Relationship between gradient and EM steps in latent variable models.

Ruslan Salakhutdinov
Sam Roweis
Department of Computer Science, University of Toronto
rsalakhu,roweis@cs.toronto.edu

Zoubin Ghahramani
Gatsby Neuroscience Unit
University College London
zoubin@gatsby.ucl.ac.uk

Abstract

We present a close relationship between Expectation - Maximization algorithm and direct optimization approaches such as gradient-based methods for parameter learning. We show that the step EM takes in the parameter space and true gradient are related by the *symmetric positive definite* P matrix, and provide an explicit form of this matrix for several widely used latent variable models. We then go on deriving a general form of the P matrix for the regular exponential family in terms of its natural parameters.

1 Introduction

The problem of Maximum Livelihood (ML) learning is known to be an important problem in the area of machine learning and pattern recognition. ML learning is generally hard problem and arises in many probabilistic models with unobserved or latent variables such as density estimation, where one seeks to find a descriptive model of data, or dimensionality reduction, where one tries to discover a compact representation of data.

A variety of methods exist for ML learning of the model parameters in the presence of latent variables. A very popular technique for ML estimation is Expectation-Maximization (EM) algorithm. The EM algorithm alternates between estimating the unobserved variables given the current model and refitting the model given the estimated, complete data. As such it takes discrete steps in parameter space similar to to first order method operating on the gradient of a locally reshaped likelihood function. Direct optimization methods for the parameter learning can be viewed as alternative to the Expectation-Maximization. These algorithms work directly with the likelihood function and its derivatives (or estimates thereof), trying to maximize or minimize it by adjusting the free parameters in a local search. This category of algorithms includes random search, standard gradient-based algorithms, line search methods such as conjugate gradient, and more computationally intensive second-order methods, such as Newton-Raphson.

In this paper we establish mathematical connection between Expectation-Maximization algorithm and direct optimization algorithms. In particular, we show that the step EM takes in the parameter space and true gradient are related by the *symmetric positive definite* matrix $P(\Theta)$, which is a function of the model parameters Θ. For a finite Gaussian mixture model this $P(\Theta)$ matrix was first described by Xu and Jordan[5]. We extend their results by deriving the explicit form of the symmetric positive definite matrix for several widely used latent variable models: Factor Analysis (FA), Probabilistic Principal Component Analysis (PPCA), mixture of FAs, mixture of PPCAs, and Hidden Markov Models (HMM). We then provide a general form of the $P(\Theta)$ matrix for the regular exponential family in terms of its natural parameters.

2 Connection between EM and gradient

2.1 Factor Analysis

Maximum likelihood Factor Analysis (FA) model seeks to specify probabilistically how a d-dimensional observed variable x is related to a p-dimensional latent variable z, where generally $p < d$. This can be viewed as a form of dimensionality reduction. The generative model is give by:

$$x = \Lambda z + \epsilon \tag{1}$$

with Λ being $d \times p$ factor loading matrix, $z \sim \mathcal{N}(0,1)$, and $\epsilon \sim \mathcal{N}(0, \Psi)$, where Ψ is diagonal matrix. In this model, the p-factors represent informative projections of the data, similar to the principal components in PCA.

The log-likelihood function for the FA model with parameters $\{\Lambda, \Psi\}$ is

$$L(\Theta) = -\frac{N}{2}\left(d \ln 2\pi + \ln|C| + \mathrm{tr}(C^{-1}S)\right) \tag{2}$$

where C is the model covariance $C = \Lambda\Lambda^T + \Psi$, and S is a sample covariance matrix $S = \frac{1}{N}\sum_n (x_n - \mu)(x_n - \mu)^T$.

