Optimal Parametric Density Estimation by Minimizing an Analytic Distance Measure

Anne Hanselmann, Oliver C. Schrempf, and Uwe D. Hanebeck
Intelligent Sensor-Actuator-Systems Laboratory
Institute of Computer Science and Engineering
Universität Karlsruhe (TH), Germany
Email: hanselm@ira.uka.de, {schrempf, uwe.hanebeck}@ieee.org

***Abstract*— In this paper, we present a novel approach to parametric density estimation from given samples. The samples are treated as a parametric density function by means of a Dirac mixture, which allows for applying analytic optimization techniques. The method is based on minimizing a distance measure between the integral of the approximation function and the empirical cumulative distribution function (EDF) of the given samples, where the EDF is represented by the integral of the Dirac mixture. Since this minimization problem cannot be solved directly in general, a progression technique is applied. Increased performance of the approach in comparison to iterative maximum likelihood approaches is shown in simulations.**

Keywords: Density Estimation, Dirac Mixture Densities, Gaussian Mixture Densities, Distance Measure

NOTATION

$\tilde{\tilde{f}}(x)$	true underlying probability density function
$\tilde{f}(x)$	representation of samples drawn from $\tilde{\tilde{f}}(x)$
$f(x)$	approximation of $\tilde{f}(x)$
$\delta(x)$	Dirac Delta function
$H(x)$	Heaviside step function
G	distance measure
$\underline{\eta}$	Dirac mixture parameter vector
$\underline{\kappa}$	parameter vector of arbitrary mixture density
γ	progression parameter
$\mathcal{N}(.,\mu,\sigma)$	Gaussian density with mean μ and standard deviation σ
$\mathrm{erf}(x)$	error function

I. Introduction

The problem of identifying an underlying density from finite sets of observations or measurements is very important in many fields. Traditionally, this problem is considered in statistical analysis, where only samples of a population can be observed and the distribution of the complete population is searched for. The methods applied for that purpose can be summarized as density estimation techniques.

In technical systems, density estimation approaches are very popular in systems identification [1]. Especially the quantification of noise terms by means of probability density functions in probabilistic filters calls for efficient density estimators.

Another application of density estimation lies in model learning for Bayesian networks [2]. Dependencies between nodes in such a network are modeled by means of conditional densities. In the case of nonlinear dependencies between continuous random variables, data driven methods for learning the stochastic models are often inevitable.

Density estimation methods can be divided into two families – parametric approaches and non-parametric approaches.

Non-parametric density estimators in general make no assumptions on the type of the density that produced the samples. The most basic approach in that context are histograms [3], where the observed data is distributed in so called bins and counted, which obviously leads to a discrete representation of the true density. A continuous density representation can be obtained by so called kernel estimators. In that approach, a continuous density function (kernel) is placed on the position where the sample occurred and summed up over all samples. The most prominent representative of this technique is the so called Parzen estimator [4]. The applied kernel is often a Gaussian normal density with a fixed standard deviation called bandwidth.

Parametric methods assume a certain type of density for approximating the true underlying density, which can be any type of parametric density representation like Gaussian, Laplacian, uniform, etc., or mixtures of such densities. This assumption does not mean that the *true* underlying density has to be of this type! The traditional method for estimating the parameters of parametric density functions from a set of samples is the Maximum Likelihood approach dating back to R.A. Fisher in 1912 [5]. The idea behind this approach is to find the parameters of the chosen density type such that the observed data has the highest possible probability. The problem of that approach is that the likelihood function may have several local maxima and finding the global maximum depends strongly on the choice of the initial parameter vector of the method applied. The most prominent algorithm to solve the maximum likelihood problem is the Expectation Maximization (EM) method [6], which is an iterative algorithm known to converge to a maximum of the likelihood function.

In this paper, we will focus on parametric density estimation and present an alternative to the maximum likelihood approach. The method that will be presented here is inspired by previous work we have done on the dual problem. In [7], [8] we have derived a method for approximating continuous probability density functions by means of so called Dirac mixture functions. These Dirac mixtures can be interpreted

as an analytical parametric representation of discrete samples. The method is based on minimization of a distance measure between the continuous function and the Dirac mixture. So the question occurred: "Is it possible to turn around the argument", i.e., is it possible to approximate a set of samples represented as a Dirac mixture by means of a parametric continuous density function. The benefit of such a consideration would be that one can switch from one representation to the other and vice versa. Typical distance measures quantifying the distance between two densities, however, are not well defined for the case of Dirac mixtures. Hence, in this paper the corresponding cumulative distribution functions of the true density and its approximation are compared in order to define an optimal approximation similar to the procedure described in [9].

It is interesting to mention that this approach finally leads to a curve fitting problem in the tradition of [10]. Curve fitting can easily be solved for basic density functions like a single Gaussian, but it is hard for Gaussian mixtures and even harder for arbitrary mixture densities.

The remainder of this paper is structured as follows. In Sec. II, we will give a precise formulation of the problem to be solved. The optimal density approximation scheme will be presented in Sec. III followed by explicit derivations for the special case of Gaussian mixture densities in Sec. IV. Sec. V has some results of the proposed method. A benefit of the connection to the dual problem is shown in Sec. VI by presenting an optimal reapproximation approach. Final conclusions are given in Sec. VII.

II. Problem Formulation

We consider a continuous probability density function $\tilde{\tilde{f}}(x)$. The characteristics of this density are unknown but a set of samples that was generated by this function is available. We assume the samples to be independent and identically distributed (i.i.d.) with respect to $\tilde{\tilde{f}}(x)$.

