements $O(LKd^2)$, where d is the number of dimensions of $\boldsymbol{x}_v$, for SRGMM's to store sufficient statistics (see [10],[9] for details on time and memory requirements).

$L = 10$.

[9] Usually, the computations dominating the time requirements are the evaluations of the model on the training data needed to compute $\hat{Q}$ in the E step.

5.1 Relations to other methods

Relations to earlier work on SRGMM's have been shown in section 1. The most interesting relations of *M step annealing using entropy projection (EP)* can probably be drawn to a very widely used technique for *vector quantization* and related problems, called *deterministic annealing (DA)* (e.g. [8], [2]). Although DA deals with a different problem, namely learning assignments *between datapoints and components* in a mixture model, it builds on the same idea as EP, namely employs an annealed version of EM after a randomization of the assignments. In short, DA anneals EM by placing entropy constraints on the *E step* (e.g. [8], [10]), while EP constrains the *M step* to deal with assignments *between different parts of the model*. From an algorithmic viewpoint, the DA form of *E step annealing* is easier than EP, due to the fact that $\lambda \mapsto P_\lambda$ is unique and can be computed analytically, so as long as we only consider distributions from $\{P_\lambda | \lambda \geq 0\}$, hysteresis effects do not occur. These relations immediately suggest to combine the two methods, for example using the SRGMM for vector quantization will be the subject of future work.

Acknowledgments

We thank Chris Williams and Amos Storkey for helpful discussions. Chris gave very valueable commments on a draft version of the paper. The author gratefully acknowledges support through a research studentship from *Microsoft Research Ltd.*

References

[1] Shun-ichi Amari. Information geometry of the EM and em algorithms for neural networks. *N. Networks*, 8(9):1379–1408, 1995.

[2] J. Buhmann. Stochastic algorithms for exploratory data analysis: Data clustering and data visualization. In M. Jordan, editor, *Learning in Graphical Models*. 1997.

[3] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, 39:1–38, 1977.

[4] Mark Gales. Semi-tied covariance matrices. In *Proceedings of ICASSP 98*, 1998.

[5] R. Gopinath. Maximum-likelihood modeling with Gaussian distributions for classification. In *Proceedings of ICASSP 98*, 1998.

[6] G. E. Hinton and R. M. Neal. A new view on the EM algorithm that justifies incremental and other variants. In M. Jordan, editor, *Learning in Graphical Models*. 1997.

[7] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecci. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[8] K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, 86(11):2210–2239, 1998.

[9] Matthias Seeger. Unterstützung eingeschränkter Mixturmodelle durch lineare Transformationen des Merkmalsraumes im Rahmen des EM-Algorithmus. Technical report, Interactive Systems Laboratory, University of Karlsruhe, 1998. See `www.kyb.tuebingen.mpg.de/bs/people/seeger`.

[10] Matthias Seeger. Annealed expectation-maximization by entropy projection. Technical report, Institute for ANC, Edinburgh, UK, 2000. See `www.kyb.tuebingen.mpg.de/bs/people/seeger`.

[11] J. R. Waldham. *The Theory of Thermodynamics*. Cambridge UP, 1985.