Density Estimation from Weighted Samples with the Localized Cumulative Distribution

Daniel Frisch and Uwe D. Hanebeck

Intelligent Sensor-Actuator-Systems Laboratory (ISAS)
Institute for Anthropomatics and Robotics (IAR)
Karlsruhe Institute of Technology (KIT), Germany
{*firstname*}.{*lastname*}@{*kit.edu, ieee.org*}

Abstract: We propose a novel method for nonparametric density estimation from weighted samples. Thereby the density is represented in discretized form as a regular grid of square roots of probabilities or density values. We define a distance measure between samples and density grid by computing the Localized Cumulative Distributions of both and then a modified Cramér–von Mises distance between them. We achieve smoothness with an additional Fisher Information regularization. The square root representation helps twofold: it enforces the nonnegativity constraint for the resulting density and simplifies computation of the Fisher Information. In a numerical evaluation, we compute the Kullback-Leibler divergence of our result to the ground truth density and demonstrate our method outperforming a conventional kernel density estimator (KDE) even for its best bandwidth choice.
Julia source code is available here: `https://github.com/KIT-ISAS/IFAC26_Frisch`

Keywords: Density estimation, weighted samples, weighted data, regular grid, square root density, Localized Cumulative Distribution (LCD), Cramér–von Mises distance, Fisher Information, regularization.

1. INTRODUCTION

1.1 Relevance of Density Estimation

Density estimation from samples, i.e., reconstructing an unknown probability density function from a finite set of observations, is a fundamental problem in statistics and machine learning, with applications across numerous scientific and engineering domains, Silverman (2018), Scott (1992), Izenman (1991). Distribution-free methods that make minimal assumptions about the underlying data-generating process are essential, e.g., to nonparametric statistical inference.

The challenge intensifies when dealing with *weighted* samples. This occurs in imbalanced classification (minority class is much rarer than majority class), Feng et al. (2021), Jiang and Nachum (2020), cost-sensitive learning (different types of errors have different cost), Araf et al. (2024), importance sampling (data generated by a different density than the target density), Tokdar and Kass (2010), e.g., survey sampling and propensity score analysis (where the exposure was not random), Ridgeway et al. (2015), active learning (human interactively assigns weights), Polyzos et al. (2024), boosting algorithms (focus on examples that previous weak learners got wrong), Schapire (2013), and outlier-robust learning (down-weigh potential outliers and adversarial examples), Zhang and Luo (2015). Efficient and accurate density estimation methods that can handle arbitrary sample weights while producing smooth, interpretable representations remain an active area of research.

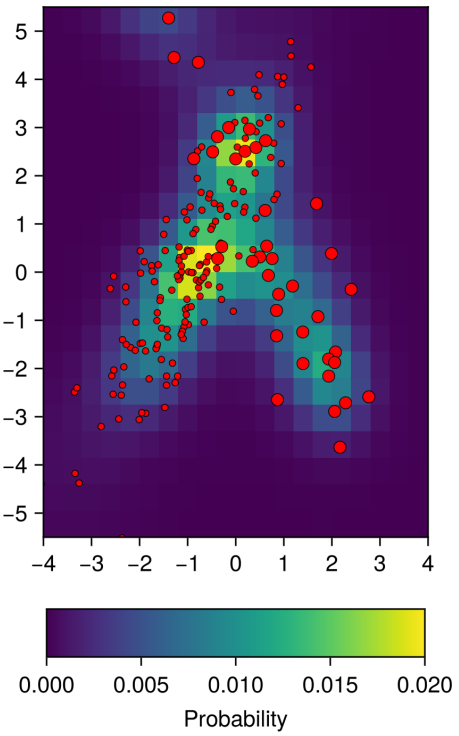


Fig. 1. Proposed density estimation on a regular grid from 200 random samples (red) with different weights (marker size). Regularization parameter $\lambda = 0.05$, computation time 1.06 s. More information in Section 4.1.

1.2 Related Work

One way to density estimation is via *parametric* densities. For exponential family densities, such as the Gaussian density, the maximum likelihood parameter estimate can be determined from the sufficient statistics. Much more flexibility in shapes of the density is possible with Gaussian mixtures. Their parameters can be estimated from weighted samples via an expectation maximization method, Frisch and Hanebeck (2021), Frisch and Hanebeck (2020).

For *nonparametric* estimation, kernel-based methods are very popular, Parzen (1962). One problem is the selection of (and restriction to) one kernel width whose choice is essential especially if little data is available, Heidenreich et al. (2013). Adaptive methods address this by varying the bandwidth locally, e.g., via balloon or sample-point estimators, Terrell and Scott (1992), or Abramson's square root law, Abramson (1982). For *weighted* samples specifically, weighted KDEs have been proposed that incorporate sample weights into the kernel sum, Gisbert (2003), but these still depend on a single global bandwidth. The Localized Cumulative Distribution (LCD) circumvents this problem by averaging over a range of kernel widths, Hanebeck and Klumpp (2008). It allows to formulate a distance measure between all combinations of continuous and discrete densities and has been used to sample deterministically from samples, Hanebeck (2015), from Gaussians, Gilitschenski and Hanebeck (2013), and Gaussian mixtures, Gilitschenski et al. (2014), Giraldo-Grueso et al. (2025), and from a regular grid that arises from the independence of noise variables, Eberhardt et al. (2010). In a sense, we are doing the opposite of Eberhardt et al. (2010): estimating a regular grid from samples.

