SIAM J. NUMER. ANAL.
Vol. 47, No. 6, pp. 4759–4792

A FINITE ELEMENT METHOD FOR DENSITY ESTIMATION WITH GAUSSIAN PROCESS PRIORS*

MICHAEL GRIEBEL† AND MARKUS HEGLAND‡

Abstract. A variational problem characterizing the density estimator defined by the maximum a posteriori method with Gaussian process priors is derived. It is shown that this problem is well posed and can be solved with Newton's method. Numerically, the solution is approximated by a Galerkin/finite element method with piecewise multilinear functions on uniform grids. Error bounds for this method are given and numerical experiments are performed for one-, two-, and three-dimensional examples.

Key words. density estimation, finite elements, Galerkin, Newton

AMS subject classifications. Primary, 65C60; Secondary, 65K10

DOI. 10.1137/080736478

1. Introduction.

> All modern theories of statistical inference take as their starting point the idea of the probability density function of the observations.
> *Emanuel Parzen, "An Approach to Time Series Analysis"* [1].

This statement by Parzen points to the fundamental importance of densities in statistical reasoning which has not diminished at all in recent years. In fact, the methods of data mining and unsupervised learning as well as the more traditional regression and classification are all approaches to extract information about densities using observed data.

Probability densities are directly applied in machine learning and data mining. There, one is interested in finding a classifier of objects (characterized by feature vectors) into one of two classes. If one knows the probability of each class for a given feature, then one can determine the classifier which minimizes a given loss function. While it is usually not feasible to determine the probability of each class for every given feature, one can often find good approximations of the density of the features for each class. The probability of the class given the feature is then provided by Bayes' theorem. An example of this approach is the "Naive Bayes classifier" [2]. Probability densities may also be used as substitutes for the data set in the estimation of features of the data or in the determination of the probability of subsets. The advantage of this approach is computational in that it does allow one to obtain the expected function values or integrals without having to go through the full data set which may be very large.

Here we consider specifically Gaussian process priors. While the general variational formulation considered in the next section may also be applied to other types of priors, a major computational advantage of Gaussian process priors results from their intimate connection with reproducing kernel Hilbert spaces or Cameron–Martin

*Received by the editors September 26, 2008; accepted for publication (in revised form) December 10, 2009; published electronically February 24, 2010. This work was partially supported by the SFB 611 "Singular Phenomena and Scaling in Mathematical Models" at the University of Bonn.
http://www.siam.org/journals/sinum/47-6/73647.html
†Institute for Numerical Simulation, University of Bonn, 53115 Bonn, Germany (griebel@ins.uni-bonn.de).
‡Mathematical Sciences Institute (MSI), The Australian National University, Canberra, ACT 0200, Australia (markus.hegland@anu.edu.au).

spaces. In this paper we investigate the well posedness and the numerical solution of the variational problem underlying the maximum a posteriori (MAP) method for density estimation. The density estimation problem consists of finding a probability density $f(t)$ from a set of data points $t_1, \ldots, t_n$. We first show that MAP estimators are penalized maximum likelihood estimators using the Cameron–Martin theory of stochastic processes. We then review existence, uniqueness, and the dual problem. We foremost show that the problem satisfies all the necessary conditions for the application of a finite element Newton method using the preconditioned conjugate gradient method for the solution of the occurring finite-dimensional linear subsystems.

The remainder of this paper is organized as follows: In section 2 we derive the penalized likelihood functional from the MAP method and review the concept of modes in infinite-dimensional spaces. In section 3 we use variational calculus to show existence and uniqueness of the estimator. Furthermore, we derive the dual equations and give a connection with maximum entropy. In section 4 we show the convergence of a Newton method and Galerkin approximation in general. In section 5 we discuss approximation spaces, approximation properties, and convergence rates. In section 6 we give the results of numerical experiments with the new method for one-, two-, and three-dimensional problems.

