On-line Learning of Parametric Mixture Models for Light Transport Simulation – Supplemental material

Jiří Vorba[1*] Ondřej Karlík[1*] Martin Šik[1*] Tobias Ritschel[2†] Jaroslav Křivánek[1‡]

[1]Charles University in Prague [2]MPI Informatik, Saarbrücken

In this document, we derive the model parameter update formulae $\overline{\theta}$ and present details regarding the implementation of our caching scheme.

1 Derivation of Update Formulae

We provide a derivation of the update formulae $\overline{\theta}$ for the model parameters in our stepwise expectation-maximization algorithm that supports weighted particles. The same formulae apply to both the *off-line* and *on-line* versions of stepwise EM.

1.1 MAP and Conjugate Priors

To alleviate over-fitting that is associated with *maximum likelihood* estimation, we pursue a *maximum a posteriori* (MAP) solution. In other words, observing a set of samples $\mathbf{S}$, we seek the mode of the *posterior* distribution $p(\theta|\mathbf{S})$ over mixture model parameters θ, given by the Bayes' theorem: $p(\theta|\mathbf{S}) \propto p(\mathbf{S}|\theta)p(\theta)$ (i.e. posterior $\propto$ likelihood $\times$ prior). To enable this Bayesian treatment, we have to express our prior beliefs about the source of our observed samples via the prior distribution $p(\theta)$. A good choice are *conjugate priors* that take the same functional form as the resulting posterior distribution and therefore lead to a greatly simplified Bayesian analysis [Bishop 2006].

A particular choice of the conjugate prior $p(\theta)$ which expresses our prior beliefs about the covariance matrix $\mathbf{\Sigma}_j$ and the mixing coefficients π_j is:

$$\text{Dir}(\pi_1,\ldots,\pi_K|\nu_1,\ldots,\nu_K)\prod_{j=1}^{K}\text{Wish}(\mathbf{\Sigma}_j|a_j,b_j\mathbf{I}). \qquad (1)$$

This is a product of a conjugate Dirichlet prior on mixing coefficients $\text{Dir}(\pi_1,\ldots,\pi_K|\nu_1,\ldots,\nu_K)$ with hyper-parameters $\nu_j > 0$ and isotropic conjugate Wishart priors $\text{Wish}(\mathbf{\Sigma}_j|a_j,b_j\mathbf{I})$ on the covariance matrix of every mixture component j. Here $\mathbf{I}$ is the identity matrix, K is the number of components in the mixture, $a_j > d-1$, $b_j > 0$ are hyper-parameters, and $d = 2$ is the dimension. Bishop [2006] provides details on the use of Dirichlet and Wishart distributions as conjugate priors.

We base our MAP solution on the prior distribution in the form of Equation (1), that is recommended by Gauvain and Lee [1994] in the context of batch EM. Unlike Gauvain and Lee, we do not take any prior assumptions about the Gaussian means μ_j, because there is no reason to a priori prefer one lobe direction over another. In our final solution, we use the same hyper-parameters a, b and ν for all components so that $\forall j \in \{1,\ldots,K\}$; $a_j = a$, $b_j = b$ and $\nu_j = \nu$. Nevertheless, for the sake of generality, we provide the derivation with possibly different hyper-parameters for each mixture component.

*{jirka,karlik,martin.sik}@cgg.mff.cuni.cz

†ritschel@mpi-inf.mpg.de

‡jaroslav.krivanek@mff.cuni.cz

1.2 Derivation Overview

We provide a derivation of the update formulae for the covariance matrix $\mathbf{\Sigma}_j$ of each Gaussian j in the mixture and for their respective mixing coefficients π_j (Equations (9) and (10) in the paper). The formula for updating the mean μ_j of each Gaussian j (Equation (9) in the paper) is straightforward – we just normalize the weighted sum of observed samples $\mathbf{s}_q$.

The derivation of the update formulae for both $\mathbf{\Sigma}_j$ and π_j follow the same steps. We start from the formulae given by Gauvain and Lee [1994] that describe a MAP update of Gaussian mixture model (GMM) parameters in the batch EM algorithm (see Sec. 3.2 in the paper). Their formulae do not account for weights of observed samples. We generalize these results to stepwise EM while taking sample weights into account. Our generalization proceeds in three steps:

a) We express the parameters of each mixture component j in terms of the batch EM sufficient statistics $\mathbf{u}^j_{N-1}$.

b) We use the fact that stepwise EM is a generalization of batch EM and replace the use of the batch EM statistics $\mathbf{u}^j_{N-1}$ with the stepwise EM statistics $\mathbf{u}^j_i$.

c) Finally, we interpret the weight w_q associated with every observed sample $\mathbf{s}_q$ as its multiplicity.