While the LCD gives a smooth distance measure, it does not perform any regularization. Thus, if the density of data points is anywhere coarser than the spacing of the grid points, we need some regularization. For this purpose we propose using the Fisher information, Hanebeck et al. (2024), an information-theoretic roughness measure for densities. Nonparametric continuous density estimation from unweighted and weighted samples with Fisher information regularization has already been done in Prossel and Hanebeck (2024) and Frisch and Hanebeck (2024), respectively, but so far only for the univariate case.

We also use an internal square root density representation, which for one thing accomplishes the nonnegativity constraint. Square root density representations on regular grids have for this purpose already been proposed on the circle, Pfaff et al. (2019), the sphere, Pfaff et al. (2020b), and the torus, Pfaff et al. (2020a). In addition, the square root representation also simplifies the Fisher information computation, Hanebeck et al. (2024).

1.3 Contributions

While the individual components (LCD with Cramér–von Mises distance, Fisher information with square root density) have been explored previously, each in a different context, our contribution lies in their novel combination for weighted density estimation on regular grids. Specifically: (i) We combine the LCD-based distance with Fisher information regularization for multidimensional grid-based density estimation, which to the best of our knowledge has not been done before. (ii) The square root density representation simultaneously enforces nonnegativity and simplifies the Fisher information to a quadratic form. (iii) We provide analytical gradients for all terms, enabling efficient gradient-based optimization. (iv) We demonstrate superior performance over KDE on both equally and non-equally weighted samples.

2. PROBLEM AND METHODS DESCRIPTION

In this section, we introduce the notation of our density representations (weighted samples and regular grid) and describe the employed density distance measure between them as well as the smoothness measure for the grid representation.

2.1 Density Representations

Given Data. The given S-dimensional data samples are represented as a weighted Dirac Mixture Density (DMD) $\tilde{f}(\underline{x})$

$$\tilde{f}(\underline{x}) = \sum_{i=1}^{L} w_i \cdot \delta(\underline{x} - \underline{\hat{x}}_i) \ , \quad \underline{\hat{x}}_i = [\hat{x}_{i,1} \ \dots \ \hat{x}_{i,S}]^\top \ , \quad (1)$$

where $\underline{\hat{x}}_i \in \mathbb{R}^S$ are the sample locations, $w_i \in \mathbb{R}_0^+$ the associated weights with $\sum_{i=1}^{L} w_i = 1$, and $\delta(\cdot)$ the Dirac delta function with $\int \delta(\underline{x}) f(\underline{x}) \, \mathrm{d}\underline{x} = f(\underline{0})$.

Desired Density. The desired density $f(\underline{x})$ is represented as a regular grid of square root density values. We first define the grid point locations $\underline{x}^{\mathrm{g}} \in \mathbb{R}^S$

$$\underline{x}^{\mathrm{g}}[\underline{n}] = \underline{x}^{\mathrm{g}}[n_1, \dots, n_S] = \left[x_1^{\mathrm{g}}[n_1] \ \dots \ x_S^{\mathrm{g}}[n_S]\right]^\top \ , \quad (2)$$

with multi-index $\underline{n}$ consisting of dimension-wise indices $n_s = 1, \dots, N_s$, for dimensions $s = 1, \dots, S$. The total number of grid points is thus $N_{\mathrm{grid}} = \prod_{s=1}^{S} N_s$.

The density on the grid can then either be represented as density function value at or probability of each grid point. In the case of equidistant spacing these two representations are equivalent up to a scaling factor. Here, we choose the interpretation of square roots of probabilities $r[\underline{n}] \in \mathbb{R}$ at each grid location $\underline{x}^{\mathrm{g}}[\underline{n}]$ and can therefore represent $f(\underline{x})$ as a DMD

$$\begin{aligned} f(\underline{x}) &= \sum_{n_1=1}^{N_1} \cdots \sum_{n_S=1}^{N_S} r^2[n_1, \dots, n_S] \cdot \delta(\underline{x} - \underline{x}^{\mathrm{g}}[n_1, \dots, n_S]) \\ &= \sum_{\underline{n}=\underline{1}}^{\underline{N}} r^2[\underline{n}] \cdot \delta(\underline{x} - \underline{x}^{\mathrm{g}}[\underline{n}]) \ , \end{aligned} \quad (3)$$

where $\underline{n}$ and $\underline{N}$ are short forms for the multi-index, $\sum_{\underline{n}=\underline{1}}^{\underline{N}}$ represents sums over each index in the multi-index, $\underline{x}^{\mathrm{g}}[\underline{n}] \in \mathbb{R}^S$ are the grid locations, and $r[\underline{n}] \in \mathbb{R}$ are the associated square root density values.

