Uwe D. Hanebeck*

Optimal Reduction of Multivariate Dirac Mixture Densities

Optimale Reduktion von multivariaten Dirac-Mischdichten

Abstract: This paper is concerned with the optimal approximation of a given multivariate Dirac mixture, i.e., a density comprising weighted Dirac distributions on a continuous domain, by a Dirac mixture with a reduced number of components. The parameters of the approximating density are calculated by numerically minimizing a smooth distance measure, a generalization of the well-known Cramér–von Mises-Distance to the multivariate case. This generalization is achieved by defining an alternative to the classical cumulative distribution, the Localized Cumulative Distribution (LCD), as a smooth characterization of discrete random quantities (on continuous domains). The resulting approximation method provides the basis for various efficient nonlinear estimation and control methods.

Keywords: Dirac mixture, mixture reduction, distance measure.

Zusammenfassung: Dieser Beitrag befasst sich mit der optimalen Approximation einer multivarianten Dirac-Mischdichte durch eine Dirac-Mischdichte mit einer geringeren Anzahl an Komponenten. Dirac-Mischdichten bestehen aus gewichteten Dirac-Distributionen auf einer kontinuierlichen Domäne. Die Parameter der approximierenden Dichte werden durch numerische Minimierung eines glatten Abstandsmaßes gewonnnen, welches eine Verallgemeinerung der bekannten Cramér–von Mises-Distanz darstellt. Diese Verallgemeinerung wird durch die Einführung einer Alternative zu klassischen kumulativen Verteilungen, den so genannten lokalisierten kumulativen Verteilungen, als eine glatte Charakterisierung von diskreten Zufallsgrößen (auf kontinuierlichen Domänen) erreicht. Die resultierende Approximationsmethode bildet die Grundlage für verschiedene effiziente nichtlineare Schätz- und Regelungsverfahren.

Schlüsselwörter: Dirac-Mischdichte, Reduktion von Mischdichten, Distanzmaß.

***Corresponding author: Uwe D. Hanebeck,** Intelligent Sensor-Actuator-Systems Laboratory (ISAS), Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology (KIT), Karlsruhe, e-mail: uwe.hanebeck@ieee.org

1 Introduction

1.1 Motivation

We consider sample sets $\mathcal{S}_x = \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_L\}$ where the locations $\underline{x}_i$, $i = 1, \ldots, L$ are arbitrarily placed in $\mathbb{R}^N$. The samples can be of arbitrary origin, e.g., be random samples from a probability density function, and are equipped with probabilities w_i^x, $i = 1, \ldots, L$ that are positive and sum up to one. These probabilities are not necessarily equal.

The sample sets are interpreted as discrete probability density functions over a continuous domain, where the individual samples correspond to locations of Dirac distributions with associated weights. The sample sets will be called Dirac mixture densities. Dirac mixture densities are a popular representation of densities in stochastic nonlinear filters such as particle filters [1]. They characterize random vectors by having a large number of components (or large weights) in regions of high density and a small number of components (or small weights) in regions of low density. Hence, approximating one Dirac mixture density by another one while maintaining the information content is equivalent to maintaining its probability mass distribution.

1.2 Applications

In model-based estimation and control methods, a model of the considered system is required. In case of stochastic

dynamic systems, a deterministic state-space model could be employed together with models for the noise sources. Alternatively, the entire system could be described by a probabilistic model, i.e., by its transition density. In addition, during processing, the uncertain state has to be represented by some type of probability density function. In many cases, only samples of the probability density functions used for describing input and output signals and samples of the probability density function for describing the considered system itself are given, so it is natural to describe all signals and systems by discrete densities. As the number of samples of these discrete densities is often too large for real-time processing, reducing the number of samples of a discrete density while maintaining its information content as much as possible is a fundamental problem.

When given noisy samples of the unknown probability density functions characterizing the noise sources of the considered system, a more compact and efficient representation is usually desired for further processing. One option is to perform density estimation with the goal of calculating an appropriate continuous probability density function explaining the samples. This requires assumptions about the type or the smoothness of the underlying density and is computationally expensive. As a cheap and simple alternative, the proposed reduction method can be employed to represent the original samples by a much smaller set of well-placed samples. These samples can either be directly used in the estimation and control methods or they can serve as input for density estimation methods that are then much cheaper to perform.

The proposed reduction method can also be used for identifying a model of an entire dynamic system including endogenous and exogenous noise sources based on input/output samples or samples of consecutive states. These samples are represented by a smaller set of well-placed samples as a discrete approximation of the transition density describing the system.

Once the model is available, estimation and control can be performed. In estimation, the state estimate can be represented by samples as, e.g., in the particle filter. A typical operation, the addition of two random variables, corresponds to the convolution of the corresponding densities. For discrete densities, this convolution requires the calculation of the Cartesian product of the two sample sets. The number of resulting samples is the product of the numbers of the two original sample sets, so that the number of samples explodes after a few steps. Usually, random selection methods are used to keep the number of samples at a constant level. The proposed reduction method can be employed to systematically and adaptively reduce the number of samples to facilitate further processing. The reduction method derived in this paper requires the samples to be explicitly given, so that the Cartesian product has to be explicitly performed, which sometimes is not practical even for a single processing step. A variant of the reduction method has been proposed in [2], where the Cartesian product of two sample sets is implicitly approximated. This method avoids the explicit calculation of all combinations of samples from the two sample sets.

Open-loop model predictive control requires the repeated propagation of the state estimate through the system model over a certain prediction horizon at every time step. For nonlinear systems, this includes a nonlinear mapping of the state density and a generalized convolution with noise densities. As these operations can only rarely performed analytically for arbitrary densities, the original densities are often approximated by discrete densities that can easily be propagated through the system. However, repeated propagation again leads to an exponential increase in samples and requires some sort of reduction. In [3], an early version of the proposed reduction method was successfully applied to keeping the number of samples in open-loop model predictive control at a manageable level by performing regular reductions.

1.3 Related work

We will now take a look at different methods that have been devised in the general context of reduction of point sets or discrete densities.

Random selection

The most common technique for reducing the number of components of a given Dirac mixture density is the random selection of certain components. It is commonly used in the prediction step of particle filters, where each prior sample is perturbed with a single sample from the noise distribution before propagation through the system model [4]. The perturbation can be viewed as generating the noise samples at once with a subsequent random selection from the Cartesian product of prior samples and noise samples.

Intermediate continuous densities

Another common technique is to replace the given Dirac mixture by a suitable continuous density in a first step [5]. In a second step, the desired number of samples is drawn from the continuous density. With this technique, it is also possible to increase the number of components as re-

quired. However, the first step is equivalent to density estimation from samples, which is by itself a complicated task and an active research topic. Furthermore, this reduction technique introduces undesired side information via the choice of the continuous smoothing density.

Clustering or vector quantization methods

Clustering or vector quantization methods also aim at representing a point set by a smaller set of representatives. For optimization purposes, a distortion measure is typically used, which sums up the (generalized) distances between the points and their representatives. Minimizing the distortion measure results in two conditions: i) Points are associated to their closest representative. ii) The representative is calculated as the average of all its associated points. As no closed-form solution for performing the minimization of the distortion measure exist, robust iterative procedures have been devised starting with Lloyd's algorithm proposed in 1957 and published later in [6], first called k-means algorithm in [7], and its extension in the form of the Linde-Buzo-Gray-algorithm [8]. Obviously, the representatives fulfilling the above two conditions do not necessarily maintain the form of the density, which will also be shown by some examples in Sect. 6 of this paper. An additional problem of clustering or vector quantization methods is that the iterative minimization procedures typically get stuck in local minima. Intuitively, the resulting samples are only influenced by samples in the corresponding part of the Voronoi diagram, while the proposed method is based upon a global distance measure.

Reapproximating continuous mixtures with continuous mixtures

Dirac mixture reduction is a special case of general mixture reduction techniques. As these techniques are usually focused on continuous densities such as Gaussian mixtures, e.g., see [9], it is worthwhile to discuss the differences. First, when continuous mixtures are reapproximated with continuous mixtures, the densities or parts of the densities can be directly compared in terms of the integral squared difference [10] or the Kullback-Leibler divergence [11]. Directly comparing densities with an integral measure is not possible when at least one of the densities is a Dirac mixture density [12]. Instead, cumulative distributions can be used in the scalar case or appropriate generalizations for the multivariate case [12]. Second, for continuous mixtures two or more critical components can be merged in order to locally reduce the number of components [13], where different criteria for identifying components are possible such as small weights. These components are then replaced by a new component with appropriate parameters, e.g., maintaining mean and covariance. Locally replacing components is not straightforward for Dirac mixture densitys as it is i) difficult to identify potential merging candidates and ii) a single replacement component does not capture the extent covered by the original components. Hence, a replacement of several Dirac components by a smaller set of Dirac components with a cardinality larger than one would be in order.