1.3 Covariance Matrices

Step a. The Gauvain's update formula for the matrix $\mathbf{\Sigma}^{'}_j$ of j-th Gaussian in the mixture reads:

$$\mathbf{\Sigma}^{'}_j = \frac{b_j\mathbf{I} + \sum_{q=0}^{N-1}\gamma_{qj}(\mathbf{s}_q-\mu_j)(\mathbf{s}_q-\mu_j)^T}{(a_j-2)+\sum_{q=0}^{N-1}\gamma_{qj}}, \qquad (2)$$

where $\mathbf{I}$ is the identity matrix, a_j and b_j are hyper-parameters of the Wishart's distribution priors and γ_{qj} is the responsibility of a component j for an observed sample $\mathbf{s}_q$ (see Equation (2) in the paper).

By using simple algebra, we get:

$$(\mathbf{s}_q-\mu_j)(\mathbf{s}_q-\mu_j)^T = \mathbf{s}_q\mathbf{s}_q^T - \mathbf{s}_q(\mu_j)^T - \mu_j\mathbf{s}_q^T + \mu_j(\mu_j)^T. \quad (3)$$

Substituting (3) into (2) and multiplying both the nominator and the denominator by $\frac{1}{N}$, the formula for $\mathbf{\Sigma}^{'}_j$ becomes:

$$\mathbf{\Sigma}^{'}_j = \mathbf{\Sigma}^{'}_j\frac{\frac{1}{N}}{\frac{1}{N}} \qquad (4)$$

$$= \frac{\frac{b_j\mathbf{I}}{N} + \frac{1}{N}\sum_{q=0}^{N-1}\gamma_{qj}\mathbf{s}_q\mathbf{s}_q^T - \mathbf{A} + \mathbf{B}}{\frac{(a_j-2)}{N}+\sum_{q=0}^{N-1}\frac{\gamma_{qj}}{N}}, \qquad (5)$$

where

$$\mathbf{A} = \frac{1}{N}\sum_{q=0}^{N-1}\gamma_{qj}\mathbf{s}_q\mu_j^T + \frac{1}{N}\sum_{q=0}^{N-1}\gamma_{qj}\mu_j\mathbf{s}_q^T$$

and

$$\mathbf{B} = \frac{1}{N}\sum_{q=0}^{N-1} \gamma_{qj}\mu_j\mu_j^T.$$

The *batch* sufficient statistics for N samples (Equation (3) in the paper) reads

$$\mathbf{u}_{N-1}^j = \frac{1}{N}\sum_{q=0}^{N-1} \gamma_{qj}\mathbf{u}(\mathbf{s}_q), \tag{6}$$

where the statistic $\mathbf{u}(\mathbf{s}_q) = (1, \mathbf{s}_q, \mathbf{s}_q\mathbf{s}_q^T)$ is based on an observed sample $\mathbf{s}_q$. By inspection of Equation (5), it is apparent that we have expressed Equation (2) in terms of the batch EM sufficient statistics $\mathbf{u}_{N-1}^j$.

Step b. The sufficient statistics in the *stepwise* EM formulation (see Equation (4) in the paper) reads

$$\mathbf{u}_i^j = (1-\eta_i)\mathbf{u}_{i-1}^j + \eta_i\,\gamma_{qj}\mathbf{u}(\mathbf{s}_q). \tag{7}$$

The sufficient statistics are expressed as a weighted sum with weights $\eta_i = i^{-\alpha}$, where α is the stepsize parameter.