2.2 Density Distance Measure

Localized Cumulative Distribution (LCD). The LCD of a density $f(\underline{x})$ is defined as

$$F(\underline{m}, b) = \int_{\mathbb{R}^S} f(\underline{x}) \cdot K(\underline{x}, \underline{m}, b) \, \mathrm{d}\underline{x} \ , \quad (4)$$

where $K(\underline{x}, \underline{m}, b)$ is a kernel function centered at $\underline{m}$ with bandwidth b. In this work, we use an isotropic Gaussian kernel $K(\underline{x}, \underline{m}, b) = \exp\left\{-\frac{1}{2b^2}\|\underline{x} - \underline{m}\|_2^2\right\}$. That gives us the LCDs of (1) and (3), respectively, as

$$\tilde{F}(\underline{m}, b) = \sum_{i=1}^{L} w_i \cdot K(\underline{m}, \hat{\underline{x}}_i, b) \ , \tag{5}$$

$$F(\underline{m}, b) = \sum_{\underline{n}=\underline{1}}^{\underline{N}} r^2[\underline{n}] \cdot K(\underline{m}, \underline{x}^{\mathrm{g}}[\underline{n}], b) \ . \tag{6}$$

Cramér–von Mises Distance. Now we can compare $\tilde{f}(\underline{x})$ and $f(\underline{x})$ by comparing $\tilde{F}(\underline{m}, b)$ and $F(\underline{m}, b)$ via the Cramér–von Mises distance $D_b(b)$ with

$$\begin{aligned} D_b^2(b) &= \int_{\mathbb{R}^S} \left[\tilde{F}(\underline{m}, b) - F(\underline{m}, b)\right]^2 \mathrm{d}\underline{m} \qquad (7)\\ &= \int_{\mathbb{R}^S} \tilde{F}^2(\underline{m}, b) - 2\tilde{F}(\underline{m}, b)F(\underline{m}, b) + F^2(\underline{m}, b)\,\mathrm{d}\underline{m} \\ &= D_{b,1}(b) - 2D_{b,2}(b) + D_{b,3}(b) \ , \end{aligned}$$

where $D_{b,1}(b)$ is constant w.r.t. $r[\underline{n}]$ and can thus be ignored for optimization, and

$$\begin{aligned} D_{b,2}(b) &= \int_{\mathbb{R}^S} \tilde{F}(\underline{m}, b)F(\underline{m}, b)\,\mathrm{d}\underline{m} \qquad (8)\\ &= \left(\sqrt{\pi}b\right)^S \sum_{i=1}^{L}\sum_{\underline{n}=\underline{1}}^{\underline{N}} w_i r^2[\underline{n}] \exp\left\{-\frac{1}{4b^2}\|\hat{\underline{x}}_i - \underline{x}^{\mathrm{g}}[\underline{n}]\|_2^2\right\} \ , \\ D_{b,3}(b) &= \int_{\mathbb{R}^S} F^2(\underline{m}, b)\,\mathrm{d}\underline{m} \qquad (9)\\ &= \left(\sqrt{\pi}b\right)^S \sum_{\underline{n}=\underline{1}}^{\underline{N}}\sum_{\underline{n}'=\underline{1}}^{\underline{N}} r^2[\underline{n}]r^2[\underline{n}'] \\ &\qquad \cdot \exp\left\{-\frac{1}{4b^2}\|\underline{x}^{\mathrm{g}}[\underline{n}] - \underline{x}^{\mathrm{g}}[\underline{n}']\|_2^2\right\} \ . \end{aligned}$$

Modified Cramér–von Mises Distance. In order to overcome the restriction to a single bandwidth b and the corresponding parameter selection, we modify the Cramér–von Mises distance by integrating over b with weight function $w(b)$, Hanebeck et al. (2009), yielding the modified Cramér–von Mises distance D with

$$D^2 = \int_0^{b_{\max}} w(b) D_b^2(b)\,\mathrm{d}b \ , \tag{10}$$

where $w(b) = \frac{1}{b^{S-1}}$. Following Hanebeck (2015), for large $b_{\max}$, we can simplify the arising exponential integral $\mathrm{Ei}(\cdot)$ via $\mathrm{Ei}(-z) \approx -\Gamma - \ln(z) - z$, where Γ is the Euler-Mascheroni constant, yielding

$$\begin{aligned} D^2 &= \frac{\pi^{S/2}}{8}(D_1 - 2D_2 + D_3) \qquad (11)\\ &\quad + \frac{\pi^{S/2}}{4}\left(\log 4b_{\max}^2 - \Gamma\right)D_E + 4b_{\max}^2 D_C \ , \end{aligned}$$

where D_1 is constant w.r.t. $r[\underline{n}]$ and can thus be ignored for optimization, and

$$D_2 = \sum_{i=1}^{L}\sum_{\underline{n}=\underline{1}}^{\underline{N}} w_i r^2[\underline{n}]\,\mathrm{xlog}\left(\|\hat{\underline{x}}_i - \underline{x}^{\mathrm{g}}[\underline{n}]\|_2^2\right) \ , \tag{12}$$