At each iteration of EM algorithm we have

$$\begin{aligned}
\mathrm{vec}[\Lambda^{(t+1)}] - \mathrm{vec}[\Lambda^{(t)}] &= P_\Lambda^{(t)} \frac{\partial L(\Theta)}{\partial \mathrm{vec}[\Lambda]}|_{\Lambda=\Lambda^{(t)}} &(3)\\
\mathrm{vec}[\Psi^{(t+1)}] - \mathrm{vec}[\Psi^{(t)}] &= P_\Psi^{(t)} \frac{\partial L(\Theta)}{\partial \mathrm{vec}[\Psi]}|_{\Psi=\Psi^{(t)}} &(4)
\end{aligned}$$

where vec(A) denotes the stacked columns of A, and

$$\begin{aligned}
P_\Lambda^{(t)} &= \left(\sum_n E^{(t)}(x_n)\right)^{-1} \otimes \Psi^{(t)} \\
P_\Psi^{(t)} &= \frac{2}{N}\mathrm{diag}^*\left[\left(\Lambda^{(t)}(\Lambda^{(t)})^T + \Psi^{(t)}\right) \otimes \Psi^{(t)}\right]
\end{aligned}$$

where $E(x_n) \equiv I - \beta\Lambda + \beta(x_n - \mu)(x_n - \mu)^T\beta^T$ with $\beta \equiv \Lambda^T(\Lambda\Lambda^T + \Psi)$, diag*(A) sets all the rows of A to zero except for rows $j(d+1) - d$, $j = 1, 2, ..., d$, and "$\otimes$" denotes the Kronecker product.

Using the notation $\Theta = \left[\mathrm{vec}[\Lambda]^T, \mathrm{vec}[\Psi]^T\right]^T$, and $P(\Theta) = \mathrm{diag}[P_\Lambda, P_\Psi]$ we can write

$$\Theta^{(t+1)} = \Theta^{(t)} + P(\Theta^t)\frac{\partial L(\Theta)}{\partial \Theta}|_{\Theta=\Theta^{(t)}} \tag{5}$$

Te validity of this symmetric positive definite matrix can be easily verified by multiplying it by the gradient of the log-likelihood function.

Restricting the covariance matrix Ψ to be spherical $\Psi = \sigma^2 I$, we arrive to so-called Probabilistic Principal Component Analysis (PPCA) [3, 4]. Here Λ spans p-dimensional principal subspace of the observed data. The P matrix for PPCA model can be easily derived in the similar way.

2.2 Mixture of Factor Analyzers

Mixture of Factor Analyzers (MFA) can be interpreted as a combination of two basic models: the standard mixture of Gaussians model together with Factor Analysis model.[1] As a result, this hybrid model simultaneously performs two tasks: clustering and local dimensionality reduction within each cluster [1].

The log-likelihood function for MFA model with parameters $\{\pi_i, \mu_i, \Lambda_i, \Psi_i\}_{i=1}^M$ is

$$L(\Theta) = \sum_n \ln \sum_{i=1}^M \pi_i \mathcal{N}(x_n|\mu_i, \Lambda_i\Lambda_i^T + \Psi_i) \tag{6}$$

with M denoting the number of clusters, and π_i, $i = 1, ..., M$ representing the mixing coefficients. At each iteration of EM algorithm we have

$$\Pi^{(t+1)} - \Pi^{(t)} = P_\Pi^{(t)} \frac{\partial L(\Theta)}{\partial \Pi}|_{\Pi=\Pi^{(t)}} \tag{7}$$

$$\mu_i^{(t+1)} - \mu_i^{(t)} = P_{\mu_i}^{(t)} \frac{\partial L(\Theta)}{\partial \mu_i}|_{\mu_i=\mu_i^{(t)}} \tag{8}$$

$$\text{vec}[\Lambda_i^{(t+1)}] - \text{vec}[\Lambda_i^{(t)}] = P_{\Lambda_i}^{(t)} \frac{\partial L(\Theta)}{\partial \text{vec}[\Lambda_i]}|_{\Lambda_i=\Lambda_i^{(t)}} \tag{9}$$