The goal is to find a parametric density representation $f(x, \underline{\kappa})$ governed by the samples that serves as an approximation of $\tilde{\tilde{f}}(x)$, where $\underline{\kappa}$ is the parameter vector of $f(\cdot)$. For $f(x, \underline{\kappa})$ we will use a finite mixture density representation [11] of the type

$$f(x, \underline{\kappa}) = \sum_{i=1}^{M} w_i f_i(x, \underline{\kappa}_i) \ ,$$

where w_i are weighting factors and $f_i(x, \underline{\kappa}_i)$ are parametric density functions with parameters $\underline{\kappa}_i$. The weights must be positive and have to sum up to one according to

$$\sum_{i=1}^{M} w_i = 1 \ .$$

Hence, we have a parameter vector

$$\underline{\kappa} = \left[w_1, w_2, \ldots, w_M, \underline{\kappa}_1^T, \underline{\kappa}_2^T, \ldots, \underline{\kappa}_M^T\right]^T \ .$$

We interpret the given samples as a density function on a continuous scale, which can be achieved by applying a so called Dirac mixture function

$$\tilde{f}(x, \underline{\tilde{\eta}}) = \sum_{i=1}^{L} \tilde{w}_i \delta(x - \tilde{x}_i) \ ,$$

where $\underline{\tilde{\eta}}$ is a parameter vector consisting of L weights $\tilde{w}_i$ and positions $\tilde{x}_i$ $(i = 1 \ldots L)$ of the samples. Since all samples are of the same importance, we have equal weights $\tilde{w}_i = \frac{1}{L}$, which yields

$$\tilde{f}(x, \underline{\tilde{\eta}}) = \frac{1}{L} \sum_{i=1}^{L} \delta(x - \tilde{x}_i) \ ,$$

with

$$\underline{\tilde{\eta}} = \left[\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_L\right]^T \ .$$

For the remainder of this paper, it is assumed that the component locations are ordered according to

$$\tilde{x}_1 < \tilde{x}_2 < \ldots < \tilde{x}_{L-1} < \tilde{x}_L \ .$$

In general, the true density $\tilde{\tilde{f}}(x)$ is unknown and can therefore not be approximated directly. Since $\tilde{f}(x, \underline{\tilde{\eta}})$ is the only information about the true density we have, we will instead approximate $\tilde{f}(x, \underline{\tilde{\eta}})$ by means of $f(x, \underline{\kappa})$.

Traditionally, this problem is addressed by maximizing the likelihood $f(\underline{\tilde{\eta}}|\underline{\kappa})$ for a fixed $\underline{\tilde{\eta}}$. In contrast to this approach, our key idea is instead to minimize a certain distance measure G between the Dirac mixture $\tilde{f}(x, \underline{\tilde{\eta}})$ and the approximation density $f(x, \underline{\kappa})$. Hence, the estimation problem is considered as an optimization problem. Furthermore, we require that this minimum also minimizes the dual problem given in [7], namely finding Dirac mixture parameters for approximating a continuous density function.

For the reason of brevity we restrict ourselves to the case of densities over scalar random variables.

III. Optimal Density Estimation

Comparing the densities directly does not make much sense for mixtures of Dirac delta functions. Hence, the key idea is to compare the corresponding (cumulative) distribution functions.

The distribution function corresponding to the density $f(x, \underline{\kappa})$ is given by

$$F(x, \underline{\kappa}) = \int_{-\infty}^{x} f(t, \underline{\kappa}) \, dt \ .$$

The distribution function corresponding to the equally weighted Dirac mixture is a staircase function and can be written as

$$\tilde{F}(x, \underline{\tilde{\eta}}) = \frac{1}{L} \sum_{i=1}^{L} H(x - \tilde{x}_i) \ ,$$

where $H(.)$ denotes the Heaviside step function defined as

$$H(y) = \begin{cases} 0, & y < 0 \\ \frac{1}{2}, & y = 0 \\ 1, & y > 0 \end{cases} \ .$$

This mixture is also called empirical distribution function (EDF) [12].

The task is now to find a parameter vector $\underline{\kappa}$ that minimizes a distance measure between the two distribution functions $F(x, \underline{\kappa})$ and $\tilde{F}(x, \underline{\tilde{\eta}})$ for a given $\underline{\tilde{\eta}}$ according to

$$\underline{\hat{\kappa}} = \arg\min_{\underline{\kappa}} G(\underline{\tilde{\eta}}, \underline{\kappa}) \ .$$

A possible distance measure is the integral quadratic distance between the distributions given by

$$G(\underline{\tilde{\eta}}, \underline{\kappa}) = \int_{-\infty}^{\infty} \left(\tilde{F}(x, \underline{\tilde{\eta}}) - F(x, \underline{\kappa}) \right)^2 dx \ . \qquad (1)$$

Since we want to find a minimum with regard to the inverse problem described in [7], we have to consider the partial derivative of (1) with respect to $\underline{\tilde{\eta}}$. By setting the derivative to zero we obtain the following system of (non-linear) equations

$$F(\tilde{x}_i, \underline{\kappa}) = \frac{2i-1}{2L} \ , \quad \text{for } i = 1, \ldots, L \ , \qquad (2)$$

which is similar to results in optimal quantization theory [13]. There is one equation per sample $\tilde{x}_i$. Since the number of samples determined by the dimension of $\underline{\tilde{\eta}}$ is in general much larger than the number of mixture components used for approximation determined by $\underline{\kappa}$, we have an overdetermined system of equations. Furthermore, these equations are nonlinear in general.

Taking a closer look at the graphical representation of the two distribution functions, (2) implies that the continuous distribution $F(x, \underline{\kappa})$ has to meet the staircase function $\tilde{F}(x, \underline{\tilde{\eta}})$ right in the middle of each step.

Solving this system of non-linear equations is basically identical to finding the root of the vector valued function

$$\underline{g}(\underline{\kappa}) = \begin{bmatrix} F(\tilde{x}_1, \underline{\kappa}) - \frac{2 \cdot 1 - 1}{2L} \\ \vdots \\ F(\tilde{x}_L, \underline{\kappa}) - \frac{2 \cdot L - 1}{2L} \end{bmatrix} \ .$$

Using standard approaches like Newton iteration or gradient descent in general yields no satisfying results. This is due to the fact that the convergence of most of these algorithms relies heavily on the choice of the initial parameters.

