Spline-Based Density Estimation Minimizing Fisher Information

Dominik Prossel and **Uwe D. Hanebeck**
Intelligent Sensor-Actuator-Systems Laboratory (ISAS)
Institute for Anthropomatics and Robotics
Karlsruhe Institute of Technology (KIT), Germany
dominik.prossel@kit.edu, uwe.hanebeck@kit.edu

***Abstract*—The construction of a continuous probability density function (pdf) that fits a set of samples is a frequently occurring task in statistics. This is an inherently underdetermined problem, that can only be solved by making some assumptions about the samples or the distribution to be estimated. This paper proposes a density estimation method based on the premise that each sample represents the same amount of probability mass of the underlying density. The estimated pdf is parameterized as the square of a polynomial spline, which makes further processing of the estimated density very efficient. This pdf is inherently non-negative, ensuring a monotone cumulative distribution function, which makes it easy to generate samples from it through inverse transform sampling. Furthermore, it is cheap to evaluate and easy to integrate, making moment calculations fast. To find the coefficients of the polynomials that make up the spline, an optimization problem is derived. The Fisher information is used as a regularizer in this problem to select the solution that contains the least amount of information. The method is shown to work on samples from a variety of different one-dimensional probability distributions.**

***Index Terms*—Density estimation, splines, polynomials, Fisher information, deterministic sampling.**

I. Introduction

Random variables and their associated probability distributions are an elemental building block in robotics to model uncertainty in sensor measurements and mismatch between ideal and real-world models. Probability distributions can be uniquely described by their probability density function (pdf) or cumulative density function (cdf).

Real-world systems are often too complicated to analytically derive the probability distribution of their output values given a distribution of inputs, but a single input value can be propagated through the system to yield a single output value. In these cases, the input distribution can be replaced by samples drawn from it. The samples are then put through the system to get a sample representation of the output distribution. This kind of processing is used for example in particle filters, stochastic simulation, and Bayesian neural networks.

A very common method is to draw samples from a given input distribution at random. This requires a rather large amount of samples to accurately match the distribution. In contrast, deterministic sampling tries to find an optimal set of samples to represent the distribution and needs fewer samples than random sampling to achieve the same accuracy [1], [2].

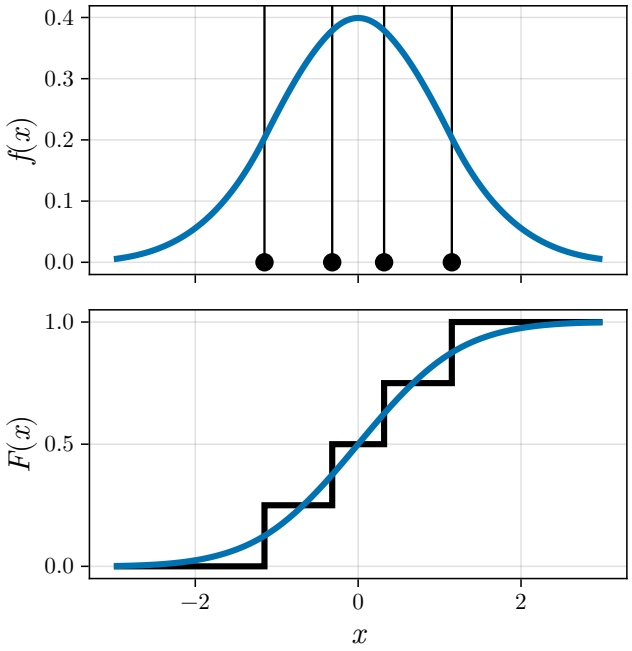

Fig. 1: Upper: The pdf (blue) estimated from four samples (black) with the proposed method. Lower: The according estimated cdf (blue) and empirical cdf (black).

If the cdf $F(x)$ of a one-dimensional probability distribution with pdf $f(x)$ is known, inverse transform sampling is a straightforward way to sample from this distribution. To draw N samples, the first step is to generate N samples from a uniform distribution between 0 and 1. These are then propagated through the inverse cdf $F^{-1}(x)$ of the target distribution, which transforms them into samples that follow this distribution. While this method can be used to generate random samples, it is especially easy to use for deterministic sampling, as the uniform samples can be given in closed form in this case, yielding the final samples ξ_i

$$\xi_i = F^{-1}\big((2i-1)/2N\big) \ , \tag{1}$$

for $i = 1, \ldots, N$.