$$D_3 = \sum_{\underline{n}=\underline{1}}^{\underline{N}}\sum_{\underline{n}'=\underline{1}}^{\underline{N}} r^2[\underline{n}]r^2[\underline{n}']\,\mathrm{xlog}\left(\|\underline{x}^{\mathrm{g}}[\underline{n}] - \underline{x}^{\mathrm{g}}[\underline{n}']\|_2^2\right) \ , \tag{13}$$

$$D_E = \sum_{s=1}^{S}\left(\sum_{i=1}^{L} w_i \hat{x}_{i,s} - \sum_{\underline{n}=\underline{1}}^{\underline{N}} r^2[\underline{n}]\underline{x}^{\mathrm{g}}[\underline{n}]\right)^2 \ , \tag{14}$$

and $D_C = \left(1 - \sum_{\underline{n}=\underline{1}}^{\underline{N}} r^2[\underline{n}]\right)^2$, where $\mathrm{xlog}(z)$ is defined as $\mathrm{xlog}(z) = \begin{cases} z \cdot \log(z) \ , & z > 0 \\ 0 \ , & z = 0 \end{cases}$. Roughly speaking, D_2 encourages placing density mass close to the samples, D_3 encourages spreading out the density mass, D_E encourages matching the means, and D_C enforces normalization of the density. So far, this is similar to Hanebeck (2015) and can be used for sample reduction, i.e., the samples should be spaced much denser than the grid points everywhere. In this work, however, we aim at high resolution density estimation from even few samples. Therefore, we will now include a smoothness measure based on the Fisher information.

2.3 Density Smoothness Measure

A common measure for smoothness of a density $f(\underline{x})$ is the Fisher information. It can be computed as described in (Toscani, 2017, Eq. 1),

$$I_F(f) = \int_{\{\mathbb{R}^S,\, f>0\}} \frac{\|\nabla f(\underline{x})\|^2}{f(\underline{x})}\,\mathrm{d}\underline{x} \ . \tag{15}$$

With the square root representation $r(\underline{x}) = \sqrt{f(\underline{x})}$, according to (Hanebeck et al., 2024, Eq. 19), this simplifies to

$$\begin{aligned} I_F(f) = I_F^R(r) &= 4\int_{\mathbb{R}^S} \|\nabla r(\underline{x})\|_2^2\,\mathrm{d}\underline{x} \qquad (16)\\ &= 4\int_{\mathbb{R}^S} \left[\frac{\partial r(\underline{x})}{\partial x_1}\right]^2 + \ldots + \left[\frac{\partial r(\underline{x})}{\partial x_S}\right]^2 \mathrm{d}\underline{x} \ . \end{aligned}$$

With the discretized density (3) on the regular grid (2) using simple forward differences, this becomes

$$I_F^R(r) = 4\sum_{s=1}^{S}\sum_{\underline{n}=\underline{1}}^{\underline{N}-\underline{e}_s}\left(\frac{r[\underline{n}+\underline{e}_s] - r[\underline{n}]}{x_s^{\mathrm{g}}[n_s+1] - x_s^{\mathrm{g}}[n_s]}\right)^2 V[\underline{n}] \tag{17}$$

$$= 4\sum_{s=1}^{S}\sum_{\underline{n}=\underline{1}+\underline{e}_s}^{\underline{N}}\left(\frac{r[\underline{n}] - r[\underline{n}-\underline{e}_s]}{x_s^{\mathrm{g}}[n_s] - x_s^{\mathrm{g}}[n_s-1]}\right)^2 V[\underline{n}-\underline{e}_s] \ , \tag{18}$$

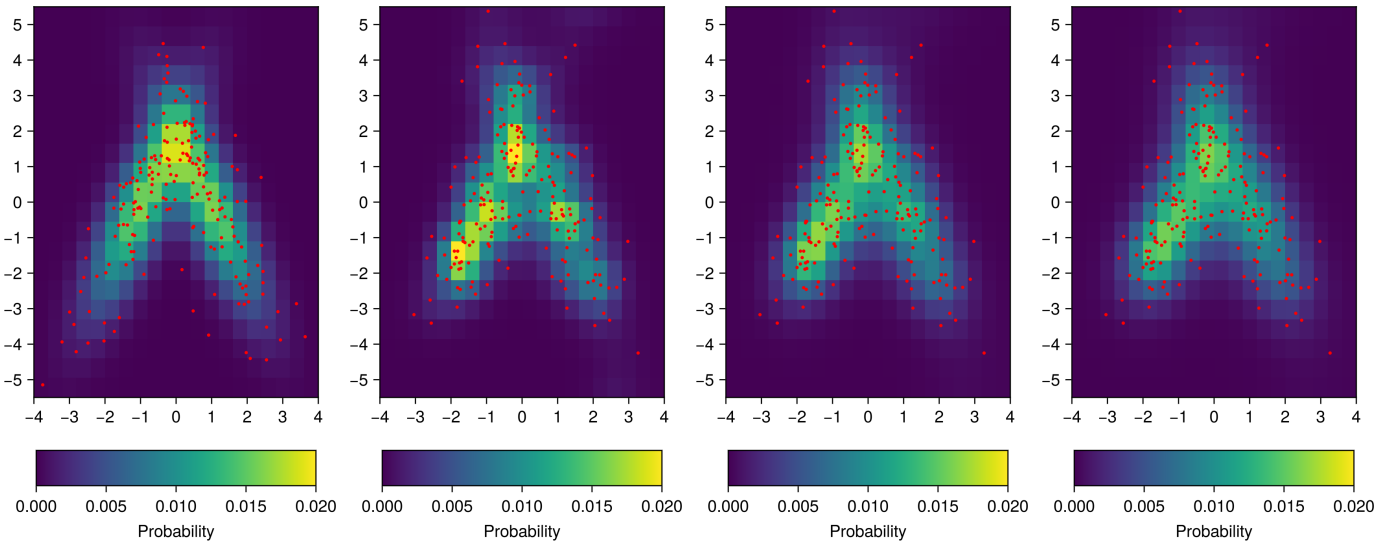
where $\underline{e}_s$ is the unit vector in the direction of dimension s, and $V[\underline{n}] = \prod_{s=1}^{S}(x_s^{\mathrm{g}}[n_s+1] - x_s^{\mathrm{g}}[n_s])$ is the volume element around grid point $\underline{x}^{\mathrm{g}}[\underline{n}]$. The alternative notation (18) is stated because it is helpful to compute the gradient (25).