In order to overcome this problem, we apply a so called progression technique similar to the method applied in [14]. The idea of this approach is to begin with some combination of $\underline{\kappa}_{Start}$ and $\underline{\eta}_{Start}$ known to be optimal, i.e., the distance between $F(x, \underline{\kappa}_{Start})$ and $\tilde{F}(x, \underline{\eta}_{Start})$ is in a global minimum. Note that $\underline{\eta}_{Start}$ is also a configuration of sample points but may be completely different from $\underline{\tilde{\eta}}$. Starting from this configuration, we move the samples gradually towards their desired final positions determined by $\underline{\tilde{\eta}}$ while tracking the minimum via adjusting $\underline{\kappa}$. The transition from $\underline{\eta}_{Start}$ to $\underline{\tilde{\eta}}$ is guided by a so called progression parameter γ, which gradually progresses from 0 to 1. The idea of the progression is to move the samples from initial positions determined by $\underline{\eta}_{Start}$ to the desired positions given by $\underline{\tilde{\eta}}$. This shift of the sample positions guided by γ can be formulated as

$$\underline{\eta}(\gamma) = \gamma \, \underline{\tilde{\eta}} + (1 - \gamma) \, \underline{\eta}_{Start} \ .$$

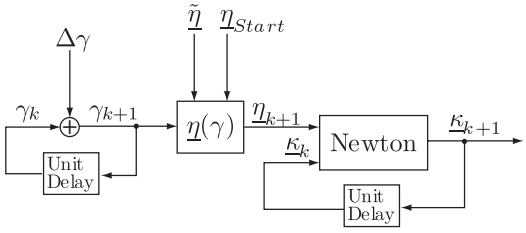

Fig. 1. Progressive solution approach: Progression parameter γ is increased by $\Delta\gamma$, which results in a modified $\underline{\eta}_{k+1}$. $\underline{\kappa}$ is updated via Newton iteration and serves as input to the next update step. We begin with $\gamma_0 = 0$ and the iteration runs while $\gamma < 1$.

While continuously increasing γ from 0 to 1, the minimum of the distance measure can be tracked by adapting the parameter vector $\underline{\kappa}$ by means of e.g. a Newton iteration. In order to calculate the parameter adaptation we increase γ only by small increments. This yields a step-wise iteration, which is depicted in Fig. 1. Each step consists of two phases. First, γ is increased by some step size $\Delta\gamma$ followed by an update of $\underline{\kappa}$ in order to minimize $G(\underline{\eta}(\gamma + \Delta\gamma), \underline{\kappa})$. We assume that the current $\underline{\kappa}$ yields the global minimum of G before γ is increased. Since only infinitesimal small steps $\gamma + \Delta\gamma$ are applied, we do not move far away from the global minimum. Hence, we can apply a Newton iteration in order to update $\underline{\kappa}$. The resulting $\underline{\kappa}$ serves as initial parameter for the next Newton iteration.

The remaining question is how to find initial parameters for the first step ($\gamma = 0$). We have to choose a pair of parameter vectors $\underline{\kappa}_{Start}$ and $\underline{\eta}_{Start}$ that minimizes the distance between $F(x, \underline{\kappa}_{Start})$ and $\tilde{F}(x, \underline{\eta}_{Start})$. For that purpose, we apply the algorithm presented in [7] in order to calculate $\underline{\eta}_{Start}$ for a given $\underline{\kappa}_{Start}$. The best results for the subsequent progression are obtained by choosing $f(x, \underline{\kappa}_{Start})$ as a density function that covers the range of the samples $\tilde{x}_i$ uniformly. Since we use finite mixtures, the uniform distribution in general can only be approximated.

The progressive procedure can be further improved by two extensions to the basic algorithm – step size control and tolerance control.

The step size control adapts the step size $\Delta\gamma$: In the first step, $\Delta\gamma$ is at some minimum step size. If the last Newton iteration converged successfully, the resulting $\underline{\kappa}$ is accepted as starting parameter for the next step and the step size is increased (until some given maximum is reached). In the case that the Newton iteration did not converge successfully, the resulting $\underline{\kappa}$ is discarded, γ is reset to the value of the previous step, and the step size $\Delta\gamma$ is decreased (until some given minimum is reached).

The tolerance control manipulates the convergence criterion of the Newton iteration: The Newton iteration has converged successfully if the parameters or the function value changed less than some ϵ before having done a certain number of iterations ($\epsilon > 0$ but very small). The tolerance control influences this ϵ in the following way: At the first step, we begin with some minimal ϵ. When performing the next steps, tolerance control takes action in two cases: If the Newton

iteration did not converge and the step size $\Delta\gamma$ is already minimal, we discard the resulting $\underline{\kappa}$, reset gamma to the value from the previous step and increase the ϵ (until some given maximum is reached). The other case occurs when the Newton iteration converged and the step size $\Delta\gamma$ is already maximal. Then the resulting $\underline{\kappa}$ is accepted as starting parameter and ϵ is decreased (until some given minimum is reached).

The whole algorithm is also shown in pseudocode notation in Alg. 1.

Algorithm 1 Progressive method with step size control and tolerance control for parameter tracking

$\gamma := 0$
$\underline{\kappa} := \mathrm{Uniform}(M, \mathrm{range}(\underline{\tilde{\eta}}))$
$\underline{\eta}_{Start} := \mathrm{DiracApprox}(\underline{\kappa}, L)$ // Alg. from [7]
$\epsilon := \epsilon_{\mathtt{min}}$
$\Delta\gamma := \gamma_{\mathtt{step\_min}}$
function $\underline{\eta} = \underline{\eta}(\gamma)$
    $\underline{\eta} = \gamma\underline{\tilde{\eta}} + (1-\gamma)\underline{\eta}_{Start}$
end function
repeat
    $\gamma := \gamma + \Delta\gamma$
    $\underline{\eta} = \underline{\eta}(\gamma)$
    $\left[\underline{\kappa}_{\mathrm{tmp}}, \mathit{success}\right] := \mathrm{NewtonApproach}\left(\underline{\kappa}, \underline{\eta}, \epsilon\right)$
    if *success* **then**
        $\underline{\kappa} := \underline{\kappa}_{\mathrm{tmp}}$
        if $\Delta\gamma < \gamma_{\mathtt{step\_max}}$ **then**
            $\mathrm{Increase}(\Delta\gamma, \gamma_{\mathtt{step\_max}})$
        else
            $\mathrm{Decrease}(\epsilon, \epsilon_{min})$
        end if
    else
        $\gamma := \gamma - \Delta\gamma$
        if $\Delta\gamma > \gamma_{\mathtt{step\_min}}$ **then**
            $\mathrm{Decrease}(\Delta\gamma, \gamma_{\mathtt{step\_min}})$
        else
            $\mathrm{Increase}(\epsilon, \epsilon_{max})$
        end if
    end if
until $\gamma = 1$