Density estimation is essentially the inverse operation of sampling from a distribution. It is an inherently underdetermined problem, as samples do not contain enough information

to reconstruct a continuous distribution uniquely. This becomes especially noticeable when working with small numbers of samples, particularly when they are random samples. This means that some assumptions either about the samples or about the underlying density have to be made to select a single solution from all the possible distributions that a set of samples could originate from. When selecting specific restrictions to impose on the estimated density, it is important to keep in mind why the density is estimated in the first place and what further processing steps are planned based on the density estimate. For example, it might be a good idea to use a Gaussian mixture density to visualize data, but generating deterministic samples from this mixture is not straightforward.

This paper introduces a novel method for density estimation based on polynomial splines. After a brief review of the existing literature, the problem of density estimation is framed as the inverse operation to inverse transform sampling. This leads to a set of constraints that the cdf of the estimated density should fulfill.

Squared polynomial splines are presented as a suitable form of parametrization of the estimated density. Their main advantages compared to other representations like Gaussian mixtures are, that they can be integrated in closed form and the cdf of the estimated distribution is comparatively easy to invert.

To select the smoothest solution fulfilling all constraints the Fisher information is added as an objective function. By working with squared functions it is possible to use a reformulation of the Fisher information, that considerably simplifies its calculation.

Finally, the proposed method is applied to some examples from different distributions to demonstrate possible results.

II. State of the Art

A. Parameter Estimation

If the family of distributions, which the samples are drawn from, is either known or assumed to be known, the free parameters of the density function need to be estimated. These parameters could be the mean and variance of a Gaussian pdf or the rate of decay of an exponential distribution. This is called parametric density estimation, and there is only a relatively small number of parameters to be estimated. The classic maximum likelihood estimator finds the parameters that maximize the likelihood that the samples were drawn from the according distribution. It is particularly straightforward to apply if the samples are independent and identically distributed. A different approach to this is called maximum spacing estimation [3]. It makes use of the same notion as inverse transform sampling, that a cdf should map the samples drawn from its distribution to uniform samples. This can be achieved by maximizing the geometric mean of the distances between values of the cdf at consecutive samples.

B. Kernel Methods

Kernel methods for density estimation have been independently introduced more than five decades ago by Parzen [4] and Rosenblatt [5] and are still widely in use today. The estimated density is represented as a sum of kernels centered at the locations of the random samples. Gaussian kernels are a popular choice and many methods have been proposed to select the variance or bandwidth of the kernels to minimize the estimation error. A popular heuristic for this is Silverman's rule of thumb [6], which is relatively easy to calculate. But as it assumes an underlying Gaussian pdf, it can easily break down when the data is multimodal. Other bandwidth selection methods are based on error metrics and minimize their expected value under different premises [7].

C. Spline-based Methods

Spline-based methods approximate the pdf by a sum of piecewise functions, each one defined on an interval. These intervals may either be defined a priori as a grid or constructed from the sample positions, in which case the estimation of the density boils down to some kind of spline interpolation. Popular families of spline functions are polynomial and exponential functions. In [8] so-called exponential epi-splines are used, as they can match important distributions like Gaussians exactly. This property comes with the disadvantage that these splines are in general difficult to integrate. The spline is fitted to the samples using maximum likelihood estimation under some optional constraints to guide the optimization based on additional information about the density.

Huber dealt with the problem of finding the pdf that minimizes the Fisher information, given some values of the corresponding cdf [9]. He showed that there is a unique solution to this problem and gave a general form of the solution. The result is a "curious and rather non-trivial type of spline interpolation" [9].

D. Fisher Information as Regularizer

If, after applying all the constraints and premises to the density estimation problem one is willing to use, there are still some degrees of freedom left, an objective function can be introduced to select a unique solution. This was done in [10], where the Fisher information was used in conjunction with maximum likelihood estimation to fit a pdf to observations. They expressed the density as a sequence of orthogonal Hermite polynomials. The Fisher information was added as a roughness penalty to the maximum likelihood estimation, to ensure that the result would be as smooth as possible. The same notion of Fisher information as roughness penalty was used in [11], to find the smoothest possible Gaussian mixture density fulfilling a set of general constraints.