3. OPTIMIZATION

3.1 Optimization Problem

We now want to find the density representation $r[\underline{n}]$ that minimizes the modified Cramér–von Mises distance D (11) while at the same time having low Fisher information $I_F^R(f)$ (17). This leads to the optimization problem

$$r^{\mathrm{opt}}[\underline{n}] = \underset{r[\underline{n}]}{\arg\min}\, J(r[\underline{n}]) \ , \tag{19}$$



(a) ground truth (b) $\lambda = 0.01$ (2.12 s) (c) $\lambda = 0.05$ (1.09 s) (d) $\lambda = 0.1$ (0.88 s)

Fig. 2. Graphical evaluation of the proposed density estimation on a regular 2D grid with various regularization parameters λ. x coordinate on horizontal, y coordinate on vertical axis. Computation times of optimizing (20) are shown in seconds. We also show the ground truth density for reference (a).

$$J(r[\underline{n}]) = D^2 + \lambda \cdot I_F^R(f) \ , \tag{20}$$

where the $N_{\text{grid}} = \prod_{s=1}^{S} N_s$ root density values $r[\underline{n}]$ are the parameters to be optimized, and $\lambda \in \mathbb{R}_0^+$ is a regularization parameter controlling the trade-off between fitting the data and smoothness.

3.2 Gradient

Very helpful for the optimization procedure are the analytical gradients w.r.t. $r[\underline{n}]$. The gradients of the individual terms of the cost function are given by

$$\frac{\partial D_2}{\partial r[\underline{n}]} = 2r[\underline{n}] \sum_{i=1}^{L} w_i \operatorname{xlog}\left(\|\hat{\underline{x}}_i - \underline{x}^{\text{g}}[\underline{n}]\|_2^2\right) \ , \tag{21}$$

$$\frac{\partial D_3}{\partial r[\underline{n}]} = 4r[\underline{n}] \sum_{\underline{n}'=\underline{1}}^{\underline{N}} r^2[\underline{n}'] \operatorname{xlog}\left(\|\underline{x}^{\text{g}}[\underline{n}] - \underline{x}^{\text{g}}[\underline{n}']\|_2^2\right) \ , \tag{22}$$

$$\frac{\partial D_E}{\partial r[\underline{n}]} = 4r[\underline{n}] \sum_{s=1}^{S} \underline{x}_s^{\text{g}}[n_s] \left(\sum_{\underline{n}'=\underline{1}}^{\underline{N}} r^2[\underline{n}'] \underline{x}_s^{\text{g}}[n'_s] - \sum_{i=1}^{L} w_i \hat{x}_{i,s} \right) \ , \tag{23}$$

$$\frac{\partial D_C}{\partial r[\underline{n}]} = -4r[\underline{n}] \left(1 - \sum_{\underline{n}'=\underline{1}}^{\underline{N}} r^2[\underline{n}'] \right) \ , \tag{24}$$

and

$$\frac{\partial I_F^R(f)}{\partial r[\underline{n}]} = 8 \sum_{s=1}^{S} \frac{r[\underline{n}] - r[\underline{n} - \underline{e}_s]}{\underline{x}^{\text{g}}[\underline{n}] - \underline{x}^{\text{g}}[\underline{n} - \underline{e}_s]} V[\underline{n} - \underline{e}_s] - \frac{r[\underline{n} + \underline{e}_s] - r[\underline{n}]}{\underline{x}^{\text{g}}[\underline{n} + \underline{e}_s] - \underline{x}^{\text{g}}[\underline{n}]} V[\underline{n}] \ . \tag{25}$$

To improve computation speed, we cached the outputs of xlog in (12) and (13), and the respective gradients (21) and (22).