IV. Special Case: Gaussian Mixtures

In this section, we discuss how to estimate the parameters of $f(x, \underline{\kappa})$ being a Gaussian mixture density

$$f(x, \underline{\kappa}) \quad = \quad \sum_{j=1}^{M} \omega_j \, \mathcal{N}(x, \mu_j, \sigma_j) \ .$$

This type of representation has good approximation capabilities and is very popular in the literature [11].

Here the parameter vector contains the weight ω_j, the mean μ_j, and the standard deviation σ_j of each Gaussian density in the mixture.

In order to obtain a normalized density, the weights are positive and sum up to one according to

$$\sum_{k=1}^{M} \omega_k = 1 \ .$$

Hence, the parameter vector contains only $M-1$ weights. The M-th weight can be calculated from the weights ω_1 to ω_{M-1} as

$$\omega_M = 1 - \sum_{k=1}^{M-1} \omega_k$$

and the parameter vector is

$$\underline{\kappa} = \left[\omega_1, \omega_2, \ldots, \omega_{M-1}, \mu_1, \mu_2, \ldots, \mu_M, \sigma_1, \sigma_2, \ldots, \sigma_M\right]^T \ .$$

The corresponding distribution function of a Gaussian mixture density can be written as

$$F(x, \underline{\kappa}) \quad = \quad \frac{1}{2} \sum_{j=1}^{M} \omega_j \cdot \mathrm{erf}\left(\frac{x - \mu_j}{\sqrt{2}\sigma_j}\right) + \frac{1}{2} \ ,$$

where $\mathrm{erf}(\,\cdot\,)$ denotes the error function.

The resulting system of nonlinear equations with one equation per sample x_i we have to solve in every step is hence

$$\frac{1}{2} \sum_{j=1}^{M} \omega_j \cdot \mathrm{erf}\left(\frac{x_i - \mu_j}{\sqrt{2}\sigma_j}\right) + \frac{1}{2} = \frac{2i-1}{2L}$$

for $i = 1, \ldots, L$. Solving this system of equations is basically the same as finding the root of the vector valued function

$$\underline{g}(\underline{\kappa}) = \begin{bmatrix} \frac{1}{2} \sum_{j=1}^{M} \omega_j \cdot \mathrm{erf}\left(\frac{x_1 - \mu_j}{\sqrt{2}\sigma_j}\right) + \frac{1}{2} - \frac{2 \cdot 1 - 1}{2L} \\ \vdots \\ \frac{1}{2} \sum_{j=1}^{M} \omega_j \cdot \mathrm{erf}\left(\frac{x_L - \mu_j}{\sqrt{2}\sigma_j}\right) + \frac{1}{2} - \frac{2 \cdot L - 1}{2L} \end{bmatrix} \ .$$

Here we apply a modification of Newton's method that uses a pseudoinverse, as the Jacobian of $\underline{g}(\underline{\kappa})$ is not square in general. In each iteration, the $\underline{\kappa}_{k+1}$ is calculated by the rule

$$\underline{\kappa}_{k+1} = \underline{\kappa}_k - \left(\mathbf{J}_{\underline{g}}(\underline{\kappa}_k)^T \mathbf{J}_{\underline{g}}(\underline{\kappa}_k)\right)^{-1} \mathbf{J}_{\underline{g}}(\underline{\kappa}_k)^T \cdot \underline{g}(\underline{\kappa}_k) \ ,$$

where $\mathbf{J}_{\underline{g}}(\underline{\kappa})$ denotes the Jacobian of the vector valued function $\underline{g}(\underline{\kappa})$. In this special case the Jacobian consists of three blocks

$$\mathbf{J}_{\underline{g}}(\underline{\kappa}) = \begin{bmatrix} \mathbf{J}_1 & \mathbf{J}_2 & \mathbf{J}_3 \end{bmatrix} \ .$$

The first of these blocks given by

$$\mathbf{J}_1(i,j) \quad = \quad \frac{1}{2}\left[\mathrm{erf}\left(\frac{x_i - \mu_j}{\sqrt{2}\sigma_j}\right) - \mathrm{erf}\left(\frac{x_i - \mu_M}{\sqrt{2}\sigma_M}\right)\right]$$

is an $(L \times M-1)$-matrix, which contains the derivatives with respect to the weights $\omega_j, \ j = 1, \ldots, M-1$. The second erf-function appears because of the normalization constraint, which says that the weights have to sum up to one. So here ω_M is calculated from $\omega_1, \ldots, \omega_{M-1}$ as stated above, which has to be considered when taking the derivative.