For $\alpha = 1$, the batch EM sufficient statistics, Equation (6), for N samples and the stepwise EM sufficient statistics, Equation (7), for the N-th sample (i.e. $i = N-1$) would be equivalent:

$$\mathbf{u}_{N-1}^j \equiv^{[\alpha=1]} \mathbf{u}_i^j. \tag{8}$$

We use this fact to obtain the MAP update formula of the covariance matrix from Equation (5). If we write the sufficient statistics $\mathbf{u}_i^j$ as a triplet $\mathbf{u}_i^j = ((u_\gamma)_i^j, (\mathbf{s})_i^j, (\mathbf{s}\mathbf{s}^T)_i^j)$ where the first component is the weighted average of all responsibilities γ_{qj}, similarly $(\mathbf{s})_i^j$ is a vector, and $(\mathbf{s}\mathbf{s}^T)_i^j$ is a matrix, the stepwise update formula reads:

$$\mathbf{\Sigma}_j' \quad = \quad \frac{\frac{b_j}{N}\mathbf{I} + (\mathbf{s}\mathbf{s}^T)_i^j - \mathbf{A} + (u_\gamma)_i^j\mathbf{B}}{\frac{a_j-2}{N} + (u_\gamma)_i^j} \tag{9}$$

where

$$\mathbf{A} = (\mathbf{s})_i^j\mu_j^T + \mu_j(\mathbf{s}^T)_i^j, \qquad \mathbf{B} = \mu_j\mu_j^T. \tag{10}$$

Note that if $\alpha < 1$, the equivalence in Equation (8) does not hold. Nonetheless, we take the liberty to generalize the result in Equations (9) and (10) to values of α other than 1. In our implementation we use $\alpha = 0.7$.

Step c. Finally, we interpret the weight w_q associated with every observed sample $\mathbf{s}_q$ as its multiplicity. The stepwise EM sufficient statistics in our weights-aware algorithm are given by Equation (7) in the paper:

$$\mathbf{u}_i^j = (1-\eta_i)\mathbf{u}_{i-1}^j + \eta_i w_q\gamma_{qj}\mathbf{u}(\mathbf{s}_q). \tag{11}$$

To obtain a correct result that takes the weights into account, we normalize these weighted statistics by the total sample weight (Equation (8) in the paper):

$$\overline{w}_i = (1-\eta_i)\overline{w}_{i-1} + \eta_i w_i. \tag{12}$$

The update formula for $\mathbf{\Sigma}_j$, that respects the observed sample weights and provides the MAP solution, becomes:

$$\mathbf{\Sigma}_j = \frac{\frac{b_j}{n}\mathbf{I} + \frac{(\mathbf{s}\mathbf{s}^T)_i^j - \mathbf{A} + (u_\gamma)_i^j\mathbf{B}}{\overline{w}_i}}{\frac{a_j-2}{n} + \frac{(u_\gamma)_i^j}{\overline{w}_i}}, \tag{13}$$

where $\mathbf{A}$ and $\mathbf{B}$ are given in Equation (10) and n is the total number of currently processed samples. (Details on n are given in the last paragraph of Sec. 4.2 in the paper.) Note, that we let the effect of priors diminish with the number of observed samples rather than with the total observed weight $\overline{w}_i$. This helps avoid over-fitting in the early stages of training, when there may only be a few observed samples with potentially enormous weights.

1.4 Mixing Coefficients

To derive the weights-aware update formula for the mixing coefficients π_j, we follow the same procedure as in the above derivation of the covariance matrices $\mathbf{\Sigma}_j$. We start with the update formula given by Gauvain and Lee,

$$\pi_j = \frac{(\nu_j - 1) + \sum_{q=0}^{N-1}\gamma_{qj}}{\sum_{j=1}^{K}(\nu_j-1) + \sum_{j=1}^{K}\sum_{q=0}^{N-1}\gamma_{qj}},$$

and after multiplying both the nominator and the denominator by $\frac{1}{N}$ (step a), using the equivalence (8) (step b), and normalizing for sample weights (step c) we finally arrive at:

$$\pi_j = \frac{\frac{(u_\gamma)_i^j}{\overline{w}_i} + \frac{\nu_j-1}{n}}{1 + \frac{\sum_j^K(\nu_j-1)}{n}}. \tag{14}$$

2 Spacing of Cached Distributions

We now present details on our distribution caching, described in Sec. 4.3 of the paper. Specifically, we detail the computation of the validity radius that determines spacing of cached directional distributions. The *validity radius* r, assigned to each distribution, is a scalar that gives the maximum spatial distance from the distribution where it can be reused. We compute the validity radius as a weighted harmonic mean of validity radii r_j of the individual mixture lobes (i.e. individual GMM components):

$$r = \frac{1}{\sum_j^K \frac{\pi_j}{r_j}}. \tag{15}$$

Here π_j are the mixing coefficients that sum to one over all K components.