3.3 Computation

All functions and gradients have been implemented in `Julia`, see Bezanson et al. (2017). We use the `Optim.jl` package by Mogensen and Riseth (2018) to solve the optimization problem. Best runtime was achieved with the algorithms `LBFGS()`, Liu and Nocedal (1989). Per optimizer iteration, evaluating the objective and gradient is dominated by the D_3 term (13) and (22) at $\mathcal{O}(N_{\text{grid}}^2)$. The cached xlog tables also dominate memory at $\mathcal{O}(N_{\text{grid}}^2)$. Both grow exponentially with S via $N_{\text{grid}} = \prod_{s=1}^{S} N_s$. The optimizer itself adds only $\mathcal{O}(m \cdot N_{\text{grid}})$ work per iteration (L-BFGS history size m), negligible compared to the objective. With K outer iterations and a small constant c of line-search evaluations per iteration, the total cost is $\mathcal{O}(K \cdot c \cdot N_{\text{grid}}^2)$.

4. EVALUATION

We define as ground truth $f^{\text{gt}}(\underline{x})$ a Gaussian mixture in 2D ($S = 2$) with two equally weighted components with parameters

$$\mu_1 = \begin{bmatrix} -1 \\ 0 \end{bmatrix} \ , \qquad \mathbf{C}_1 = \begin{bmatrix} 1^2 & 1.6 \\ 1.6 & 2^2 \end{bmatrix} \ , \tag{26}$$

$$\mu_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \ , \qquad \mathbf{C}_2 = \begin{bmatrix} 1^2 & -1.6 \\ -1.6 & 2^2 \end{bmatrix} \ , \tag{27}$$

from which we draw 100 samples, respectively, that are then combined. The resulting $L = 200$ samples constitute the reference density $\tilde{f}(\underline{x})$ whose underlying continuous density we estimate on a regular grid with $N_1 = N_2 = 20$, yielding 400 grid points, equally spaced in the range $[-4, 4] \times [-5.5, 5.5]$. We perform the proposed density estimation, with parameter $b_{\max} = 30$. The estimated densities $f(\underline{x})$, for various regularization parameters λ, are shown in Fig. 1 and Fig. 2. With increasing λ, the density gets smoother, but also starts to lose details of the original density.

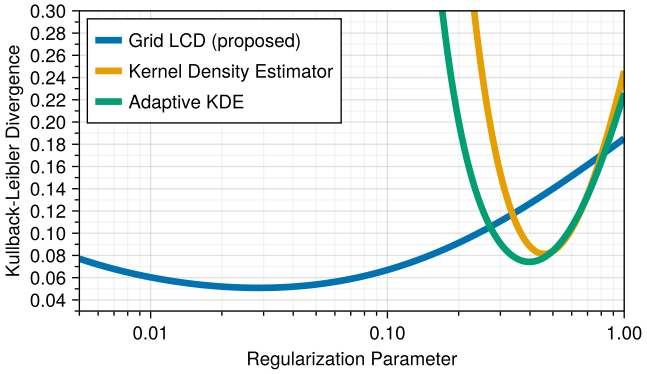


Fig. 3. Numerical evaluation. Abscissa is regularization parameter λ for proposed method (blue) or bandwidth for KDE (orange) or pilot bandwidth for adaptive KDE (green). Proposed method outperforms KDE for a wide range of λ.

For a quantitative evaluation, we compute the Kullback-Leibler Divergence (KLD) between the ground truth density $f^{\mathrm{gt}}(\underline{x})$ and the estimated density $f(\underline{x})$ that is represented by $r[\underline{n}]$, via

$$\mathrm{KLD}(f^{\mathrm{gt}}, f) = \sum_{\underline{n}=\underline{1}}^{\underline{N}} f^{\mathrm{gt}}(\underline{x}^{\mathrm{g}}[\underline{n}]) \cdot \log\left(\frac{f^{\mathrm{gt}}(\underline{x}^{\mathrm{g}}[\underline{n}])}{r^2[\underline{n}]}\right) . \quad (28)$$

The lowest KLD of 0.0509 is achieved for $\lambda = 0.0285$, see blue line in Fig. 3.

We also compare our method to a conventional KDE with Gaussian kernels and various bandwidths. The best KLD of this estimate is 0.12, at a kernel bandwidth of 0.51, see orange line in Fig. 3. Additionally, we compare against an adaptive KDE using Abramson's square root law, Abramson (1982), where each sample receives a local bandwidth inversely proportional to the square root of a pilot density estimate. This adaptive approach is shown in green in Fig. 3. Our proposed method outperforms both the fixed-bandwidth and adaptive KDE. Furthermore, the valley of suitable λ is flatter for the proposed method than the valley of suitable kernel widths for the KDEs, therefore our method is less sensitive to the parameter choice.

4.1 Weighted Experiment

To validate the method on genuinely weighted data, we also performed an importance sampling experiment: samples were drawn from a biased proposal density (an 80/20 mixture instead of the 50/50 target) and corrected via importance weights that restored the 50/50 target. The proposed method successfully recovers the target density from these non-uniformly weighted samples, see Fig. 1, confirming its applicability to the weighted setting motivated in the introduction.