III. Problem Formulation

A. Assumptions on Samples

We consider the estimation of a pdf $f(x)$ from a set of $N > 1$ one-dimensional samples ξ_i, $i = 1, \ldots, N$ of an unknown underlying probability distribution $\tilde{f}(x)$. The samples are expected to be distinct from each other and sorted in ascending order

$$-\infty < \xi_1 < \xi_2 < \cdots < \xi_{N-1} < \xi_N < \infty \ . \tag{2}$$

Additionally, each sample is interpreted as a Dirac delta function that represents an equal amount of the probability mass of the continuous underlying distribution $\tilde{f}(x)$. This leads to the assumption that the intervals defined by the locations of the samples should also contain the same amount of mass of $\tilde{f}(x)$

$$\int_{x_i}^{x_{i+1}} \tilde{f}(x)\,dx = \frac{1}{N} \quad i = 1, \ldots, N-1 \tag{3}$$

and the two border intervals should contain half the mass

$$\int_{-\infty}^{x_1} \tilde{f}(x)\,dx = \int_{x_N}^{\infty} \tilde{f}(x)\,dx = \frac{1}{2N}. \tag{4}$$

In terms of the cdf $\tilde{F}(x)$ this is equivalent to

$$\tilde{F}(x_i) = \int_{-\infty}^{x_i} \tilde{f}(x)\,dx = \frac{2i-1}{2N} \quad i = 1, \ldots, N, \tag{5}$$

which gives the formula for inverse transform sampling when solved for x_i. These assumptions cause the estimated cdf to map the samples to equidistant values on the ordinate. This is similar to the procedure in maximum spacing estimation where the distances between the cdf values of consecutive samples are maximized to achieve the mapping to a uniform distribution.

Regardless of the method used to generate the samples, it is assumed that they approximately satisfy (3) to (5). Samples obtained from inverse transform sampling are drawn to do this by design. Unfortunately, this is not true for random samples, which can lead to some unexpected results, when using them as input to our proposed method. For example, the pdf estimated from a set of random Gaussian samples approaches a Gaussian density only with a relatively large number of samples. To mitigate these effects from randomness, we propose to smooth out random samples before applying our method for example by replacing them with fewer samples, that encode the same distribution. A fast way to do this is to replace the original set of samples with its q_i-quantiles where $q_i = (2i-1)/(2N)$. A more accurate, but computationally more demanding alternative would be to use Dirac mixture reduction techniques based on Localized Cumulative Distributions (LCDs) [12].

B. Density Representation

We seek to parameterize the estimated pdf $f(x)$ in a way that lends itself to further processing like moment calculation or sampling, while still being able to approximate arbitrary density functions accurately.

These requirements are fulfilled very well by polynomial splines. A polynomial spline is defined as

$$S(x) = \sum_{i=1}^{M-1} s_i(x) \tag{6}$$

with functions $s_i(x)$ defined on consecutive intervals between knots $m_1, \ldots, m_{M+1}$

$$s_i(x) = \begin{cases} p_i(x) & \text{if } m_i < x \leq m_{i+1} \\ 0 & \text{else} \end{cases} \quad i = 1, \ldots, M \tag{7}$$

and polynomials of fixed degree d with coefficients $c_{i,k}$

$$p_i(x) = \sum_{k=0}^{d} c_{i,k}\, x^k. \tag{8}$$

In contrast to exponential epi-splines [8] or the spline functions proposed in [9], polynomial splines are straightforward to integrate, enabling the efficient calculation of the cdf and moments of the distribution. Also, polynomial root finding is a well-studied area of mathematics making fast inverse transform sampling possible, which is more complex for mixture models as used in [11]. Lastly, the number of degrees of freedom and with that the expressiveness of the representation can be adjusted by adjusting the degree of the used polynomials.

However, one of the necessary conditions of a pdf is that it needs to be non-negative. This means that each polynomial $p_i^{(d)}(x)$ needs to be non-negative on its relevant interval. For polynomials of low degree, some constraints of the coefficients can be found to ensure this [13]–[16]. To avoid these additional constraints in the final optimization problem and to be able to easily change the degree of the used polynomials without also adapting the constraints, we chose to work with squared polynomials

$$p_i(x) = r_i(x)^2, \tag{9}$$

that are completely defined by the polynomials $r_i(x)$. This has the advantage that the resulting spline is non-negative without needing to enforce further constraints, at the cost of sacrificing some degrees of freedom of the original polynomials.

We define the knots for the spline function based on the given samples (2). First, two additional artificial points $x_0 < x_1$ and $x_{N+1} > x_N$ are added to the sample set. These two points serve as outer bounds for the support of the spline. As all polynomials tend to $\pm\infty$ for $x \to \pm\infty$, these bounds are required to ensure that the integral over the spline is finite. We can then set

$$m_i = x_{i-1} \quad i = 1, \ldots, N+2 \tag{10}$$

to get $M = N+2$ knots and the corresponding $N+1$ intervals and polynomials defined on these intervals.