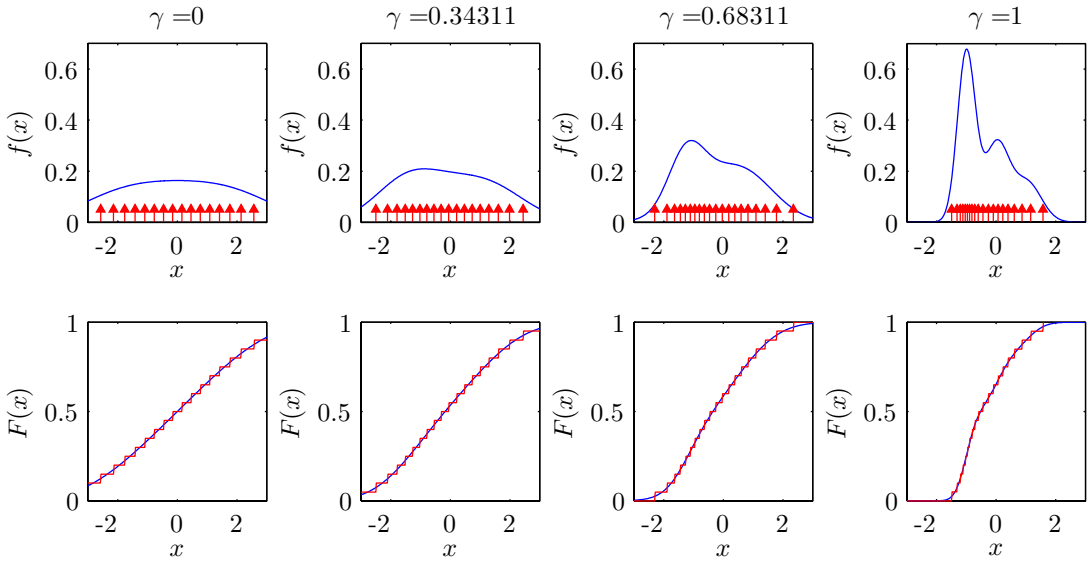

Fig. 2. Example for progression: Estimating a Gaussian mixture density with three components from 20 optimal samples at four different values for γ. Top: Density function. Bottom: Distribution function.

The second of these blocks given by

$$\mathbf{J}_2(i,j) \quad = \quad -\omega_j \mathcal{N}(x_i, \mu_j, \sigma_j)$$

is an $(L \times M)$-matrix. It contains the derivatives with respect to the means $\mu_j,\ j = 1, \ldots, M$.

The third block given by

$$\mathbf{J}_3(i,j) \quad = \quad -\omega_j \frac{x_i - \mu_j}{\sigma_j} \mathcal{N}(x_i, \mu_j, \sigma_j)$$

is an $(L \times M)$-matrix as well. It contains the derivatives with respect to the standard deviations $\sigma_j,\ j = 1, \ldots, M$.

V. Experimental Results

Before we present some results of our approach, there is an important remark to be made: We distinguish between two different types of sample sets. The first type contains purely random samples. They are drawn randomly from a given probability density function. The other type consists of so called optimal samples. They are not drawn randomly but calculated as an optimal approximation of the given probability density function using the algorithm given in [7].

A. Progression

Here we assume the underlying true density $\tilde{\tilde{f}}(x)$ to be a Gaussian mixture with parameter vector

$$\underline{\tilde{\kappa}} = [0.5,\ 0.3,\ -1,\ 0,\ 1,\ 0.3,\ 0.4,\ 0.5]^T \quad ,$$

which implies that there are three not equally weighted components ($\omega_1 = 0.5,\ \omega_2 = 0.3,\ \omega_3 = 1-(\omega_1+\omega_2)$). The means are $\mu_1 = -1,\ \mu_2 = 0$, and $\mu_3 = 1$ and standard deviations $\sigma_1 = 0.3,\ \sigma_2 = 0.4$ and $\sigma_3 = 0.5$.

From this density we calculated 20 optimal samples using the algorithm explained in [7] in order to obtain $\underline{\tilde{\eta}}$.

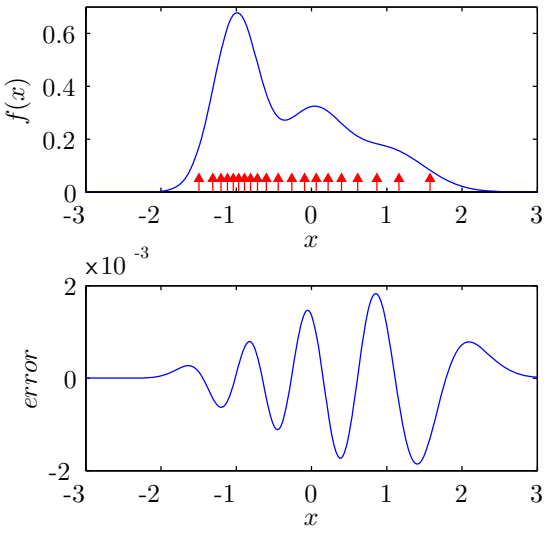

Fig. 3. The resulting density $f(x, \underline{\kappa})$ can be seen in the upper plot. Below the difference $\tilde{\tilde{f}}(x) - f(x, \underline{\kappa})$ is shown.

At $\gamma = 0$ we start with some density function with three components that covers the interval [-3, 3] uniformly, since the samples in $\underline{\tilde{\eta}}$ are located in this range. Its parameter vector $\underline{\kappa}_{Start}$ contains the values

$$\underline{\kappa}_{Start} = \left[\frac{1}{3},\ \frac{1}{3},\ -2,\ 0,\ 2,\ 1.4,\ 1.4,\ 1.4\right]^T \quad .$$

$\underline{\eta}_{Start}$ is computed from $\underline{\kappa}_{Start}$ by the algorithm given in [7] and is hence known to minimize $G(\underline{\eta}_{Start}, \underline{\kappa}_{Start})$.