5. CONCLUSION

We proposed a method that can learn a discretized probability density (defined on a regular grid) from given data points or samples. The connection between sample locations and density function value is established via the LCD and a modified Cramér–von Mises distance, so the method is independent of a specific choice of kernel width. Regularization is done via a Fisher information penalty term, an information-theoretic roughness measure defined specifically for such purposes. Our estimated density is closer to the ground truth density in terms of KLD than standard KDEs, even for an explicitly optimized kernel bandwidth.

Due to the regular grid representation, this method in its current form is feasible for low-dimensional problems only, say, 1D, 2D, and 3D. Furthermore, choosing the regularization parameter λ must be done via multiple trials and manual inspection, or via cross-validation, shrinking the available data points.

In the future, we will seek to introduce a low-rank tensor representation, Govaers et al. (2019), Frisch and Hanebeck (2025), to handle higher-dimensional problems. In such a representation, the root density grid $r[\underline{n}]$ is decomposed into a sum of rank-one terms. Both the LCD and the Fisher information Frisch and Hanebeck (2026) can be decomposed along the tensor factors. We will also seek to automatically determine the optimal regularization parameter λ. Finally, using a suitable kernel from previous works, Frisch et al. (2020), this method may be applicable also on manifolds.

REFERENCES

Abramson, I.S. (1982). On Bandwidth Variation in Kernel Estimates-A Square Root Law. *The Annals of Statistics*, 10(4), 1217–1223.

Araf, I., Idri, A., and Chairi, I. (2024). Cost-Sensitive Learning for Imbalanced Medical Data: A Review. *Artificial Intelligence Review*, 57(4), 80. doi:10.1007/s10462-023-10652-8.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V.B. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM review*, 59(1), 65–98.

Eberhardt, H., Klumpp, V., and Hanebeck, U.D. (2010). Optimal Dirac Approximation by Exploiting Independencies. In *Proceedings of the 2010 American Control Conference (ACC 2010)*. Baltimore, Maryland, USA. doi:10.1109/acc.2010.5530503.

Feng, Y., Zhou, M., and Tong, X. (2021). Imbalanced Classification: A Paradigm-Based Review. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 14(5), 383–406. doi:10.1002/sam.11538.

Frisch, D. and Hanebeck, U.D. (2020). Progressive Bayesian Filtering with Coupled Gaussian and Dirac Mixtures. In *Proceedings of the 23rd International Conference on Information Fusion (Fusion 2020)*. Virtual. doi:10.23919/FUSION45008.2020.9190540.

Frisch, D. and Hanebeck, U.D. (2021). Gaussian Mixture Estimation from Weighted Samples. In *Proceedings of the 2021 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2021)*. Karlsruhe, Germany. doi:10.1109/MFI52462.2021.9591161.

Frisch, D. and Hanebeck, U.D. (2024). Gaussian Mixture Particle Filter Step based on Method of Moments. In *Proceedings of the 27th International Conference on Information Fusion (FUSION 2024)*. Venice, Italy. doi:10.23919/FUSION59988.2024.10706458.

Frisch, D. and Hanebeck, U.D. (2025). Fokker-Planck Prediction on the Cylindric Manifold using Tensor Decomposition of a Regular Grid. In *17th Symposium Sensor Data Fusion: Trends, Solutions, Applications (SDF*

2025). Bonn, Germany. doi:10.1109/SDF67080.2025.11330240.

Frisch, D. and Hanebeck, U.D. (2026). Interpolation of Probability Densities on a Grid in Tensor Decomposition Representation. In *Proceedings of the 29th International Conference on Information Fusion (FUSION 2026)*. Trondheim, Norway.

Frisch, D., Li, K., and Hanebeck, U.D. (2020). Optimal Reduction of Dirac Mixture Densities on the 2-Sphere. In *Proceedings of the 1st Virtual IFAC World Congress (IFAC-V 2020)*. doi:10.1016/j.ifacol.2020.12.1856.

Gilitschenski, I. and Hanebeck, U.D. (2013). Efficient Deterministic Dirac Mixture Approximation of Gaussian Distributions. In *Proceedings of the 2013 American Control Conference (ACC 2013)*. Washington D.C., USA. doi:10.1109/acc.2013.6580197.

Gilitschenski, I., Steinbring, J., Hanebeck, U.D., and Simandl, M. (2014). Deterministic Dirac Mixture Approximation of Gaussian Mixtures. In *Proceedings of the 17th International Conference on Information Fusion (Fusion 2014)*. Salamanca, Spain.

Giraldo-Grueso, F., Popov, A.A., Hanebeck, U.D., and Zanetti, R. (2025). Optimal Grid Point Sampling for Point Mass Filtering. In *Proceedings of the AAS/AIAA Space Flight Mechanics Meeting, Kaua'i, Hawaii*. Kaua'i, Hawaii.

Gisbert, F.J.G. (2003). Weighted Samples, Kernel Density Estimators and Convergence. *Empirical Economics*, 28(2), 335–351. doi:10.1007/s001810200134.