$$\text{vec}[\Psi_i^{(t+1)}] - \text{vec}[\Psi_i^{(t)}] = P_{\Psi_i}^{(t)} \frac{\partial L(\Theta)}{\partial \text{vec}[\Psi_i]}|_{\Psi_i=\Psi_i^{(t)}} \tag{10}$$

where Π denotes mixing coefficients, $\Pi = [\pi_1, ..., \pi_M]^T$ and

$$\begin{aligned}
P_\Pi^{(t)} &= \frac{1}{N}\left[\text{diag}[\pi_1^{(t)}, ..., \pi_M^{(t)}] - \Pi^{(t)}(\Pi^{(t)})^T\right] \\
P_{\mu_i}^{(t)} &= \frac{\Psi_i^{(t)}}{\sum_n h_i^{(t)}(x_n)} \\
P_{\Lambda_i}^{(t)} &= \left(\sum_n h_i^{(t)}(x_n) E_i^{(t)}(x_n)\right)^{-1} \otimes \Psi_i^{(t)} \\
P_{\Psi_i}^{(t)} &= \frac{2}{\sum_n h_i^{(t)}(x_n)} \text{diag}^*\left[(\Lambda_i^{(t)}(\Lambda_i^{(t)})^T + \Psi_i^{(t)}) \otimes \Psi_i^{(t)}\right]
\end{aligned}$$

where $E_i(x_n) \equiv I - \beta_i\Lambda_i + \beta_i(x_n - \mu_i)(x_n - \mu_i)^T\beta_i^T$ with $\beta_i \equiv \Lambda_i^T(\Lambda_i\Lambda_i^T + \Psi_i)$, $h_i(x_n)$ are the responsibilities, diag*(A) sets all the rows of A to zero except for rows $j(d+1) - d$, $j = 1, 2, ..., d$, where d is the dimensionality of data, and "$\otimes$" denotes the Kronecker product.

[1] In regular Mixture of Factor Analyzers model, the the isotropic noise covariance Ψ is fixed across all component densities. In our derivation we have different noise models across different component densities.

Using the notation $\Theta = \left[\Pi^T, \mu_1^T, ..., \mu_M^T, \text{vec}[\Lambda_1]^T, ..., \text{vec}[\Lambda_M]^T, \text{vec}[\Psi_1]^T, ..., \text{vec}[\Psi_M]^T\right]^T$, and $P(\Theta) = \text{diag}[P_\Pi, P_{\mu_1}, ..., P_{\mu_M}, P_{\Lambda_1}, ..., P_{\Lambda_M}, P_{\Psi_1}, ..., P_{\Psi_M}]$ we can write

$$\Theta^{(t+1)} = \Theta^{(t)} + P(\Theta^t)\frac{\partial L(\Theta)}{\partial \Theta}|_{\Theta=\Theta^{(t)}} \tag{11}$$

One can easily verify the validity of this symmetric positive definite matrix by multiplying it by the gradient of the log-likelihood function.

The symmetric positive definite matrix for Mixture of Probabilistic Principal Component Analyzers model [4] can be easily derived in the analogous way.

2.3 Hidden Markov Model

Hidden Markov Model (HMM) can be interpreted as a dynamical mixture model, or a mixture model evolving over time [2].

The log-likelihood of observing the data under this model with parameters $\Theta = \{\pi, A, H\}$ is

$$L(\Theta) = \log \sum_{s_1}\sum_{s_2}...\sum_{s_T}\pi_{s_1}\prod_{t=1}^{T-1}a_{s_t,s_{t+1}}\prod_{t=1}^{T}h_{s_t,x_t} \tag{12}$$

where

- π_i is the probability of state s_i at time t=1.
- A is $M \times M$ matrix with its elements a_{ij} denoting the transition probability from state s_i to state s_j, and M denoting the number of states.
- H is $M \times R$ matrix with its elements h_{ij} denoting the probability of state s_i to generate observation x_j, and R denoting the alphabet size.