Reapproximating continuous mixtures with discrete mixtures

The reduction problem can be viewed as approximating a given (potentially continuous) density with a Dirac mixture density. Several options are available for performing this approximation. Moment-based approximations have been proposed in the context of Gaussian densities and Linear Regression Kalman Filters (LRKFs), see [14]. Examples are the Unscented Kalman Filter (UKF) in [15] and its scaled version in [16], its higher-order generalization in [17], and a generalization to an arbitrary number of deterministic samples placed along the coordinate axes introduced in [18]. For circular probability density functions, a first approach to Dirac mixture approximation in the vein of the UKF is introduced in [19] for the von Mises distribution and the wrapped Normal distribution. Three components are systematically placed based on matching the first circular moment. This Dirac mixture approximation of continuous circular probability density functions has already been applied to sensor scheduling based on bearings-only measurements [20]. In [21], the results are used to perform recursive circular filtering for tracking an object constrained to an arbitrary one-dimensional manifold. For the case that only a finite set of moments of a random vector is given and the underlying density is unknown, an algorithm is proposed in [22] for calculating multivariate Dirac mixture densities with an arbitrary number of arbitrarily placed components maintaining these moments while providing a homogeneous coverage of the state space. This method could also be used for the reduction problem by calculating the moments of the given point set. Methods that are based on distance measures between the given density and its Dirac mixture approximation have been proposed for the case of scalar continuous densities in [23, 24]. They are based on distance measures between cumulative distribution functions. These distance-based approximation methods are generalized to the multi-dimensional case by defining an alternative to the classical cumula-

tive distribution, the Localized Cumulative Distribution (LCD) [12], which is unique and symmetric. Based on the LCD, multi-dimensional Gaussian densities are approximated by Dirac mixture densities in [25]. A more efficient method for the case of standard normal distributions with a subsequent transformation to arbitrary Gaussian densities is given in [26]. The LCD-based methods will be extended to the reduction of Dirac mixture densities in this paper.

1.4 Key ideas and results of the paper

The key idea of this paper is the systematic reapproximation of Dirac mixture densities by minimization of a novel distance measure. The distance measure compares the probability masses of both densities under certain kernels for all possible kernel locations and widths, which allows the use of integral measures for the mass functions. This approximation method is similar to the approximation of multivariate Gaussian densities by Dirac mixtures in [25]. However, calculating the distance measure between multivariate Gaussians and Dirac mixture densities in [25] requires a one-dimensional numerical integration, while the distance measure for comparing Dirac mixture densities with Dirac mixture densities proposed in this paper is given in closed form.

The resulting distance measure is smooth and does not suffer from local minima, so that standard optimization methods can be used for calculating the desired Dirac mixture approximation. The optimization results are deterministic and reproducible, which is in contrast to random selection procedures and most clustering methods.

The results for approximating 2000 samples from a standard normal distribution by a Dirac mixture approximation with $L = 10$, $L = 20$, and $L = 30$ components are shown in Figure 1.

1.5 Organization of the paper

In the next section, a rigorous formulation of the considered approximation problem is given. For comparing Dirac mixture densities, an alternative to the classical cumulative distribution, the so called Localized Cumulative Distribution (LCD) is introduced in Sect. 3. Based on this LCD, a generalization of the Cramér–von Mises-Distance, which is the integral squared distance between the LCD of the given density and the LCD of the approximate density is given in Sect. 4. This new distance measure is used for analysis purposes, i.e., for comparing the approximate Dirac mixture to the given one. The synthesis problem, i.e., determining the parameters of the approximate Dirac mixture in such a way that it is as close as possible to the given Dirac mixture according to the new distance measure is the topic of Sect. 5. Minimization is performed with a quasi-Newton method. The required gradient is derived in [27]. Examples of using the new reduction method on specific sample sets are given in Sect. 6. The new approach is discussed in Sect. 7 and an outlook to future work is given.

2 Problem formulation

We consider an N–dimensional Dirac mixture density with M components given by

$$\tilde{f}(\underline{x}) = \sum_{i=1}^{M} w_i^y \, \delta(\underline{x} - \underline{y}_i) \ , \tag{1}$$

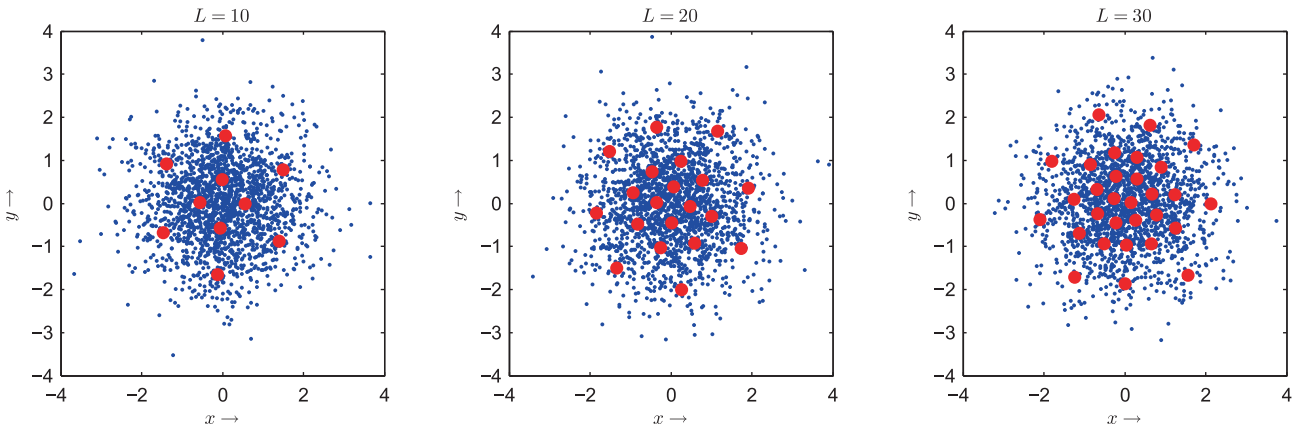

Figure 1: Dirac mixture approximation of 2000 samples from a standard normal distribution with $L = 10$, $L = 20$, and $L = 30$. Blue: Random samples representing the standard normal distribution. Red: Reduced point set.

with positive weights $w_i^y > 0$ for $i = 1, \ldots, M$, that sum up to one and M locations

$$\underline{y}_i = \left[y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(N)} \right]^T$$

for $i = 1, \ldots, M$. This density is approximated by another N–dimensional Dirac mixture density with L components given by

$$f(\underline{x}) = \sum_{i=1}^{L} w_i^x \, \delta(\underline{x} - \underline{x}_i) \ , \tag{2}$$

with positive weights $w_i^x > 0$ for $i = 1, \ldots, L$, that sum up to one and L locations

$$\underline{x}_i = \left[x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(N)} \right]^T$$

for $i = 1, \ldots, L$, where we assume $L \leq M$.

The goal is to select the weights w_i^x, $i = 1, \ldots, L$ and location parameters $\underline{x}_i$, $i = 1, \ldots, L$ of the approximating density $f(\underline{x})$ in such a way that $f(\underline{x})$ is as close as possible to the original density $\tilde{f}(\underline{x})$. For simplifying the notation, we define a parameter vector comprising weights and locations as

$$\underline{\eta} = \left[w_1^x, \ldots, w_L^x, \underline{x}_1^T, \ldots, \underline{x}_L^T \right]^T$$

and denote the dependence of the approximating density from the parameter vector by $f(\underline{x}, \underline{\eta})$. The number of parameters is $L - 1$ for the weights as they have to sum up to one plus $N \cdot L$ for the locations, giving a total of $(N+1) \cdot L - 1$ parameters.