To find the desired pdf $f(x)$ given the samples x_i, the spline function is fit to the samples fulfilling (5) and making sure that the pdf and its first derivative are continuous. This gives a system of nonlinear equations consisting of

$$\int_{x_1}^{x_2} p_i(x)\,dx = \frac{1}{2N}, \tag{11}$$

$$\int_{x_{i-1}}^{x_{i+1}} p_i(x)\,dx = \frac{1}{N} \quad \text{for } i = 2, \ldots, N, \tag{12}$$

$$\int_{x_N}^{x_{N+1}} p_i(x)\,dx = \frac{1}{2N}, \tag{13}$$

and the continuity constraints

$$p_i(x_i) = p_{i+1}(x_i) \quad \forall i \in 1, \ldots, N, \tag{14}$$

$$p_i'(x_i) = p_{i+1}'(x_i) \quad \forall i \in 1, \ldots, N. \tag{15}$$

These constraints on $p_i(x)$ directly result in a set of constraints on $r_i(x)$ with $3N+1$ equations to be solved for $(d+1)(N+1)$ coefficients of all the spline polynomials. This means that polynomials of degree two or higher have to be used for $r_i(x)$ to have enough degrees of freedom to get feasible solutions, which results in squared polynomials of at least degree four. The system then has more coefficients than equations making it underdetermined. This is addressed by introducing the Fisher information as a regularizer to select the least informative solution from the feasible set. This solution contains the minimum amount of additional information compared to the original samples, while simultaneously fulfilling the given continuity constraints.

C. Fisher Information

Fisher information is commonly used as a roughness measure for pdfs in the literature [10], [11], [17]. It is based on the information-theoretic notion that pdfs become more informative the more peaks they have, and the more pronounced these are. It is commonly defined as

$$I_F[f(x)] = \int_{-1}^{1} \frac{f'(x)^2}{f(x)}\,dx. \tag{16}$$

This integral typically is not solvable in closed form. By setting $g(x) = \sqrt{f(x)}$ with $f(x) \neq 0$ and exploiting

$$g'(x) = \sqrt{f(x)}' = \frac{f'(x)}{2\sqrt{f(x)}} \tag{17}$$

the Fisher information can be reformulated as

$$I_F[f(x)] = I_F[g(x)^2] = 4\int_{-1}^{1} g'(x)^2\,dx. \tag{18}$$

This reformulation eliminates the division by $f(x)$ from the integral but requires the square root of the function. As we plug the proposed squared polynomial spline into the Fisher information we discover another reason that we chose to work with squares of the polynomials in (9), as this form is perfectly suited to be used with (18).

D. Final Optimization Problem

We model the desired pdf $f(x)$ as the square of a polynomial spline, that is itself a polynomial spline of higher degree. The unsquared spline $r(x)$ is parameterized by the vector of coefficients $C = [c_{1;0}, \ldots, c_{1;d}, \ldots, c_{N+1;0}, \ldots, c_{N+1;d}]^\top$ of its constituent polynomials $s_i(x)$. We now find the vector of coefficients C, that minimizes the Fisher information (18), while fulfilling the constraints (11) to (15) imposed on the squared spline. The optimal coefficients of $r(x)$ are obtained by solving the optimization problem

$$C = \arg\min_C I_F[r(x;C)^2] \tag{19}$$

subject to (11) to (15). The corresponding coefficients of the squared polynomials $p_i(x)$ can then easily be calculated from these, yielding the final pdf $f(x;C)$.

When using unsquared polynomials of degree two the constraints admit only one solution when the two border points x_0 and x_{N+1} are fixed, as the number of parameters is exactly equal to the number of equations. If the borders are part of the optimization parameters or a higher-order polynomial is used, the system is underdetermined and the Fisher information selects the unique least informative solution.

It is possible to introduce additional constraints to the optimization problem if more details about the shape of the unknown underlying density are known. This could for example be function values or derivative information. In [8] this additional information is called soft information.

IV. Examples

The proposed method was implemented in Julia using the JuMP framework [18]. The type of results that were produced are visualized in Fig. 1. The empirical cdf given by the samples is smoothed by the squared polynomial spline, intersecting the cdf at the values specified in the constraints (11) to (13).