Four different stages of the progression are shown in Fig. 2 for illustrating this example. The two leftmost plots show the density function (above) and the distribution function (below)

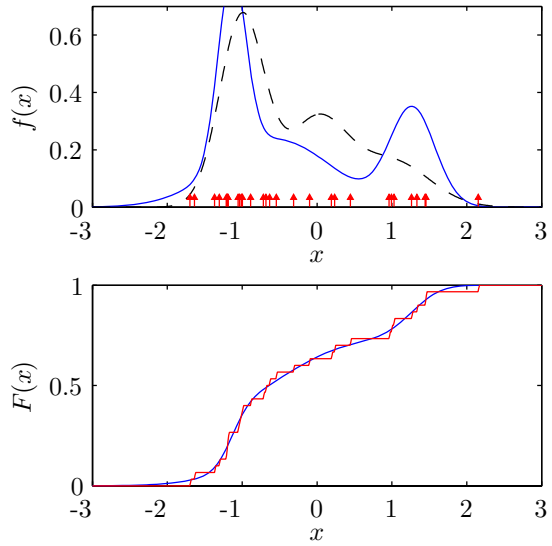

Fig. 4. Estimating a Gaussian mixture density with three components from 30 random samples. Top: True density function $\tilde{\tilde{f}}(x)$ (dashed black line), random samples $\tilde{f}(x, \underline{\tilde{\eta}})$ (red), estimated density function $f(x, \underline{\kappa})$ (solid blue line). Bottom: Staircase function resulting from the samples $\tilde{F}(x, \underline{\tilde{\eta}})$ (red), estimated distribution function $F(x, \underline{\kappa})$ (blue)

at $\gamma = 0$. The two rightmost pictures show the density function (above) and the distribution function (below) at $\gamma = 1$. The two columns in the middle show stages after approximately one third respectively two thirds of the progression.

The resulting density $f(x, \underline{\kappa})$ can be seen in the upper plot in Fig. 3. This is identical to the $\gamma = 1$ stage of Fig. 2. Below, the difference $\tilde{\tilde{f}}(x) - f(x, \underline{\kappa})$ to the true density is shown.

In this first example we used deterministically computed samples instead of random samples in order to explain the basic principle of the progression. An example with random samples is shown in the next section.

B. Random Samples

Here we assume the same underlying true Gaussian mixture density $\tilde{\tilde{f}}(x)$ with parameter vector

$$\underline{\tilde{\tilde{\kappa}}} = [0.5,\ 0.3,\ -1,\ 0,\ 1,\ 0.3,\ 0.4,\ 0.5]$$

as in the section above. But in this example we draw 30 random samples from $\tilde{\tilde{f}}(x)$ in order to obtain $\underline{\tilde{\eta}}$.

We start again with the same parameters as in the section above: some density function with three components that covers the interval $[-3, 3]$ uniformly. From this density we compute 30 optimal samples to obtain $\underline{\eta}_{Start}$. The resulting density and distribution function for $\gamma = 1$ are shown in Fig. 4. In addition, the true density function $\tilde{\tilde{f}}(x)$ is shown as a dashed line in the upper plot. As can be seen from the distribution plot of Fig. 4 the continuous distribution $F(x, \underline{\kappa})$ does not meet the staircase function $\tilde{F}(x, \underline{\tilde{\eta}})$ exactly in the middle of each step. This means the condition (2) is not fulfilled. The reason for this is that the system of nonlinear equations is overdetermined and there exists no exact solution for this system of equations for finite Gaussian mixtures. Nevertheless, the proposed method yields the best possible parameter vector $\underline{\kappa}$ from a least squares perspective.

C. Performance

In this section, we compare the progressive method to the Expectation Maximization method which is the state-of-the-art algorithm in parametric density estimation. In order to obtain comparable results, we draw random samples $\underline{\eta}$ from a Gaussian mixture density function with parameter vector

$$\underline{\tilde{\tilde{\kappa}}} = [0.5,\ 0.3,\ -1,\ 0,\ 1,\ 0.3,\ 0.4,\ 0.5]^T$$

and choose for both algorithms the starting parameters

$$\underline{\kappa}_{Start} = \left[\frac{1}{3},\ \frac{1}{3},\ -2,\ 0,\ 2,\ 1.4,\ 1.4,\ 1.4\right]^T ,$$

which implies a density function that covers the range $[-3, 3]$ uniformly. In order to compare the approaches for different numbers of samples, we draw 20 sets with the same number of random samples, estimate the densities, and measure the error using the integral quadratic distance measure of the resulting densities to the true density $\tilde{\tilde{f}}(x)$ from which the random samples were generated. This distance measure can be written

$$G(\underline{\tilde{\tilde{\kappa}}}, \underline{\kappa}_{EM}) = \int_{-\infty}^{\infty} \left(\tilde{\tilde{f}}(x) - f(x, \underline{\kappa}_{EM})\right)^2 dx \ ,$$

and

$$G(\underline{\tilde{\tilde{\kappa}}}, \underline{\kappa}_{P}) = \int_{-\infty}^{\infty} \left(\tilde{\tilde{f}}(x) - f(x, \underline{\kappa}_{P})\right)^2 dx \ .$$

for EM and the progressive approach respectively.

The root mean square error (e_{RMS}) of the distances generated by both algorithms is shown in the right part of Fig. 5.

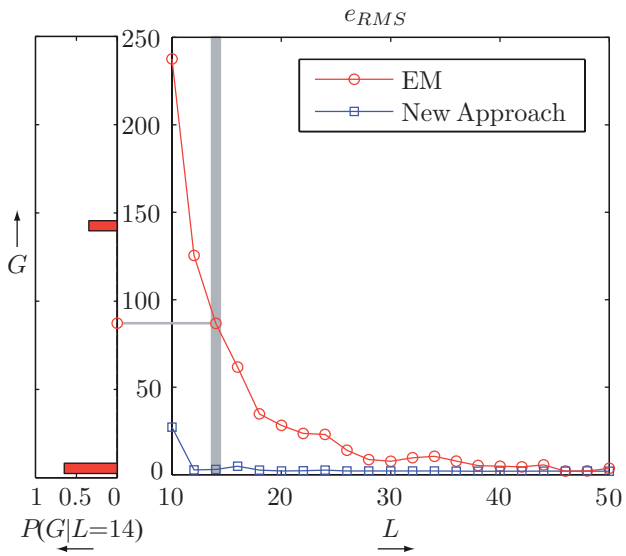

Fig. 5. Left: Histogram of the distribution of $G(\underline{\tilde{\tilde{\kappa}}}, \underline{\kappa}_{EM})$ for $L = 14$ samples. Right: Root Mean Square error of the progressive method (blue) and EM (red) over the distance between the resulting densities and the true density $\tilde{\tilde{f}}(x)$ for 20 runs per L.