For comparing the original density $\tilde{f}(\underline{x})$ and its approximation $f(\underline{x}, \underline{\eta})$, a distance measure for probability density functions is required, see [28] for an overview. Distance measures designed for directly comparing densities such as the integral squared distance, the Kullback-Leibler distance [29], or the Hellinger distance [28] are obviously ill-defined for comparing Dirac mixture densities. Distance measures comparing cumulative distributions instead of densities such as the Kolmogorov-Smirnov distance [30, p. 623] or the Cramér–von Mises-Distance [31] can be used for continuous densities and for Dirac mixture densities. However, they are difficult to use for $N \geq 2$, see [12] for details. For point sets, many distance measures are available [32] such as the Hausdorff distance. Typically, they do not view the point sets as densities and, therefore, do not consider weighted points. In addition, they perform an optimization themselves so that they have a high computational complexity and are not differentiable. A transport metric [33] defines the distance between two densities as the minimum cost of transferring one density into the other. It is also called Wasserstein distance [34]. This distance is well defined for Dirac mixture densities [35]. As the calculation of the distance includes an optimization itself, computational complexity is high and the distance is not easily differentiable. In this paper, the squared integral distance between transformations of the discrete densities by the so called Localized Cumulative Distribution (LCD) is used as a distance measure. This distance that we will denote by $D(\tilde{f}(\underline{x}), f(\underline{x}, \underline{\eta}))$ is given in closed form, efficient to compute, and differentiable.

The desired optimal parameters $\underline{\eta}^*$ are now obtained by minimizing the distance measure

$$\begin{aligned} \underline{\eta}^* &= \arg\min_{\underline{\eta}} D(\tilde{f}(\underline{x}), f(\underline{x}, \underline{\eta})) \\ &\text{s.t.} \quad w_i^x > 0 \text{ for } i = 1, \ldots, L \text{ and } \sum_{i=1}^{L} w_i^x = 1 \ , \end{aligned} \tag{3}$$

which is a constrained optimization problem.

When the weights of the approximating density are fixed a priori, the parameter vector is given by

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

is of length $N \cdot L$, and the optimization problem is unconstrained

$$\underline{\eta}^* = \arg\min_{\underline{\eta}} D(\tilde{f}(\underline{x}), f(\underline{x}, \underline{\eta})) \ . \tag{4}$$

Equally weighted approximating densities are an important special case of fixed weights. Here, we can distinguish two cases for the original density: i) The original Dirac mixture density might already have equally weighted components, so that the information is solely stored in the component locations. In this case, the goal of the approximation is a pure reduction of the number of components. ii) On the other hand, the components of the original Dirac mixture density might have different weights. This could be the result of, e.g., weighting a prior Dirac mixture density by a likelihood function in a Bayesian filtering setup. In that case, the approximation replaces an arbitrarily weighted Dirac mixture density by an equally weighted one. In the latter case, an equal number of components, i.e., $L = M$, can be useful.

So far, we considered the number L of components of the approximating density as given. When an upper bound $\bar{D}$ of the distance measure D is prespecified, the number of components will be adapted in such a way that the resulting distance D^* after performing the optimization is below this bound, i.e., $D^* < \bar{D}$.

3 Localized Cumulative Distribution

For the systematic reduction of the number of components of a given Dirac mixture density, a distance measure for comparing the original density and its approximation is required. However, Dirac mixture densities cannot be directly compared as they typically do not even share a common support. Typically, their corresponding cumulative distributions are used for comparison purposes, as is the case in certain statistical tests such as the Kolmogov-Smirnov test [30, p. 623]. However, it has been shown in [12] that although the cumulative distribution is well suited for comparing scalar densities, it exhibits several problems in higher-dimensional spaces: It is non-unique and non-symmetric. In addition, integral measures for comparing two cumulative distributions do not converge over infinite integration domains when the underlying Dirac mixture densities differ.

As an alternative transformation of densities, the Localized Cumulative Distribution (LCD) introduced in [12] is employed here in a generalized form. An LCD is an integral measure proportional to the mass concentrated in a region with a size parametrized by a vector $\underline{b}$ around test points $\underline{m}$. These regions are defined by kernels $K(\underline{x}-\underline{m},\underline{b})$ centered around $\underline{m}$ with size $\underline{b}$.

Definition 1 Let $\underline{\boldsymbol{x}}$ be a random vector with $\underline{\boldsymbol{x}} \in \mathbb{R}^N$, which is characterized by an N–dimensional probability density function $f: \mathbb{R}^N \to \mathbb{R}_+$. The corresponding Localized Cumulative Distribution (LCD) is defined as

$$F(\underline{m},\underline{b}) = \int_{\mathbb{R}^N} f(\underline{x})\, K(\underline{x}-\underline{m},\underline{b})\, d\underline{x}$$

with $\underline{b} \in \mathbb{R}_+^N$ and $F: \Omega \to [0,1], \Omega \subset \mathbb{R}^N \times \mathbb{R}_+^N$.

Definition 2 As a shorthand notation, we will denote the relation between the density $f(\underline{x})$ and its LCD $F(\underline{x},\underline{b})$ by

$$f(\underline{x}) \circ\!\!-\!\!\bullet\, F(\underline{m},\underline{b}) \ .$$

In this paper, we focus attention on separable kernels of the type

$$K(\underline{x}-\underline{m},\underline{b}) = \prod_{k=1}^{N} K(x^{(k)}-m^{(k)},b^{(k)}) \ .$$

Furthermore, we consider kernels with equal width in every dimension, i.e., $b^{(k)} = b$ for $k = 1,\ldots,N$, which gives

$$K(\underline{x}-\underline{m},b) = \prod_{k=1}^{N} K(x^{(k)}-m^{(k)},b) \ .$$

Rectangular, axis-aligned kernels as used in [12] are the obvious choice as they yield the probability mass of the considered density in a rectangular region centered around $\underline{m}$. They are well suited for analysis purposes and are used, e.g., when assessing the discrepancy of a sample set from a uniform distribution. However, for synthesizing a suitable approximation for a given (nonuniform) Dirac mixture with a smaller number of components, smooth kernels lead to simpler optimization problems. Here, we consider kernels of Gaussian type according to

$$K(\underline{x}-\underline{m},b) = \prod_{k=1}^{N} \exp\left(-\frac{1}{2}\frac{\left(x^{(k)}-m^{(k)}\right)^2}{b^2}\right) \ .$$

Based on a Gaussian kernel, an N-dimensional Dirac component $\delta(\underline{x}-\hat{\underline{x}})$ at location $\hat{\underline{x}}$ corresponds to its LCD $\Delta(\underline{m},b)$

$$\delta(\underline{x}-\hat{\underline{x}}) \circ\!\!-\!\!\bullet\, \Delta(\underline{m},b)$$

with

$$\begin{aligned}\Delta(\underline{m},b) &= \int_{\mathbb{R}^N} \delta(\underline{x}-\hat{\underline{x}})\, K(\underline{x}-\underline{m},\underline{b})\, d\underline{x} \\ &= \prod_{k=1}^{N} \exp\left(-\frac{1}{2}\frac{\left(\hat{x}^{(k)}-m^{(k)}\right)^2}{b^2}\right) \ .\end{aligned}$$

With this LCD of a single Dirac component, the LCD of the Dirac mixture in (2) is given by

$$F(\underline{m},b) = \sum_{i=1}^{L} w_i^x \prod_{k=1}^{N} \exp\left(-\frac{1}{2}\frac{\left(x_i^{(k)}-m^{(k)}\right)^2}{b^2}\right) \ . \quad (5)$$

For the original Dirac mixture $\tilde{f}(\underline{x})$ in (1), we obtain a similar result

$$\tilde{F}(\underline{m},b) = \sum_{i=1}^{M} w_i^y \prod_{k=1}^{N} \exp\left(-\frac{1}{2}\frac{(y_i^{(k)}-m^{(k)})^2}{b^2}\right) \ . \quad (6)$$

4 A Modified Cramér–von Mises-Distance

The Localized Cumulative Distribution (LCD) defined previously can now be used to derive a modified version of the Cramér–von Mises-Distance suitable for comparing Dirac

Mixtures. This new distance is defined as the integral of the square of the difference between the LCD of the true density $\tilde{f}(\underline{x})$ and the LCD of its approximation $f(\underline{x})$.

Definition 3 (Modified Cramér–von Mises-Distance) The distance D between two densities $\tilde{f}\colon \mathbb{R}^N \to \mathbb{R}_+$ and $f\colon \mathbb{R}^N \to \mathbb{R}_+$ is given in terms of their corresponding LCDs $\tilde{F}(\underline{x}, b)$ and $F(\underline{x}, b)$ as

$$D^2 = \int_{\mathbb{R}_+} w(b) \int_{\mathbb{R}^N} \left(\tilde{F}(\underline{m}, b) - F(\underline{m}, b)\right)^2 \, d\underline{m}\, db \ , \qquad (7)$$

where $w\colon \mathbb{R}_+ \to \mathbb{R}_+$ is a suitable weighting function.

A weighting function $w(b)$ has been introduced that controls how kernels of different sizes influence the resulting distance. This provides some degrees of freedom during the design of an approximation algorithm. Alternatively, a unit weighting function could be used while modifying the kernels accordingly.