At each iteration of EM algorithm we have

$$\begin{aligned}
\Pi^{(t+1)} - \Pi^{(t)} &= P_\Pi^{(t)}\frac{\partial L(\Theta)}{\partial \Pi}|_{\Pi=\Pi^{(t)}} &(13)\\
\text{vec}[A^{(t+1)}] - \text{vec}[A^{(t)}] &= P_A^{(t)}\frac{\partial L(\Theta)}{\partial \text{vec}[A]}|_{A=A^{(t)}} &(14)\\
\text{vec}[H^{(t+1)}] - \text{vec}[H^{(t)}] &= P_H^{(t)}\frac{\partial L(\Theta)}{\partial \text{vec}[H]}|_{H=H^{(t)}} &(15)
\end{aligned}$$

where Π denotes initial state priors, $\Pi = [\pi_1, ..., \pi_M]^T$, and

$$\begin{aligned}
P_\Pi^{(t)} &= \text{diag}[\Pi^{(t)}] - \Pi^{(t)}(\Pi^{(t)})^T\\
P_A^{(t)} &= \text{diag}\big[\text{diag}(\text{vec}(E^{(t)}))\text{vec}(A^{(t)})\big] - \frac{1}{M}\text{vec}(A^{(t)})\text{vec}(A^{(t)})^T\text{diag}(\text{vec}(E^{(t)}))\\
P_H^{(t)} &= \text{diag}\big[\text{diag}(\text{vec}(F^{(t)}))\text{vec}(H^{(t)})\big] - \frac{1}{M}\text{vec}(H^{(t)})\text{vec}(H^{(t)})^T\text{diag}(\text{vec}(F^{(t)}))
\end{aligned}$$

where we have defined the following:

- $E^{(t)}$ is $M \times M$ matrix with elements $e_{ij} = 1/\sum_{t=1}^{T-1}\gamma_t(i)$, where $\gamma_t(i)$ denotes the probability of being in state s_i at time t. (Note that $e_{i1} = e_{i2} = ... = e_{iM}$.)
- $F^{(t)}$ is $M \times R$ matrix with elements $f_{ij} = 1/\sum_{t=1}^{T}\gamma_t(i)$.

Using the notation $\Theta = \left[\Pi^T, \mathrm{vec}[A]^T, \mathrm{vec}[H]^T\right]^T$, and $P(\Theta) = \mathrm{diag}[P_\Pi, P_A, P_H]$ we can write

$$\Theta^{(t+1)} = \Theta^{(t)} + P(\Theta^t)\frac{\partial L(\Theta)}{\partial \Theta}|_{\Theta=\Theta^{(t)}} \tag{16}$$

Once again, the reader can easily verify the validity of this symmetric positive definite matrix by multiplying it by the gradient of the log-likelihood function.

3 Exponential Family Models

Let us assume that the exponential family takes the following form:

$$p(x, z|\Theta) = f(x, z)\exp\{\Theta^T T(x, z)\}/g(\Theta) \tag{17}$$

where x are the observed variables, z are the latent variables, Θ is the vector of natural parameters and T is the vector of sufficient statistics. We are seeking the general form of the transformation matrix $P(\Theta)$, as a function of Θ:

$$\Theta^{t+1} - \Theta^t = P(\Theta^t)\frac{\partial L(\Theta)}{\partial \Theta}|_{\Theta=\Theta^t} \tag{18}$$

where $L(\Theta)$ is the log-likelihood function, and $\Theta^{t+1} - \Theta^t$ represents the step EM performs in the parameter space. We also define the expected complete log-likelihood term as:

$$Q(\Theta|\Theta^t) = \int p(z|x, \Theta^t)\log p(x, z|\Theta)dz \tag{19}$$

For exponential family models we get from (17) that

$$\begin{aligned}\frac{\partial L(\Theta)}{\partial \Theta}|_{\Theta=\Theta^t} &= \frac{\partial Q(\Theta|\Theta^t)}{\partial \Theta}|_{\Theta=\Theta^t} \\ &= \int p(z|x, \Theta^t)T(x, z)dz - \int p(z, x|\Theta^t)T(x, z)dx\,dz \end{aligned}\tag{20}$$