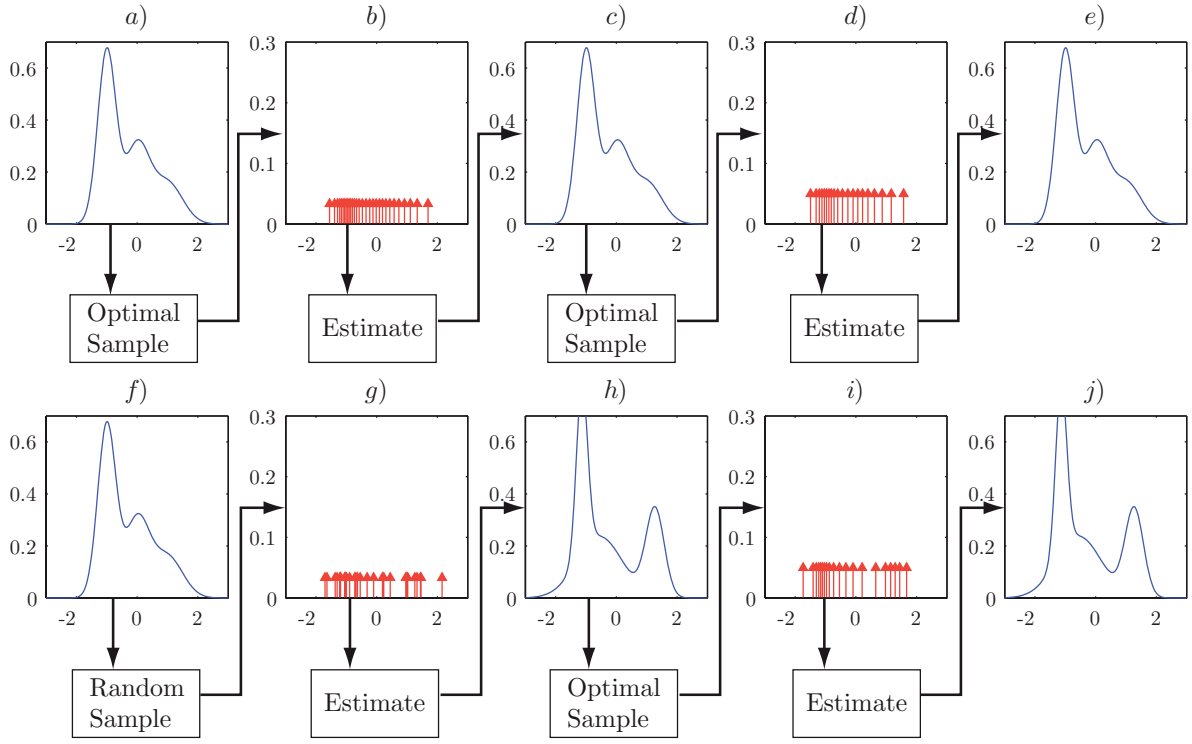

Fig. 6. Top row: a) True density function $\tilde{\tilde{f}}(x)$. b) 30 optimal samples generated from $\tilde{\tilde{f}}(x)$. c) Density estimated from the samples in b). d) 20 from optimal samples computed the density in c). e) Density estimated from the samples in d). Second row: f) True density function $\tilde{\tilde{f}}(x)$, the same as in a). g) 30 random samples drawn from $\tilde{\tilde{f}}(x)$. h) Density estimated from the samples in g). i) 20 optimal samples computed from the density function in h). j) density function estimated from the samples in i)

It can be seen that the progressive method performs much better than EM – especially for a small number of samples. The reason for this is motivated by the left plot in Fig. 5. Here we show a histogram of the distribution of $G(\underline{\tilde{\tilde{\kappa}}}, \underline{\kappa}_{EM})$ for $L = 14$ samples. It shows that for some sample configurations the EM algorithm converges to the global maximum, but for a large number of configurations it converges to a local maximum.

VI. Reapproximation

We will now present a reapproximation scheme as an application of the combination of the approach presented here and the dual one presented in [7]. Reapproximation by optimal samples can for example be applied instead of random resampling in sample-based filtering techniques like particle filters, where sample degradation is a problem.

The idea of the reapproximation scheme we present here is simple and straightforward: Given a set of samples, a continuous probability density function can be estimated by the approach presented in this paper. From the resulting density an arbitrary number of optimally placed samples with respect to this density can be generated by the dual algorithm.

Since [7] is dual to the new progressive method introduced in this paper, one can switch back and forth between the continuous density function and its discrete representation by optimal samples without causing a major loss. This means some initial density function $\tilde{\tilde{f}}(x)$ can be approximated optimally by means of Dirac mixtures, then be reestimated by means of Gaussian mixtures, and again be approximated by means of Dirac mixtures – even with a different number of samples – and reestimated again, and the error towards the original function $\tilde{\tilde{f}}(x)$ increases just marginally.

A. Example

In Fig. 6 an illustrative example is presented. The first row shows the completely deterministic case, in the second row the case where random samples are involved is displayed.

In plot a), the true continuous probability density function $\tilde{\tilde{f}}(x)$ given by a Gaussian mixture with parameters

$$\underline{\tilde{\tilde{\kappa}}} = [0.5,\ 0.3,\ -1,\ 0,\ 1,\ 0.3,\ 0.4,\ 0.5]^T \quad .$$

is depicted. Plot b) shows 30 optimal samples generated from $\tilde{\tilde{f}}(x)$ using the algorithm from [7]. From these samples a density is estimated using the approach presented in this paper. The resulting density is shown in in plot c). This density is again represented by 20 optimal samples displayed in plot d) followed by an estimation shown in plot e).