Theorem 1 *By inserting the LCDs $\tilde{F}(\underline{m}, b)$ from (6) and $F(\underline{m}, b)$ from (5) into (7) and by using the weighting function*

$$w(b) = \begin{cases} \frac{1}{b^{N-1}} & b \in [0, b_{max}] \\ 0 & \textit{elsewhere} \end{cases} \ ,$$

the following expressions for the distance D

$$\begin{aligned} D^2 = &\sum_{i=1}^{M}\sum_{j=1}^{M} w_i^y w_j^y \, \gamma\left(\sum_{k=1}^{N}\left(y_i^{(k)} - y_j^{(k)}\right)^2\right) \\ &-2\sum_{i=1}^{L}\sum_{j=1}^{M} w_i^x w_j^y \, \gamma\left(\sum_{k=1}^{N}\left(x_i^{(k)} - y_j^{(k)}\right)^2\right) \\ &+\sum_{i=1}^{L}\sum_{j=1}^{L} w_i^x w_j^x \, \gamma\left(\sum_{k=1}^{N}\left(x_i^{(k)} - x_j^{(k)}\right)^2\right) \end{aligned}$$

with

$$\gamma(z) = \frac{\pi^{\frac{N}{2}}}{8}\left\{4\, b_{max}^2 \, \exp\left(-\frac{1}{2}\frac{z}{2\, b_{max}^2}\right) + z\, Ei\left(-\frac{1}{2}\frac{z}{2\, b_{max}^2}\right)\right\},$$

are obtained, where $Ei(z)$ denotes the exponential integral.

Proof. For the given specific weighting function $w(b)$, the distance measure is given by

$$D^2 = \int_0^{b_{\max}} \frac{1}{b^{N-1}} \int_{\mathbb{R}^N} \left(\tilde{F}(\underline{m}, b) - F(\underline{m}, b)\right)^2 \, d\underline{m}\, db \ . \qquad (8)$$

Inserting the LCDs $\tilde{F}(\underline{m}, b)$ and $F(\underline{m}, b)$ leads to

$$\begin{aligned} D^2 = \int_0^{b_{\max}} \frac{1}{b^{N-1}} \int_{\mathbb{R}^N} &\sum_{i=1}^{M}\sum_{j=1}^{M} w_i^y w_j^y \prod_{k=1}^{N} \exp\left(-\frac{1}{2}\frac{(y_i^{(k)} - m^{(k)})^2}{b^2}\right) \\ &\prod_{k=1}^{N} \exp\left(-\frac{1}{2}\frac{(y_j^{(k)} - m^{(k)})^2}{b^2}\right) \\ &-2\sum_{i=1}^{L}\sum_{j=1}^{M} w_i^x w_j^y \prod_{k=1}^{N} \exp\left(-\frac{1}{2}\frac{(x_i^{(k)} - m^{(k)})^2}{b^2}\right) \\ &\prod_{k=1}^{N} \exp\left(-\frac{1}{2}\frac{(y_j^{(k)} - m^{(k)})^2}{b^2}\right) \\ &+\sum_{i=1}^{L}\sum_{j=1}^{L} w_i^x w_j^x \prod_{k=1}^{N} \exp\left(-\frac{1}{2}\frac{(x_i^{(k)} - m^{(k)})^2}{b^2}\right) \\ &\prod_{k=1}^{N} \exp\left(-\frac{1}{2}\frac{(x_j^{(k)} - m^{(k)})^2}{b^2}\right) \\ &d\underline{m}\, db \ . \end{aligned}$$

Exchanging integration and summation gives

$$\begin{aligned} D^2 = &\sum_{i=1}^{M}\sum_{j=1}^{M} w_i^y w_j^y \int_0^{b_{\max}} \frac{1}{b^{N-1}} \prod_{k=1}^{N} \int_{\mathbb{R}} \exp\left(-\frac{1}{2}\frac{(y_i^{(k)} - m^{(k)})^2}{b^2}\right) \\ &\exp\left(-\frac{1}{2}\frac{(y_j^{(k)} - m^{(k)})^2}{b^2}\right) dm^{(k)}\, db \\ &-2\sum_{i=1}^{L}\sum_{j=1}^{M} w_i^x w_j^y \int_0^{b_{\max}} \frac{1}{b^{N-1}} \prod_{k=1}^{N} \int_{\mathbb{R}} \exp\left(-\frac{1}{2}\frac{(x_i^{(k)} - m^{(k)})^2}{b^2}\right) \\ &\exp\left(-\frac{1}{2}\frac{(y_j^{(k)} - m^{(k)})^2}{b^2}\right) dm^{(k)}\, db \\ &+\sum_{i=1}^{L}\sum_{j=1}^{L} w_i^x w_j^x \int_0^{b_{\max}} \frac{1}{b^{N-1}} \prod_{k=1}^{N} \int_{\mathbb{R}} \exp\left(-\frac{1}{2}\frac{(x_i^{(k)} - m^{(k)})^2}{b^2}\right) \\ &\exp\left(-\frac{1}{2}\frac{(x_j^{(k)} - m^{(k)})^2}{b^2}\right) dm^{(k)}\, db \end{aligned}$$

For further simplification, the following closed-form expression for the occurring integrals

$$\begin{aligned} &\int_{\mathbb{R}} \exp\left(-\frac{1}{2}\frac{(z_i - m)^2}{b^2}\right) \exp\left(-\frac{1}{2}\frac{(z_j - m)^2}{b^2}\right) dm \\ &\quad = \sqrt{\pi}\, b \, \exp\left(-\frac{1}{2}\frac{(z_i - z_j)^2}{2\, b^2}\right) \ , \qquad (9) \end{aligned}$$

is used. This gives

$$D^2 = \sum_{i=1}^{M}\sum_{j=1}^{M} w_i^y w_j^y \int_0^{b_{\max}} b\,\pi^{\frac{N}{2}} \prod_{k=1}^{N} \exp\left(-\frac{1}{2}\frac{\left(y_i^{(k)} - y_j^{(k)}\right)^2}{2\,b^2}\right) db$$
$$-2\sum_{i=1}^{L}\sum_{j=1}^{M} w_i^x w_j^y \int_0^{b_{\max}} b\,\pi^{\frac{N}{2}} \prod_{k=1}^{N} \exp\left(-\frac{1}{2}\frac{\left(x_i^{(k)} - y_j^{(k)}\right)^2}{2\,b^2}\right) db$$
$$+\sum_{i=1}^{L}\sum_{j=1}^{L} w_i^x w_j^x \int_0^{b_{\max}} b\,\pi^{\frac{N}{2}} \prod_{k=1}^{N} \exp\left(-\frac{1}{2}\frac{\left(x_i^{(k)} - x_j^{(k)}\right)^2}{2\,b^2}\right) db.$$

or

$$D^2 = \sum_{i=1}^{M}\sum_{j=1}^{M} w_i^y w_j^y \int_0^{b_{\max}} b\,\pi^{\frac{N}{2}} \exp\left(-\frac{1}{2}\frac{\sum_{k=1}^{N}\left(y_i^{(k)} - y_j^{(k)}\right)^2}{2\,b^2}\right) db$$
$$-2\sum_{i=1}^{L}\sum_{j=1}^{M} w_i^x w_j^y \int_0^{b_{\max}} b\,\pi^{\frac{N}{2}} \exp\left(-\frac{1}{2}\frac{\sum_{k=1}^{N}\left(x_i^{(k)} - y_j^{(k)}\right)^2}{2\,b^2}\right) db$$
$$+\sum_{i=1}^{L}\sum_{j=1}^{L} w_i^x w_j^x \int_0^{b_{\max}} b\,\pi^{\frac{N}{2}} \exp\left(-\frac{1}{2}\frac{\sum_{k=1}^{N}\left(x_i^{(k)} - x_j^{(k)}\right)^2}{2\,b^2}\right) db.$$