When computing the validity radius, we assume that each single lobe of a distribution corresponds to a single highlight and that the lobes are isotropic. These assumptions are made only when computing the validity radius while the training and sampling from the distributions still uses the full anisotropic model.

2.1 Estimating and Limiting Distribution Change

To determine the validity radius r_j of a single lobe l_j, we first predict how the lobe l_j of a distribution at the position $\mathbf{x}$ would change if we observed the corresponding highlight from a slightly different position $\mathbf{x}'$ (see Fig. 1). Because of this change, using a distribution constructed at $\mathbf{x}$ at the different point $\mathbf{x}'$ decreases the importance sampling quality and thus increases variance of the result.

This effect will be small if we ensure that a significant mass of each pair of an original lobe l_j and its predicted image l_j' overlap. We measure this overlap by Kullback-Leiber (KL) divergence [Bishop 2006], a tool for measuring difference between distributions. By imposing a limit on the KL divergence between l_j and l_j' we compute the maximum acceptable angle

$$\alpha_{\max} = \arccos(\omega_\mu \cdot \omega_\mu') \tag{16}$$

between the lobe mean directions ω_μ and ω'_μ. Details about the computation of ω'_μ from KL divergence are given below. Using trigonometry between the original distribution position $\mathbf{x}$, the new position $\mathbf{x}'$ and the alleged highlight position $\mathbf{y}$ (see Fig. 1), the validity radius r_j is then computed as

$$r_j = d_{\text{avg}} \tan(\alpha_{\max}), \tag{17}$$

where d_{avg} is the distance between $\mathbf{x}$ and the position $\mathbf{y}$ of the alleged highlight. We estimate d_{avg} from all particles that were used to train all the lobes in the distribution. Specifically, we take the average of the distance that the particles traveled from their last bounce. We have decided to use the single common estimate d_{avg} for all lobes in a given mixture because it is, according to our experience, a more robust solution than having independent estimates for every lobe.

2.2 KL Divergence Limit

As we said, we determine ω'_μ by imposing a limit on the KL divergence between the lobes l_j and l'_j. This is computed in the unit square domain (see Fig. 1 right) where all the directions on the hemisphere $\mathcal{H}^+$ are projected through the area preserving mapping $\mathcal{S}$ of Shirley and Chiu [1997]. To keep the notation uncluttered, we omit writing the component index j if there is no danger of confusion. In the unit square domain, a lobe l corresponds to a Gaussian $\mathcal{N}(\mathbf{s}|\mu, \mathbf{\Sigma})$ and its shifted image l' to a Gaussian $\mathcal{N}(\mathbf{s}|\mu', \mathbf{\Sigma})$, where $\mu = \mathcal{S}(\omega_\mu)$ and $\mu' = \mathcal{S}(\omega'_\mu)$. Note that these two normal distributions differ only in their means.

The KL divergence formula $\mathrm{KL}(\mu', \mathbf{\Sigma}' \,||\, \mu, \mathbf{\Sigma})$ for two bivariate normal distributions [Duchi 2014] reads

$$\frac{1}{2}\left(\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}') + (\mu-\mu')^T\mathbf{\Sigma}^{-1}(\mu-\mu') - 2 - \ln\frac{|\mathbf{\Sigma}'|}{|\mathbf{\Sigma}^{-1}|}\right),$$

where $|\cdot|$ is the determinant and $\mathrm{tr}(\cdot)$ the trace of a matrix. Since in our case $\mathbf{\Sigma}$ and $\mathbf{\Sigma}'$ are identical, this reduces to one half of a square of Mahalanobis distance Δ between the two means:

$$\mathrm{KL}(\mu', \mathbf{\Sigma}' \,||\, \mu, \mathbf{\Sigma}) = \frac{1}{2}\Delta^2 = \frac{1}{2}(\mu-\mu')^T\mathbf{\Sigma}^{-1}(\mu-\mu').$$