which can be interpreted as the difference in the expected sufficient statistic vector when the observed data is clamped and unclamped. Define the following vector-valued functions:

$$\bar{T}(\Theta) = \int p(z, x|\Theta)T(x, z)dx\,dz \tag{21}$$

$$\bar{T}_z(\Theta) = \int p(z|x, \Theta)T(x, z)dz \tag{22}$$

The M step of the EM algorithm for the exponential family models then solves:

$$\frac{\partial Q(\Theta|\Theta^t)}{\partial \Theta}|_{\Theta^{t+1}} = \bar{T}_z(\Theta^t) - \bar{T}(\Theta^{t+1}) = 0 \tag{23}$$

Since $\bar{T}(\Theta)$ is an invertible function, we can write $\Theta^{t+1} = \bar{T}^{-1}(\bar{T}_z(\Theta^t))$. We now have all the ingredients to re-write (18) as:

$$\bar{T}^{-1}(\bar{T}_z(\Theta^t)) - \Theta^t = P(\Theta^t)\left[\bar{T}_z(\Theta^t) - \bar{T}(\Theta^t)\right] \tag{24}$$

One way, out of many, to write the general form of the transformation matrix $P(\Theta^t)$ that satisfies equation (24) is the following:

$$P(\Theta^t) = \frac{v(\Theta^t)v(\Theta^t)^T}{v(\Theta^t)^T u(\Theta^t)} \tag{25}$$

where $v(\Theta^t) = \bar{T}^{-1}(\bar{T}_z(\Theta^t)) - \Theta^t$, and $u(\Theta^t) = \bar{T}_z(\Theta^t) - \bar{T}(\Theta^t)$. Note that this transformation matrix $P(\Theta^t)$ is symmetric positive definite. Indeed, the denominator of (25) is written as:

$$u(\Theta^t)^T v(\Theta^t) = \frac{\partial Q(\Theta|\Theta^t)}{\partial \Theta}|_{\Theta^t}(\Theta^{t+1} - \Theta) > 0 \quad \Theta^{t+1} \neq \Theta^t \tag{26}$$

The above term can be regarded as a directional derivative of function $Q(\Theta|\Theta^t)$ in the direction of $\Theta^{t+1} - \Theta^t$. This quantity is always positive because the expected complete log-likelihood function $Q(\Theta|\Theta^t)$ for exponential family models is well-defined and concave, attaining its maximum at the point Θ^{t+1}.

4 Discussion

In this paper we have built up the link between EM algorithm and gradient based methods for ML learning by showing that the EM step in the parameter space can be obtained from the gradient via the transformation symmetric positive definite P matrix. The important consequence of the above analysis is that EM has the appealing quality of always taking a step $\Theta^{(t+1)} - \Theta^t$ having positive projection onto the true gradient of the likelihood function $L(\Theta^t)$. This makes EM similar to the first order methods operating on the gradient of a locally reshaped likelihood function.

We could now study the convergence of the EM algorithm by analyzing the structure of this transformation P matrix and relating it to the convergence rate matrix. One could also analyze the effect that P matrix has on the likelihood surface by examining its special properties [5]. This will help us in getting deeper understanding of the nature of the EM algorithm and identify analytic conditions under which it is superior or inferior to other direct optimization methods in terms of convergence.

References

[1] Zoubin Ghahramani and Geoffrey Hinton. The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, Dept. of Computer Science, University of Toronto, May 1996.

[2] Lawrence Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *In Proceedings of IEEE*, 77(2):257–286, February 1989.

[3] S. T. Roweis. EM algorthms for PCA and SPCA. In *Advances in neural information processing systems*, volume 10, pages 626–632, Cambridge, MA, 1998. MIT Press.

[4] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482, 1999.

[5] L. Xu and M. I. Jordan. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, 8(1):129–151, 1996.