With

$$\int_0^{b_{\max}} b\,\exp\left(-\frac{1}{2}\frac{z}{2\,b^2}\right) db =$$
$$\frac{1}{8}\left\{4\,b_{\max}^2\,\exp\left(-\frac{1}{2}\frac{z}{2\,b_{\max}^2}\right) + z\,\mathrm{Ei}\left(-\frac{1}{2}\frac{z}{2\,b_{\max}^2}\right\}\right)$$

for $z > 0$, the final result is obtained. □

Remark 1 The exponential integral $\mathrm{Ei}(z)$ is defined as

$$\mathrm{Ei}(z) = \int_{-\infty}^{z} \frac{e^t}{t} dt \ .$$

For $z > 0$, $\mathrm{Ei}(z)$ is related to the incomplete gamma function $\Gamma(0, z)$ according to

$$\mathrm{Ei}(-z) = -\Gamma(0, z) \ .$$

Theorem 2 *For large b_{max}, the distance D is described by*

$$D^2 = \frac{\pi^{\frac{N}{2}}}{8}\left(D_{\underline{y}} - 2D_{\underline{xy}} + D_{\underline{x}}\right) + \frac{\pi^{\frac{N}{2}}}{4} C_b\, D_E \ , \qquad (10)$$

with the constant $C_b = \log(4\,b_{max}^2) - \Gamma$. Here, only the last term depends upon b_{max} and we have

$$D_{\underline{y}} = \sum_{i=1}^{M}\sum_{j=1}^{M} w_i^y w_j^y\, xlog\left(\sum_{k=1}^{N}\left(y_i^{(k)} - y_j^{(k)}\right)^2\right) \ ,$$

$$D_{\underline{xy}} = \sum_{i=1}^{L}\sum_{j=1}^{M} w_i^x w_j^y\, xlog\left(\sum_{k=1}^{N}\left(x_i^{(k)} - y_j^{(k)}\right)^2\right) \ ,$$

$$D_{\underline{x}} = \sum_{i=1}^{L}\sum_{j=1}^{L} w_i^x w_j^x\, xlog\left(\sum_{k=1}^{N}\left(x_i^{(k)} - x_j^{(k)}\right)^2\right) \ ,$$

with $xlog(z) = z \cdot \log(z)$, where $xlog(0) = 0$, and

$$D_E = \sum_{k=1}^{N}\left(\sum_{i=1}^{L} w_i^x x_i^{(k)} - \sum_{i=1}^{M} w_i^y y_i^{(k)}\right)^2 \ .$$

Proof. For small $z > 0$, the exponential integral can be approximated by

$$\mathrm{Ei}(-z) \approx \Gamma + \log(z) - z \ , \qquad (11)$$

where $\Gamma \approx 0.5772$ is the Euler gamma constant. As a result, the function $\gamma(z)$ can be approximated according to

$$\begin{aligned}\gamma(z) &\approx \frac{\pi^{\frac{N}{2}}}{8}\left\{4\,b_{\max}^2\,\exp\left(-\frac{1}{2}\frac{z^2}{2\,b_{\max}^2}\right)\right.\\ &\quad \left. + z\left(\Gamma + \log\left(\frac{1}{2}\frac{z}{2\,b_{\max}^2}\right) - \frac{1}{2}\frac{z}{2\,b_{\max}^2}\right)\right\}\\ &\approx \frac{\pi^{\frac{N}{2}}}{8}\left\{4\,b_{\max}^2 + z\left(\Gamma - \log(4\,b_{\max}^2) + \log(z)\right)\right\}\\ &= \frac{\pi^{\frac{N}{2}}}{8}\left\{4\,b_{\max}^2 - C_b\,z + \mathrm{xlog}(z)\right\} \ .\end{aligned}$$

Inserting the first term into the distance measure D in Theorem 1 cancels due to the fact that

$$\frac{\pi^{\frac{N}{2}}}{2} b_{\max}^2 \left\{\sum_{i=1}^{M}\sum_{j=1}^{M} w_i^y w_j^y - 2\sum_{i=1}^{L}\sum_{j=1}^{M} w_i^x w_j^y + \sum_{i=1}^{L}\sum_{j=1}^{L} w_i^x w_j^x\right\}$$
$$= \frac{\pi^{\frac{N}{2}}}{2} b_{\max}^2 \left\{\sum_{i=1}^{M} w_i^y - \sum_{i=1}^{L} w_i^x\right\}^2 = 0 \ .$$

Inserting the second term according to

$$-\frac{\pi^{\frac{N}{2}}}{8} C_b \sum_{k=1}^{N}\left\{\sum_{i=1}^{M}\sum_{j=1}^{M} w_i^y w_j^y \left(y_i^{(k)} - y_j^{(k)}\right)^2\right.$$
$$-2\sum_{i=1}^{L}\sum_{j=1}^{M} w_i^x w_j^y \left(x_i^{(k)} - y_j^{(k)}\right)^2$$
$$\left. + \sum_{i=1}^{L}\sum_{j=1}^{L} w_i^x w_j^x \left(x_i^{(k)} - x_j^{(k)}\right)^2\right\} \ ,$$

can be written as

$$
\begin{aligned}
&-\frac{\pi^{\frac{N}{2}}}{8} C_b \sum_{k=1}^{N} \Bigg\{ \sum_{i=1}^{M} w_i^y \left(y_i^{(k)}\right)^2 - 2 \sum_{i=1}^{M} \sum_{j=1}^{M} w_i^y w_j^y y_i^{(k)} y_j^{(k)} \\
&+ \sum_{i=1}^{M} w_i^y \left(y_i^{(k)}\right)^2 - 2 \Bigg[\sum_{i=1}^{L} w_i^x \left(x_i^{(k)}\right)^2 \\
&- 2 \sum_{i=1}^{L} \sum_{j=1}^{M} w_i^x w_j^y x_i^{(k)} y_j^{(k)} + \sum_{i=1}^{M} w_i^y \left(y_i^{(k)}\right)^2 \Bigg] \\
&+ \sum_{i=1}^{L} w_i^x \left(x_i^{(k)}\right)^2 - 2 \sum_{i=1}^{L} \sum_{j=1}^{L} w_i^x w_j^x x_i^{(k)} x_j^{(k)} + \sum_{i=1}^{L} w_i^x \left(x_i^{(k)}\right)^2 \Bigg\}.
\end{aligned}
$$

Canceling corresponding terms finally gives

$$
\frac{\pi^{\frac{N}{2}}}{4} C_b \sum_{k=1}^{N} \left(\sum_{i=1}^{M} w_i^y y_i^{(k)} - \sum_{i=1}^{L} w_i^x x_i^{(k)} \right)^2 .
$$

Inserting the third term gives the remaining expressions. □

Remark 2 For equal expected values of the densities $\tilde{f}(\underline{x})$ and $f(\underline{x})$, the distance measure in Theorem 2 does not depend upon $b_{\max}$ anymore.

Remark 3 Thanks to symmetry, the summation in $D_{\underline{y}}$ can be reduced to $i = 1, \ldots, M$ and $j = i+1, \ldots, M$ by removing redundant terms, resulting in $M(M-1)/2$ operations. A similar argument holds for $D_{\underline{x}}$.

5 Reduction

The goal is to find the optimal L weights w_i^x, $i = 1, \ldots, L$ and L locations $\underline{x}_i$, $i = 1, \ldots, L$ of the approximating Dirac mixture density such that the distance measure D in (10) in Theorem 2 is minimized according to (3). The last term in (10) can be viewed as a penalty term for different means of the densities $\tilde{f}$ and f, where the penalty function is given by D_E and the penalty coefficient depends on $b_{\max}$. $b_{\max}$ is set to a large value. Alternatively, the penalty term can be removed in (10) and a set of N constraints

$$
\sum_{i=1}^{M} w_i^y y_i^{(k)} - \sum_{i=1}^{L} w_i^x x_i^{(k)} = 0
$$

for $k = 1, \ldots, N$ is used instead for assuring equal means of the densities $\tilde{f}$ and f. When weights and locations are optimized according to (3), this just adds more constraints to the constrained optimization problem. In the case of just optimizing the locations, however, the unconstrained optimization problem in (4) becomes a constrained optimization problem when removing the penalty term.

The distance measure D is a smooth and twice continuously differentiable function, with the gradient G given in closed form in [27]. Standard optimization methods can be used for finding the minimum of D in (3) and (4). For the unconstrained optimization, we use a quasi-Newton method, specifically the Matlab R2014b implementation of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm included in `fminunc`. BFGS is an optimization method independently proposed by Broyden [36, 37], Fletcher [38], Goldfarb [39], and Shanno [40]. In contrast to Newton methods, quasi-Newton methods do not explicitly require the Hessian matrix and directly estimate the required inverse of the Hessian, thus avoiding the costly inversion. For constrained optiization, we use the Matlab R2014b function `fmincon`.