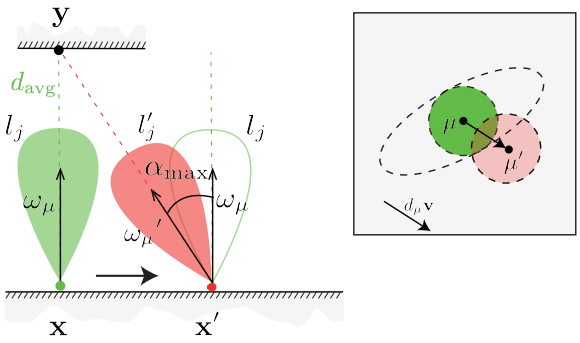

Figure 1: *The assumed geometry used for the computation of the validity radius associated with the lobe of a distribution placed at the position* $\mathbf{x}$*. On the left, the lobe* l_j *with its mean direction* ω_μ *is shifted to a position* $\mathbf{x}'$ *and is compared to its predicted image* l'_j *with its mean direction* ω'_μ*. On the right, the same situation is depicted in the unit square domain. The normal anisotropic distribution with the mean* μ *can be shifted at most about the distance* d_μ *in any direction* $\mathbf{v}$*. The ellipse depicts the real distribution shape while the circle centered in* μ *suggests its conservative isotropic approximation. The distance* d_μ *is determined by the imposed KL divergence limit between the two normal distributions.*

Recall that we assume both normal distributions to be isotropic. This allows us to replace the covariance matrix inverse $\mathbf{\Sigma}^{-1}$ by its eigenvalue λ and thus to write $\Delta^2 = \lambda d_\mu^2$, where d_μ is the distance between the two vectors μ and μ'. It follows that

$$d_\mu = \sqrt{\frac{\Delta^2}{\lambda}}. \tag{18}$$

We impose a maximum threshold $\Delta^2_{\text{thr}} = 5$, which then yields a maximum allowed value of d_μ, denoted $d_{\mu,\text{thr}}$, for a given lobe.

Because our normal distributions are actually anisotropic, we set λ to be the higher from the two eigenvalues of $\mathbf{\Sigma}^{-1}$. This choice is conservative because it causes the resulting lobe validity radius r_j to be smaller than if we chose the other eigenvalue.

Finally, the direction ω'_μ, that appears in Equation (16), is computed as

$$\omega'_\mu = \mathcal{S}^{-1}(\mu'), \tag{19}$$

where $\mu' = d_\mu \mathbf{v}$ and $\mathcal{S}^{-1}$ is the inverse mapping of Shirley and Chiu. The direction $\mathbf{v}$ in the unit square domain can be chosen arbitrarily because we assume isotropic normal distributions (see Fig. 1). So the computed validity radius does not depend on the selected direction $\mathbf{v}$.

2.3 Summary

To summarize, the steps in calculating the validity radius r_j for a single lobe l_j are:

a) Compute the maximum distance $d_{\mu,\text{thr}}$, in which the lobe l_j can be shifted without exceeding the user specified threshold Δ^2_{thr}:

$$d_{\mu,\text{thr}} = \sqrt{\frac{\Delta^2_{\text{thr}}}{\lambda}}. \tag{20}$$

The parameter λ is the larger of the two eigenvalues of $\mathbf{\Sigma}_j^{-1}$, that is the inverse of the lobe covariance matrix $\mathbf{\Sigma}_j$.

b) Select an arbitrary direction $\mathbf{v}$ in the unit square domain and compute the 3D direction ω'_μ using the inverse mapping of Shirley and Chiu:

$$\omega'_\mu = \mathcal{S}^{-1}(d_{\mu,\text{thr}}\mathbf{v}). \tag{21}$$

c) Calculate the validity radius r_j using Equations (16) and (17):

$$r_j = d_{\text{avg}} \tan\left(\arccos\left(\omega_\mu \cdot \omega'_\mu\right)\right). \tag{22}$$

References

BISHOP, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer.

DUCHI, J., 2014. Derivations for linear algebra and optimization. http://www.cs.berkeley.edu/∼jduchi/projects/general_notes.pdf (retrieved in year 2014).

GAUVAIN, J., AND LEE, C.-H. 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans Audio Speech Lang Processing 2*, 2, 291–298.

SHIRLEY, P., AND CHIU, K. 1997. A low distortion map between disk and square. *J. Graph. Tools 2*, 3 (Dec.), 45–52.