Govaers, F., Demissie, B., Khan, A., Ulmke, M., and Koch, W. (2019). Tensor Decomposition-Based Multitarget Tracking in Cluttered Environments. *Journal of Advances in Information Fusion*, 14(1), 86.

Hanebeck, U.D. (2015). Optimal Reduction of Multivariate Dirac Mixture Densities. *at – Automatisierungstechnik*, 63(4), 265–278. doi:10.1515/auto-2015-0005.

Hanebeck, U.D., Frisch, D., and Prossel, D. (2024). Closed-Form Information-Theoretic Roughness Measures for Mixture Densities. In *Proceedings of the 2024 American Control Conference (ACC 2024)*. Toronto, Canada. doi:10.23919/ACC60939.2024.10645015.

Hanebeck, U.D., Huber, M.F., and Klumpp, V. (2009). Dirac Mixture Approximation of Multivariate Gaussian Densities. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC 2009)*. Shanghai, China. doi:10.1109/CDC.2009.5400649.

Hanebeck, U.D. and Klumpp, V. (2008). 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)*, 33–39. Seoul, Republic of Korea. doi:10.1109/MFI.2008.4648104.

Heidenreich, N.B., Schindler, A., and Sperlich, S. (2013). Bandwidth Selection for Kernel Density Estimation: A Review of Fully Automatic Selectors. *AStA Advances in Statistical Analysis*, 97(4), 403–433. doi:10.1007/s10182-013-0216-y.

Izenman, A.J. (1991). Recent Developments in Nonparametric Density Estimation. *Journal of the American Statistical Association*, 86(413), 205–224.

Jiang, H. and Nachum, O. (2020). Identifying and Correcting Label Bias in Machine Learning. In S. Chiappa and R. Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, 702–712. PMLR.

Liu, D.C. and Nocedal, J. (1989). On the Limited Memory BFGS Method for Large Scale Optimization. *Mathematical Programming*, 45(1), 503–528. doi:10.1007/BF01589116.

Mogensen, P.K. and Riseth, A.N. (2018). Optim: A Mathematical Optimization Package for Julia. *Journal of Open Source Software*, 3(24), 615. doi:10.21105/joss.00615.

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3), 1065–1076.

Pfaff, F., Li, K., and Hanebeck, U.D. (2019). Fourier Filters, Grid Filters, and the Fourier-Interpreted Grid Filter. In *Proceedings of the 22nd International Conference on Information Fusion (Fusion 2019)*. Ottawa, Canada. doi:10.23919/FUSION43075.2019.9011318.

Pfaff, F., Li, K., and Hanebeck, U.D. (2020a). Estimating Correlated Angles Using the Hypertoroidal Grid Filter. In *Proceedings of the 2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2020)*. Virtual. doi:10.1109/MFI49285.2020.9235220.

Pfaff, F., Li, K., and Hanebeck, U.D. (2020b). The Spherical Grid Filter for Nonlinear Estimation on the Unit Sphere. In *Proceedings of the 1st Virtual IFAC World Congress (IFAC-V 2020)*. doi:10.1016/j.ifacol.2020.12.031.

Polyzos, K.D., Lu, Q., and Giannakis, G.B. (2024). Weighted Ensembles for Adaptive Active Learning. *IEEE Transactions on Signal Processing*, 72, 4178–4190. doi:10.1109/TSP.2024.3450270.

Prossel, D. and Hanebeck, U.D. (2024). Spline-Based Density Estimation Minimizing Fisher Information. In *Proceedings of the 27th International Conference on Information Fusion (FUSION 2024)*. Venice, Italy. doi:10.23919/FUSION59988.2024.10706290.

Ridgeway, G., Kovalchik, S.A., Griffin, B.A., and Kabeto, M.U. (2015). Propensity Score Analysis with Survey Weighted Data. *Journal of Causal Inference*, 3(2), 237–249. doi:10.1515/jci-2014-0039.

Schapire, R.E. (2013). Explaining AdaBoost. In B. Schölkopf, Z. Luo, and V. Vovk (eds.), *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, 37–52. Springer Berlin Heidelberg, Berlin, Heidelberg. doi:10.1007/978-3-642-41136-6_5.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, Ltd. doi:10.1002/9780470316849.

Silverman, B.W. (2018). *Density Estimation for Statistics and Data Analysis*. Routledge. doi:10.1201/9781315140919.

Terrell, G.R. and Scott, D.W. (1992). Variable Kernel Density Estimation. *The Annals of Statistics*, 20(3), 1236–1265.

Tokdar, S.T. and Kass, R.E. (2010). Importance Sampling: A Review. *WIREs Computational Statistics*, 2(1), 54–60. doi:10.1002/wics.56.

Toscani, G. (2017). Score Functions, Generalized Relative Fisher Information and Applications. *Ricerche di Matematica*, 66(1), 15–26. doi:10.1007/s11587-016-0281-0.

Zhang, K. and Luo, M. (2015). Outlier-Robust Extreme Learning Machine for Regression Problems. *Neurocomputing*, 151, 1519–1527. doi:10.1016/j.neucom.2014.09.022.