Evaluation of the distance D in (10) requires $\mathcal{O}\left((M^2 + M \cdot L + L^2) \cdot N\right)$ operations, with M the number of Dirac components in the original Dirac mixture and L the number of Dirac components used for the approximation. N is the number of dimensions. The first term $D_{\underline{y}}$ only depends on parameters of the original density $\tilde{f}$. As it does not depend on the desired parameter vector $\underline{\eta}$, its calculation is only required when the absolute value of the distance is of interest, e.g., when comparing approximations for different numbers L of samples. It is calculated once before the optimization. During optimization, only changes of the distance measure caused by the parameter vector $\underline{\eta}$ are needed. When the value of the distance measure is not required, the first term is not calculated at all. Calculating changes of the distance with respect to changes in the parameter vector $\underline{\eta}$ costs $\mathcal{O}\left((M \cdot L + L^2) \cdot N\right)$ operations. When the number of components L of the approximation is much smaller than the number of Dirac mixture components M of the given original density, i.e., we have $L \ll M$, the complexity of calculating the third term in (10) can be neglected. In that case, we obtain a complexity of $\mathcal{O}\left(M \cdot L \cdot N\right)$ operations, which is linear in M, L, and N.

The optimization is not sensitive to starting values for the parameter vector $\underline{\eta}$, so initialization is usually performed by simply drawing the locations $\underline{x}_i$, $i = 1, \ldots, L$ of the approximating Dirac mixture density from a Gaussian density with variances given by the variances of the original density $\tilde{f}$. Random samples can be used for this purpose. When a fully deterministic optimization is desired, deterministic samples precomputed with the method in [25] are used. As an alternative, starting values for the locations of f are obtained by a random selection of L samples from the M samples of the original density $\tilde{f}$. When

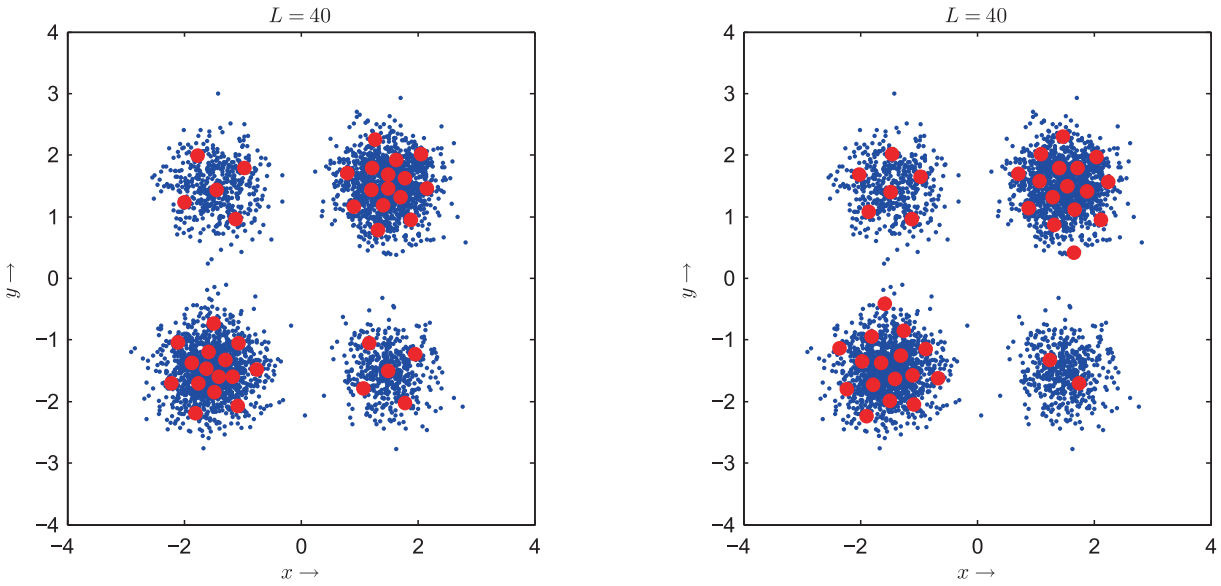

Figure 2: Reduction of a Gaussian mixture density with four components and a varying number of samples per component from 4000 points to 40 points. Blue: Random samples representing the Gaussian mixture density. Red: Reduced point set. (left) LCD reduction. (right) Result of k-means clustering.

the weights are also optimized, equal weights are used as starting values.

Arbitrary additional constraints can be added either as penalty functions or as explicit constraints for the optimization problem. This includes constraints for maintaining certain moments of the original density $\tilde{f}$ or constraints for avoiding certain regions of the state space.

6 Numerical Evaluation

The proposed method for the optimal reduction of Dirac mixture densities is evaluated and compared to a standard clustering technique, the k-means algorithm [7]. The implementation shipped with Matlab R2014b is used. The focus is on equally weighted approximating densities f.

The results of approximating random samples from a standard normal distribution have already been shown in Figure 1 in the introduction. We now approximate random samples from a Gaussian mixture density with four isotropic components placed at $[\pm 1.5, \pm 1.5]^T$ with standard deviation 0.4, see Figure 2. It is important to note that we have a total of $M = 4000$ samples, but the number of samples differs for each component: We have 500 samples for components $(1, 1)$ and $(2, 2)$ and 1500 samples for components $(1, 2)$ and $(2, 1)$. After the reduction from $M = 4000$ samples down to $L = 40$ samples, we would expect that the probability masses for each component of the Gaussian mixture density are maintained. This is exactly the case for the proposed LCD reduction as can be seen in Figure 2 on the left side, where we end up with 5 samples for components $(1, 1)$ and $(2, 2)$ and 15 samples for components $(1, 2)$ and $(2, 1)$. For k-means clustering, shown on the right side in Figure 2, this is not the case, so the original distribution is not maintained. In addition, the results of k-means clustering are not reproducible. As the Matlab-implementation employs randomly selected initial starting points for the representatives, several runs produce different results even for the same sample set. This not only means different point positions, but also a different number of points associated to each Gaussian mixture component. For different sample sets, k-means clustering produces significantly different results. In contrast, the LCD reduction method produces very similar results when run several times on the same sample set even when using random initialization. For different sample sets, the positions of the points change only slightly while the number of points associated to each Gaussian mixture component stays the same.

Another way to demonstrate that the proposed reduction method maintains the probability mass distribution is to compare histograms of the samples before and after reduction. To simplify visualization, histograms are calculated for the marginals in x-direction. Figure 3 shows the histogram of the originals samples on the left side. The histogram after reduction with the proposed LCD reduction method is shown in the middle, while the histogram of the results obtained with k-means are shown on the right side.

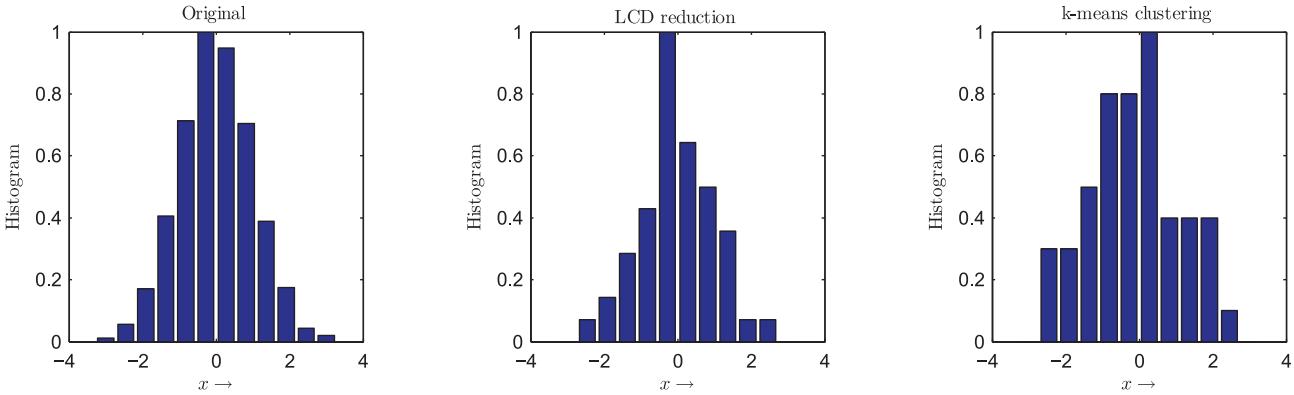

Figure 3: Normalized histograms of projections onto x-axis for reducing $M = 5000$ samples of a two-dimensional standard normal distribution to $L = 50$ samples. (left) Marginal of original samples. (middle) Marginal of LCD reduction result. (right) Marginal of result of k-means clustering.