2. The MAP framework for density estimation. In this section we review the basic framework for density estimation used in the later sections. A key component is that the unknown probability (or the log thereof) is modeled by a stochastic process. In the case of Gaussian processes, this approach was pioneered by Parzen in a paper [1] on signal processing. This approach has then led to what is now called "the penalized maximum likelihood method" of density estimation which was established by Good and Gaskins in the statistical literature [3]. The approach was further developed by Leonard in [4] using ideas of Parzen. While Parzen does acknowledge the importance of stochastic modeling, he does not relate the models to the Cameron–Martin theory which was developed in the 1940s [5]. This connection of the maximum a posteriori method with stochastic modeling can be found in a book by Bogachev on Gaussian processes [6] and was applied to the MAP method in a recent paper [7].

In the following we review this approach and, in particular, derive the variational problem (the "penalized maximum likelihood problem") from the maximum a posteriori method in the case of stochastic process priors using tools from stochastic analysis. In particular, we will use the Radon–Nikodym theorem, the Cameron–Martin spaces, and the Cameron–Martin derivatives.

Let $T \subset \mathbb{R}^d$ be a domain, typically the unit cube. The unknown probability density $f_u(t)$ on T is assumed to be of the form

$$f_u(t) = \exp(u(t) - \gamma(u)), \tag{1}$$

where u is an unknown function on T and $\gamma(\cdot)$ is the functional given by

$$\gamma(u) = \log \int_T \exp(u(t))\, dt. \tag{2}$$

Combining these two equations one gets

$$f_u(t) = \frac{e^{u(t)}}{\int_T e^{u(t)}\, dt}. \tag{3}$$

While f_u is uniquely determined by u and (3), the converse does not hold, as for all constants c the functions $u + c$ lead to the same probability density $f_u = f_{u+c}$.[1] The functions f defined as in (3) are probability densities; i.e., they are nonnegative and integrate to one.[2]

In many statistical applications, and, indeed, also in numerical computations, the function $u(t)$ is selected from a finite-dimensional function space, such that for some vector $\theta \in \mathbb{R}^m$ and a function $\phi : T \to \mathbb{R}^m$ one has

$$u(t) = \theta^T \phi(t) = \sum_{i=1}^{m} \theta_i \phi_i(t).$$

In statistics, $\phi(t)$ is called a sufficient statistic, γ is the log partition function, and the family of densities f_u thus defined is called the exponential family. Many popular probability distributions are elements of this family. A prominent member is the normal distribution where $m = 2$ and the basis functions are

$$\phi_1(t) = t, \quad \phi_2(t) = t^2$$

in $d = 1$ dimension. In this case the coefficients turn out to be

$$\theta_1 = \frac{\mu}{\sigma^2}, \quad \theta_2 = -\frac{1}{2\sigma^2},$$

where μ and σ^2 denote the mean and variance of the normal distribution.

The determination of θ, and hence $u(t)$, from data is often done by the maximum likelihood method (see, e.g., [8, 9]). There, typically the negative log likelihood function

$$l(u) = -\sum_{i=1}^{n} u(t_i) + n\,\gamma(u) \tag{4}$$

of the data $t_1, \ldots, t_n$ is minimized. This approach has been introduced by Fisher in [10]. It is known that, under some weak conditions, the minimizer of $l(u)$ is asymptotically (in n) unbiased, achieves an optimal error bound (the Cramer/Rao bound), and is asymptotically normal. This method is closely related to the maximum entropy method.[3]

The maximum likelihood method breaks down when the space of considered functions u is infinite dimensional as in this case the problem becomes ill posed. A popular way to deal with ill posedness is to use a penalized maximum likelihood approach where instead of the negative log likelihood $l(u)$ one minimizes a penalized negative log likelihood $j(u)$ which is of the form

$$j(u) = \frac{\alpha}{2}\|Lu\|^2 + \frac{1}{n}l(u) = \frac{\alpha}{2}\|Lu\|^2 + \log \int_T \exp(u(t))\,dt - \frac{1}{n}\sum_{i=1}^{n} u(t_i) \tag{5}$$

with some operator L and L_2-norm $\|\cdot\|$. Here $\alpha \geq 0$ is a regularization parameter which balances the fit to the data with the regularity of the solution. This functional

[1] We assume here that the Lebesgue measure of T is finite.