The error with respect to the true density function $\tilde{\tilde{f}}(x)$ that arises from this reapproximation is charted in the first row of Fig. 7. The error is quantified by the difference between the

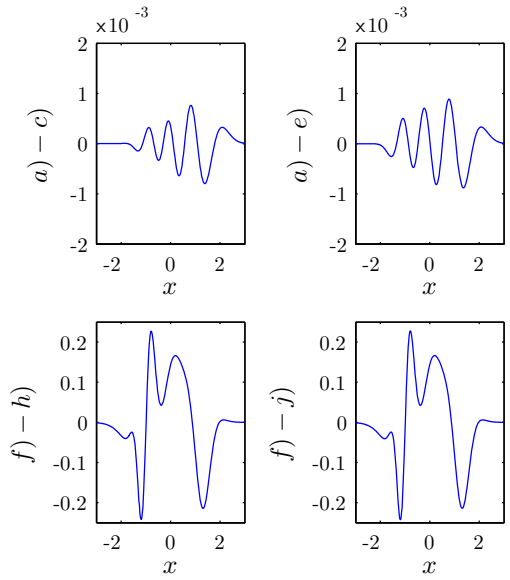

Fig. 7. Top row, left plot: Difference between the density of Fig. 6.a) and the density of Fig. 6.c). Top row, right plot: Difference between the density of Fig. 6.a) and the density of Fig. 6.e). Second row, left plot: Difference between the density of Fig. 6.f) and the density of Fig. 6.h). Second row, right plot: Difference between the density of Fig. 6.f) and the density of Fig. 6.j).

density of plot a) and the densities in c) and e) respectively. It can be seen that the error is very small and increases only marginally.

In the second row basically the same actions are taken beginning in plot f) with the same density as in plot a), but now there are 30 *random* samples drawn. The result is shown in g). From these samples the density function depicted in plot h) is estimated by application of our approach. The samples shown in i) are optimal samples computed from the density function in h). The estimated density based on these samples is shown in j). The arising error with respect to the true density function $\tilde{\tilde{f}}(x)$ is given in the second row of Fig. 7. The left plot shows the error after the random sampling-and-estimation step. This error is determined by the quality of the random samples. The right plot gives the error after the subsequent optimal sampling-and-estimation step. It is to be recognized that the error did not increase.

VII. Conclusions

We have presented an approach for parametric density estimation by minimizing the distance between the distribution functions of a set of samples represented by means of a Dirac mixture density and a parametric continuous mixture density. The presented approach is strongly tied to the dual problem of approximating a continuous density by means of a Dirac mixture density. Considering the necessary condition of the dual problem, we have shown that the minimization of the distance measure boils down to a curve fitting problem.

The new approach leads to higher computational effort than the popular EM method. However, the experiments show that the proposed method yields significantly better results for a small number of random samples. The reason for this lies in the fact that the EM method tends to converge to local maxima. Furthermore, the higher computational effort is a minor issue, as the new approach is designed to be used offline.

For the general case, where random samples are given, our approach yields the best possible solution from a least squares perspective due to the applied distance measure. Since the problem is overdetermined, an exact solution in general is not available.

The simulations have also shown that our approach yields very good results for samples generated by the algorithm in [7]. This observation emphasizes that we indeed solve the dual problem, which allows us to switch back and forth between a Dirac mixture and a Gaussian mixture representation

The coexistence of the two approaches given here and in [7] can be used to implement a reapproximation scheme which yields a desired number of optimally placed samples from a given number of purely random samples as shown in Sec. VI.

Due to the fact that in many applications samples appear sequentially, a modified progression scheme allowing for successive insertion of samples would be beneficial, resulting in a sequential density estimator.

References

[1] V. Peterka, "Bayesian System Identification," *Automatica*, vol. 17, no. 1, pp. 41–53, Jan. 1981.

[2] M. Ramoni and P. Sebastiani, "Parameter Estimation in Bayesian Networks from Incomplete Databases." *Intell. Data Anal.*, vol. 2, no. 1-4, pp. 139–160, 1998.

[3] D. Freedman and P. Diaconis, "On the Histogram as a Density Estimator: L2 Theory," *Probability Theory and Related Fields*, vol. V57, no. 4, pp. 453–476, Dec. 1981.

[4] E. Parzen, "On Estimation of a Probability Density Function and Mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, Sep. 1962.

[5] J. Aldrich, "R. A. Fisher and the Making of Maximum Likelihood 1912-1922," *Statistical Science*, vol. 12, no. 3, pp. 162–176, aug 1997.

[6] Dempster, A. P., Laird, N. M., and Rubin, D. B., "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[7] O. C. Schrempf, D. Brunn, and U. D. Hanebeck, "Density Approximation Based on Dirac Mixtures with Regard to Nonlinear Estimation and Filtering," in *Proceedings of the 45th IEEE Conference on Decision and Control (CDC'06), San Diego, California, USA*, Dec. 2006.

[8] ——, "Dirac Mixture Density Approximation Based on Minimization of the Weighted Cramér–von Mises Distance," in *Proceedings of the International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2006), Heidelberg, Germany*, Sep. 2006, pp. 512–517.

[9] M. D. Weber, L. M. Leemis, and R. K. Kincaid, "Minimum Kolmogorov-Smirnov Test Statistic Parameter Estimates," *Journal of Statistical Computation and Simulation*, vol. 76, no. 3, pp. 196–206, 2006.

[10] K. Pearson, "On the Systematic Fitting of Curves to Observations and Measurements," *Biometrika*, vol. 1, no. 3, pp. 265–303, Apr. 1902.

[11] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*, ser. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1985.

[12] R. B. D'Agostino and M. A. Stephens, Eds., *Goodness-of-Fit Techniques*, ser. Statistics, textbooks and monographs. Marcel Dekker, Inc., 1986, vol. 68.

[13] S. P. Lloyd, "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.

[14] U. D. Hanebeck, K. Briechle, and A. Rauh, "Progressive Bayes: A New Framework for Nonlinear State Estimation," in *Proceedings of SPIE*, vol. 5099, Orlando, Florida, 2003, pp. 256–267, AeroSense Symposium.