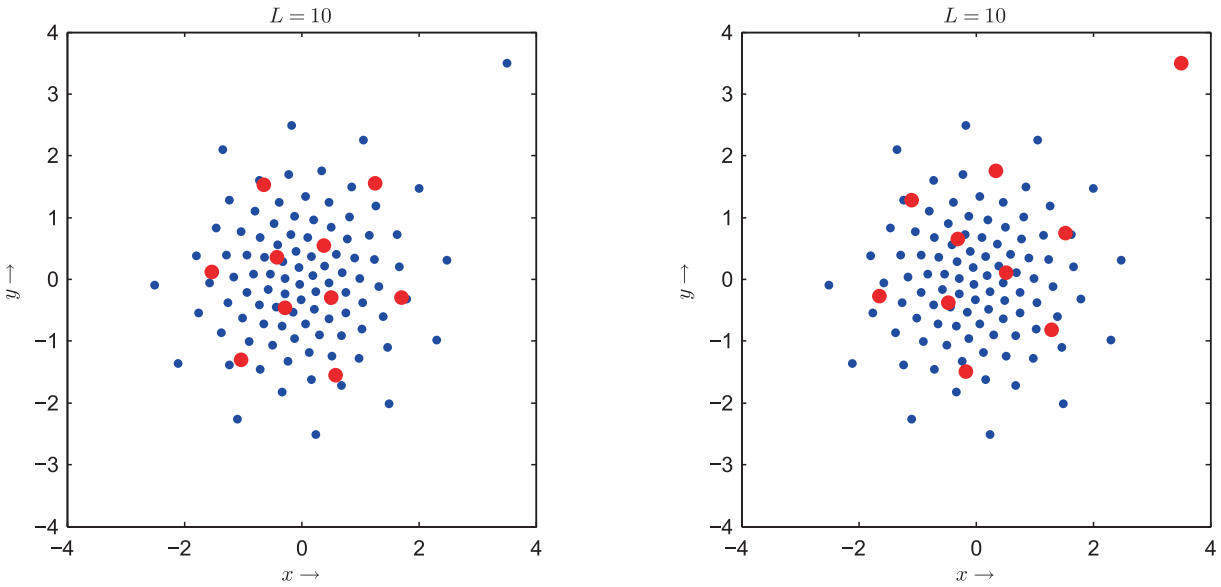

Figure 4: Blue: Deterministic samples representing a standard normal distribution. One sample is replaced by an outlier at $[3.5, 3.5]^T$. Red: Reduced point set. (left) LCD reduction. (right) k-means clustering.

It is obvious that the histogram of the LCD reduction is much closer to the original histogram than the histogram of k-means.

We now consider the reduction of deterministic samples from a standard normal distribution corrupted by a single outlier. $M = 100$ deterministic samples of a standard normal distribution are shown in Figure 4. The samples are calculated with the method from [25]. One sample is replaced with an outlier located at $[3.5, 3.5]^T$. The point set is reduced to $L = 10$ samples. The left side shows the result of the LCD reduction. The samples are well placed and only slightly shifted due to the outlier. On the right side, k-means clustering produces a result heavily disturbed by the outlier. In fact, one sample of the reduced point set is placed directly on the outlier, which significantly changes the mass distribution and the moments. Instead of representing 1 % of the distribution as before the reduction, the outlier now allocates 10 %.

Finally, we investigate the robustness of the reduction methods with respect to missing data. For that purpose, we generate 2500 samples and remove samples located within three vertical strips, see Figure 5. The remaining samples are reduced down to $L = 25$ samples. Figure 5 left shows the result of the LCD reduction, which almost gives the same results as before. The right side shows the result of k-means clustering, where it is obvious that samples are more or less placed along lines and the original mass distribution is not well maintained.

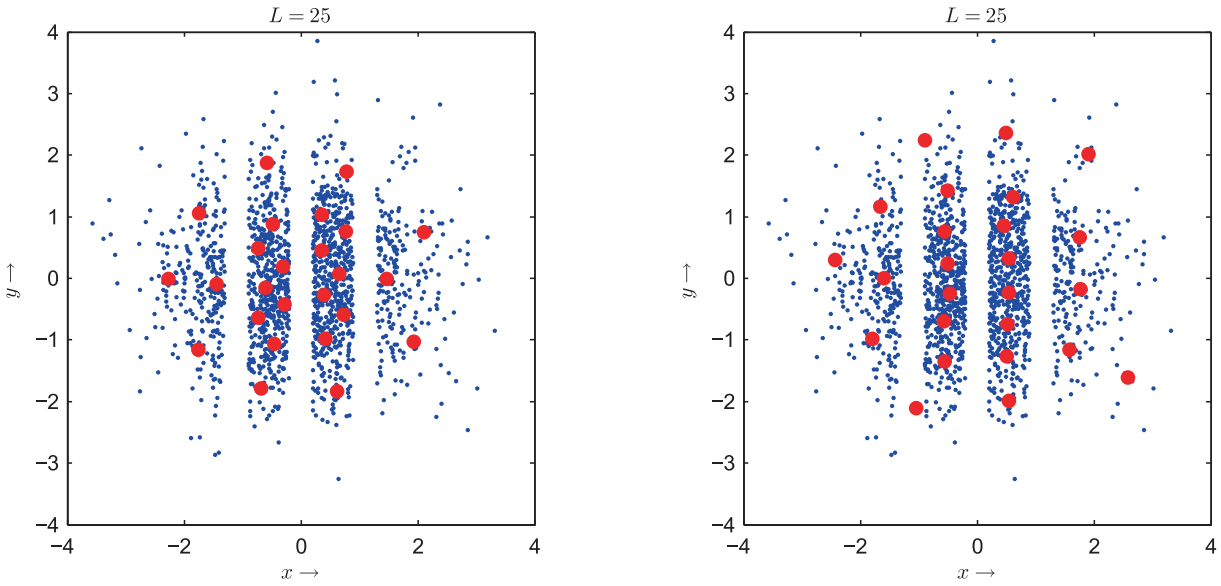

Figure 5: Blue: Random samples representing a standard normal distribution with some samples removed along three vertical lines. Red: Reduced point set. (left) LCD reduction. (right) k-means clustering.

7 Discussion

A systematic approach for approximating a given Dirac mixture density by another one with less components has been introduced that is radically different from current clustering or vector quantization approaches. The (weights and) locations of the approximating density are calculated by minimizing a global distance measure, a generalization of the well-known Cramér–von Mises-Distance to the multivariate case. This generalization is obtained by defining an alternative to the classical cumulative distribution, the Localized Cumulative Distribution (LCD), as a characterization of discrete random quantities, which is unique and symmetric also in the multivariate case.

Although kernels are used to define the LCD, this is not a kernel method. The distance measure is obtained by integrating over all possible kernels with all locations and widths, so that the final expression does not contain any kernel.

The given Dirac mixture might be the result from random sampling or from certain processing steps involving deterministic Dirac mixtures. In any case, the resulting approximating Dirac mixture is fully deterministic and the optimization process gives reproducible results.

Compared to clustering methods that find cluster heads minimizing the distance to their nearest neighbors, which is a local method, the LCD reduction globally matches the mass distributions of the given point set and its approximation. This leads to a smooth distance measure with almost no local minima that can be efficiently minimized with standard optimization procedures. However, it is important to note that due to its operating principle, the proposed reduction method does not provide an explicit mapping from old components to new components.

Constraints on the state variables can easily be considered when performing the approximation of the given density. An obvious application is the explicit avoidance of certain regions in the state space in order to obey physical constraints. Another application is to maintain certain moments of the original density $\tilde{f}$ during the reduction.

Large data sets occur when performing Dirac mixture based state estimation in high–dimensional spaces or when considering product spaces of Dirac mixture densities. For a very large number of components, the computational effort for performing a direct reduction might be too large. For coping with this complexity issue, the proposed approach offers the unique feature of hierarchical approximation. For that purpose, the data set is decomposed into several smaller sets that are individually approximated. The resulting Dirac components of the individual approximations are then collected into a single approximating Dirac mixture, which subsequently is further approximated to yield the desired number of components. Of course, this approximation hierarchy may consist of more intermediate approximation steps.

Acknowledgement: The author would like to thank Dipl.-Inform. Henning Eberhardt for many fruitful discussions

on this topic and the nice ideas for visualizing the performance of the proposed new reduction algorithm.

The author would also like to thank the anonymous reviewers for helpful comments that led to significant changes in this manuscript.