To show the usefulness of the proposed density estimation method, it was tested on samples from four different probability distributions. The samples were generated from these distributions with inverse transform sampling based on deterministic uniform samples. It is important to note that the only inputs to optimization are the positions of the samples. The function that was used to generate the samples is completely unknown to the optimizer. For all experiments, polynomial splines with $d = 3$ were used and the border points x_0 and x_{N+1} were part of the optimization parameters and not fixed beforehand.

The first two distributions showcased are a Gaussian distribution with zero mean and unit variance (Fig. 2) as well as a Laplace distribution with zero mean and scale of one (Fig. 3). Even for a low number of samples the estimated pdf matches the underlying density closely and for $N = 10$ samples there is almost no difference between them. For one, this shows the power of deterministic samples to accurately represent densities with few samples, but also that our method makes good use of this power by enforcing (5).

Next, we have a look at a multimodal pdf as it may arise in the filter step of a Bayesian filter. The example in Fig. 4 results from multiplying a Gaussian prior with the likelihood associated with a quadratic measurement equation. Our method manages to give good estimates for the density that produced these samples. As the least informative estimate given by our method is quite different from the actual underlying density for 5 samples, it is clear this number of samples is insufficient to capture all the information of the underlying density. With 10 samples, however, the original density is matched much more closely as more information is available.

As a last distribution, we have a look at the Gamma distribution with shape parameter 2 and scale 1, see Fig. 5. The special feature of this distribution is, that it is only defined for positive numbers. We built this information into the optimization by fixing the left border point $x_0 = 0.0$. While the falling flank of the density is estimated relatively accurately, there are some deviations on the rising flank and around the maximum of the distribution. Through numerical experiments, we could validate, that these are not errors of the

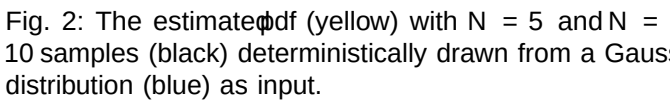

Fig. 2: The estimated pdf (yellow) with $N = 5$ and $N = 10$ samples (black) deterministically drawn from a Gaussian distribution (blue) as input.

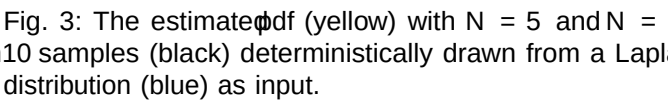

Fig. 3: The estimated pdf (yellow) with $N = 5$ and $N = 10$ samples (black) deterministically drawn from a Laplace distribution (blue) as input.

proposed method. It seems, that these are the functions with the lowest Fisher information satisfying the set of constraints. As the code was not optimized for speed and the time to solve the optimization problem seems to depend heavily on the exact sample positions, the initial guess, and the degree of the polynomials used, no detailed benchmarking regarding the execution time was performed. All example problems discussed in the paper terminated after about 100 ms on one core of an AMD Ryzen 7 PRO 4750U notebook CPU.

V. Conclusion

We introduced a new method to estimate a pdf from a set of samples. The density is represented by a squared polynomial spline, which facilitates further processing of the estimated pdf. Specifically, it is guaranteed to be non-negative, fast to evaluate, and easy to integrate. The fitting is done by imposing constraints on the cdf of the solution and selecting the least informative solution based on Fisher information. The method can be applied to deterministic and, with suitable preprocessing, random samples. The results show that even polynomials of relatively low degree can approximate the optimal solution very well.

Using different objective functions and additional constraints can be used to guide the optimization to different solutions. The impact that specific objectives or constraints have on the solution is still to be investigated. We also hope to extend the proposed method to two or more dimensions, though this is a challenging problem to solve. The key component for it is to find a high-dimensional analogue to the intervals between samples.

VI. Acknowledgement

This work has been supported by the Carl Zeiss Foundation under the JuBot project.

References

[1] R. E. Caflisch, "Monte Carlo and quasi-Monte Carlo methods," *Acta Numerica*, vol. 7, pp. 1–49, 1998.

[2] H. Niederreiter and A. Winterhof, "Quasi-Monte Carlo Methods," in *Applied Number Theory*, H. Niederreiter and A. Winterhof, Eds., Cham: Springer International Publishing, 2015, pp. 185–306.

[3] B. Ranneby, "The Maximum Spacing Method. An Estimation Method Related to the Maximum Likelihood Method," *Scandinavian Journal of Statistics*, vol. 11, no. 2, pp. 93–112, 1984.

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

[5] M. Rosenblatt, "Remarks on Some Nonparametric Estimates of a Density Function," *The Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 832–837, Sep. 1956.