[2] Note that only probability densities which do not have zero values can be modeled in this way. By adjusting T to be the support of f_u one gets around this problem.

[3] Here the entropy of f is the expectation of the random variable defined by the function $u(t)$ with respect to the probability density $f(t)$.

$j(u)$ shall be derived from the maximum a posteriori method in the following. The discussion is based on [7].

For simplicity, we assumed in the following that the prior does have a zero mean. This is the simplest and most widely used case in practice. If required, the inclusion of a nonzero mean can also be included with no need to change the theory.

2.1. Stochastic processes, their measures, and their modes. In principle, the maximum a posteriori method is simple: First one assumes a probability distribution over a class of models which represents the prior knowledge of the problem. Then the likelihood of the data is interpreted as a conditional probability of the data given a model. From this, one can derive the posterior probability which is the probability of the model given the data using Bayes' theorem.

While a full Bayesian analysis then proceeds by further analyzing the posterior distribution, the MAP method merely determines the maximum or main mode of the posterior. The approach works well and is used, e.g., to determine for every t the most likely $u(t)$ given the data. But in the case of infinite-dimensional function spaces for u this approach has a problem with the definition of the mode due to the fact that the Lebesgue measure does not exist here [11]. This problem, while sometimes forgotten, was well known in the original statistical literature by Parzen and Leonard mentioned previously. However, it is not discussed in the newer literature based on kernels, and, in particular, not in the literature on machine learning using Gaussian process priors; see, e.g., [12, 13].

According to the textbook [14] on stochastic processes by Adler, there are two "virtually distinct" but equivalent approaches to define random fields or stochastic processes. The first approach is *measure theoretic* and a stochastic process is defined as a measure μ on a set $X \subset \mathbb{R}^T$ of functions $T \to \mathbb{R}$. In the second, *probabilistic* approach a stochastic process is defined as a collection of (real) random variables $U(t)$ parameterized by $t \in T$. While the second approach is dominant in modeling and simulation, we here use the first one to derive the variational characterization of the MAP estimator. For the Gaussian case a thorough discussion of this approach can be found in the book [6] by Bogachev.

To illustrate the measure theoretic approach for the reader who might be less familiar with stochastic processes, we consider the case of a Gaussian process which can be described by a series expansion:

$$U(t) = \sum_{n=1}^{\infty} Z_n M e_n(t).$$

In the following, random variables and stochastic processes will be denoted by capital letters, e.g., $U(t)$, while samples and (nonrandom) functions will be denoted by lowercase letters, e.g., $u(t)$. The expansion of $U(t)$ is closely related to the Karhunen–Loeve expansion. Here the Z_n are pairwise independent standard Gaussian random variables and the e_n form a Hilbert basis of a reproducing kernel Hilbert space $H \subset \mathbb{R}^T$. This Hilbert space is associated with the stochastic process and is usually called *Cameron–Martin* space. The linear operator M is bounded on H. In the case discussed later, H is a tensor product of Sobolev spaces. Clearly, the truncated series

$$U_m(t) = \sum_{n=1}^{m} Z_n M e_n(t)$$

defines a collection of normally distributed random variables indexed by t and at the same time a Gaussian measure (or probability distribution) on the m-dimensional

linear space of functions spanned by the $e_1, \ldots, e_m$. As a sample path depends on the actual sample values z_n of the random variables Z_n, the measure defining the stochastic process $U_m(t)$ is thus the standard m-dimensional normal distribution $\nu(z_1, \ldots, z_m)$. One can see [6] that these measures converge weakly as measures on $\mathbb{R}^T$ with $m \to \infty$. More generally, for a general choice of basis b_n of V_m, a stochastic process is approximated in an m-dimensional subspace V_m of H by an expansion of the form

$$U_m(t) = \sum_{n=1}^{m} Y_n b_n(t), \tag{6}$$

where the random vector $(Y_1, \ldots, Y_m)$ is Gaussian with a joint probability density $\rho_0(y_1, \ldots, y_m)$. Recall here that we denote random variables by capitals Y_i and numbers by lowercase letters y_i.