References

1. J. S. Liu and R. Chen, "Sequential Monte Carlo Methods for Dynamic Systems," *Journal of the American Statistical Association*, vol. 93, no. 443, pp. 1032–1044, Sep. 1998. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01621459.1998.10473765
2. H. Eberhardt, V. Klumpp, and U. D. Hanebeck, "Optimal Dirac Approximation by Exploiting Independencies," in *Proceedings of the 2010 American Control Conference (ACC 2010)*, Baltimore, Maryland, USA, Jun. 2010.
3. C. Chlebek, A. Hekler, and U. D. Hanebeck, "Stochastic Nonlinear Model Predictive Control Based on Progressive Density Simplification," in *Proceedings of the 51st IEEE Conference on Decision and Control (CDC 2012)*, Maui, Hawaii, USA, Dec. 2012.
4. M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, Feb. 2002.
5. O. Straka, J. Dunik, and M. Šimandl, "Measures of Non-Gaussianity in Unscented Kalman Filter Framework," in *Proceedings of the 17th International Conference on Information Fusion (Fusion 2014)*, Salamanca, Spain, Jul. 2014.
6. S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.
7. J. MacQueen, "Some methods for classification and analysis of multivariate observations." The Regents of the University of California, 1967. [Online]. Available: http://projecteuclid.org/euclid.bsmsp/1200512992
8. Y. Linde, A. Buzo, and R. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, Jan. 1980.
9. D. Crouse, P. Willett, K. Pattipati, and L. Svensson, "A look at Gaussian mixture reduction algorithms," in *2011 Proceedings of the 14th International Conference on Information Fusion (FUSION)*, Jul. 2011, pp. 1–8.
10. J. Williams and P. Maybeck, "Cost-Function-Based Gaussian Mixture Reduction for Target Tracking," in *Proceedings of the Sixth International Conference of Information Fusion, 2003*, vol. 2, Jul. 2003, pp. 1047–1054.
11. A. Runnalls, "Kullback-Leibler Approach to Gaussian Mixture Reduction," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 43, no. 3, pp. 989–999, Jul. 2007.
12. U. D. Hanebeck and V. Klumpp, "Localized Cumulative Distributions and a Multivariate Generalization of the Cramér-von Mises Distance," in *Proceedings of the 2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2008)*, Seoul, Republic of Korea, Aug. 2008, pp. 33–39.
13. M. West, "Approximating Posterior Distributions by Mixture," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 55, no. 2, pp. 409–422, Jan. 1993. [Online]. Available: http://www.jstor.org/stable/2346202
14. T. Lefebvre, H. Bruyninckx, and J. De Schutter, "The Linear Regression Kalman Filter," in *Nonlinear Kalman Filtering for Force-Controlled Robot Tasks*, ser. Springer Tracts in Advanced Robotics, 2005, vol. 19.
15. S. Julier, J. Uhlmann, and H. F. Durrant-Whyte, "A New Method for the Nonlinear Transformation of Means and Covariances in Filters and Estimators," *IEEE Transactions on Automatic Control*, vol. 45, no. 3, pp. 477–482, Mar. 2000.
16. S. J. Julier, "The Scaled Unscented Transformation," in *Proceedings of the 2002 IEEE American Control Conference (ACC 2002)*, vol. 6, Anchorage, Alaska, USA, May 2002, pp. 4555–4559.
17. D. Tenne and T. Singh, "The Higher Order Unscented Filter," in *Proceedings of the 2003 IEEE American Control Conference (ACC 2003)*, vol. 3, Denver, Colorado, USA, Jun. 2003, pp. 2441–2446.
18. M. F. Huber and U. D. Hanebeck, "Gaussian Filter based on Deterministic Sampling for High Quality Nonlinear Estimation," in *Proceedings of the 17th IFAC World Congress (IFAC 2008)*, vol. 17, no. 2, Seoul, Republic of Korea, Jul. 2008.
19. G. Kurz, I. Gilitschenski, and U. D. Hanebeck, "Recursive Nonlinear Filtering for Angular Data Based on Circular Distributions," in *Proceedings of the 2013 American Control Conference (ACC 2013)*, Washington D. C., USA, Jun. 2013.
20. I. Gilitschenski, G. Kurz, and U. D. Hanebeck, "Bearings-Only Sensor Scheduling Using Circular Statistics," in *Proceedings of the 16th International Conference on Information Fusion (Fusion 2013)*, Istanbul, Turkey, Jul. 2013.
21. G. Kurz, F. Faion, and U. D. Hanebeck, "Constrained Object Tracking on Compact One-dimensional Manifolds Based on Directional Statistics," in *Proceedings of the Fourth IEEE GRSS International Conference on Indoor Positioning and Indoor Navigation (IPIN 2013)*, Montbeliard, France, Oct. 2013.
22. U. D. Hanebeck, "Truncated Moment Problem for Dirac Mixture Densities with Entropy Regularization," *arXiv preprint: Systems and Control (cs.SY)*, Aug. 2014. [Online]. Available: http://arxiv.org/abs/1408.7083
23. 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 2006 IEEE Conference on Decision and Control (CDC 2006)*, San Diego, California, USA, Dec. 2006.
24. ——, "Dirac Mixture Density Approximation Based on Minimization of the Weighted Cramér-von Mises Distance," in *Proceedings of the 2006 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2006)*, Heidelberg, Germany, Sep. 2006, pp. 512–517.
25. U. D. Hanebeck, M. F. Huber, and V. Klumpp, "Dirac Mixture Approximation of Multivariate Gaussian Densities," in *Proceedings of the 2009 IEEE Conference on Decision and Control (CDC 2009)*, Shanghai, China, Dec. 2009.
26. I. Gilitschenski and U. D. Hanebeck, "Efficient Deterministic Dirac Mixture Approximation," in *Proceedings of the 2013 American Control Conference (ACC 2013)*, Washington D. C., USA, Jun. 2013.

27. U. D. Hanebeck, "Optimal Reduction of Multivariate Dirac Mixture Densities," *arXiv preprint: Systems and Control (cs.SY)*, Nov. 2014. [Online]. Available: http://arxiv.org/abs/1411.4586
28. A. L. Gibbs and F. E. Su, "On Choosing and Bounding Probability Metrics," *International Statistical Review*, vol. 70, no. 3, pp. 419–435, 2002. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1751-5823.2002.tb00178.x/abstract
29. S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, Mar. 1951. [Online]. Available: http://www.jstor.org/stable/2236703
30. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C (2Nd Ed.): The Art of Scientific Computing*. New York, NY, USA: Cambridge University Press, 1992.
31. D. D. Boos, "Minimum Distance Estimators for Location and Goodness of Fit," *Journal of the American Statistical Association*, vol. 76, no. 375, pp. 663–670, Sep. 1981. [Online]. Available: http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1981.10477701
32. T. Eiter and H. Mannila, "Distance measures for point sets and their computation," *Acta Informatica*, vol. 34, no. 2, pp. 109–133, Feb. 1997. [Online]. Available: http://link.springer.com/article/10.1007/s002360050075
33. C. Villani, *Optimal Transport – Old and New*, ser. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2009, no. 338.
34. M. Baum, P. Willett, and U. D. Hanebeck, "On Wasserstein Barycenters and MMOSPA Estimation (to appear)," *IEEE Signal Processing Letters*, 2015.
35. X. Nguyen, "Convergence of Latent Mixing Measures in Finite and Infinite Mixture Models," *The Annals of Statistics*, vol. 41, no. 1, pp. 370–400, Feb. 2013, arXiv: 1109.3250. [Online]. Available: http://arxiv.org/abs/1109.3250
36. C. G. Broyden, "The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations," *IMA Journal of Applied Mathematics*, vol. 6, no. 1, pp. 76–90, Mar. 1970. [Online]. Available: http://imamat.oxfordjournals.org/content/6/1/76
37. ——, "The Convergence of a Class of Double-rank Minimization Algorithms 2. The New Algorithm," *IMA Journal of Applied Mathematics*, vol. 6, no. 3, pp. 222–231, Sep. 1970. [Online]. Available: http://imamat.oxfordjournals.org/content/6/3/222
38. R. Fletcher, "A New Approach to Variable Metric Algorithms," *The Computer Journal*, vol. 13, no. 3, pp. 317–322, Jan. 1970. [Online]. Available: http://comjnl.oxfordjournals.org/content/13/3/317
39. D. Goldfarb, "A Family of Variable-Metric Methods Derived by Variational Means," *Mathematics of Computation*, vol. 24, no. 109, pp. 23–26, Jan. 1970. [Online]. Available: http://www.jstor.org/stable/2004873
40. D. F. Shanno, "Conditioning of Quasi-Newton Methods for Function Minimization," *Mathematics of Computation*, vol. 24, no. 111, pp. 647–656, 1970. [Online]. Available: http://www.ams.org/mcom/1970-24-111/S0025-5718-1970-0274029-X/