In the following we assume that the Cameron–Martin space H is compactly embedded in the Banach space $C(T)$ of continuous functions and that the triple $(\mathcal{J}, H, C(T))$ forms an *abstract Wiener space* [6, p. 137]. As usual, the norm in $C(T)$ is $\|u\|_\infty = \sup_{t\in T} |u(t)|$ and $\mathcal{J}$ denotes the embedding of H in $C(T)$. It is known that in this case the process $U(t)$ defines a Gaussian measure λ on $C(T)$; see, e.g., [6, Thm. 3.9.5].

The posterior measures which occur in the MAP method are in general not Gaussian. Such measures take the form

$$\mu(A) = \int_A \rho(u)\, d\lambda(u), \quad A \subset C(T), \text{ measurable;} \tag{7}$$

i.e., $\rho = d\mu/d\lambda$ is the Radon–Nikodým derivative or density of μ with respect to λ.[4] In MAP λ is the prior (in our case Gaussian) and ρ is the likelihood of the data. As the likelihood $\rho(u)$ depends only on a finite number of function values at the data points and on the log partition function $\gamma(u)$ introduced previously, one can show that ρ is a continous nonlinear functional on $C(T)$; see also the discussion in section 3.1. It then follows that ρ is in $L_1(\lambda)$.

For the following definition of the mode of μ one needs the *shifted measure* λ_v which is given by

$$\lambda_v(A) = \lambda(v + A), \quad A \subset C(T), \text{ measurable.}$$

If v is such that λ_v is absolutely continuous with respect to λ (which we denote by $\lambda_v \ll \lambda$), then the Radon–Nikodým derivative or $d\lambda_v/d\lambda \in L_1(\mu)$ exists.

For the case of Gaussian λ it is known [6] that $\lambda_v \ll \lambda$ if and only if v is an element of the Cameron–Martin space H. In this case one has an explicit representation for the derivative:

$$\frac{d\lambda_v}{d\lambda}(w) = \exp\left(-\langle \psi_v, w\rangle - \frac{1}{2}\|v\|_H^2\right), \quad v \in H,\ w \in C(T). \tag{8}$$

Equation (8) is called the *Cameron–Martin formula*. The linear functional ψ_v in this formula is such that $w(t) = \langle \psi_v, w\rangle$ if $v = k_t$ is the reproducing kernel of H at t. Consequently, for $w \in H$ one has

$$\frac{d\lambda_v}{d\lambda}(w) = \exp\left(-(v, w)_H - \frac{1}{2}\|v\|_H^2\right), \quad v, w \in H.$$

[4] The "density" ρ is defined on the function space $C(T)$ and should not be confused with the "density" f_u defined on T.

It follows that $\frac{d\lambda_v}{d\lambda}(w)$ is a continuous nonlinear functional on H which is in $L_1(\lambda)$. However, it is in general not a continuous functional on $C(T)$ except when $v = \sum_{j=1}^m c_j k_{t_j}$ holds for some finite m. In these special cases one has

$$\langle \psi_v, w \rangle = \sum_{j=1}^m c_j w(t_j),$$

which is continuous in $C(T)$ and the continuity of $\frac{d\lambda_v}{d\lambda}(w)$ as a function of w follows.

The mode of a stochastic process with measure μ of the form given in (7) is then defined as follows.

DEFINITION 1. *Let μ and λ be measures on $C(T)$ and let the Radon–Nikodým derivative $\rho = d\mu/d\lambda$ be a continuous linear functional on $C(T)$. Furthermore, let $u \in C(T)$.*

(a) *We call an element $v \in C(T)$ admissible if $\|v\|_\infty < \epsilon$ for some $\epsilon > 0$, if $\lambda_v \ll \lambda$, and if $d\lambda_v/d\lambda(w)$ is continuous at $w = u$.*

(b) *u is a mode of the measure μ if for all admissible v one has*

$$\rho(u) \geq \frac{d\lambda_v}{d\lambda}(u)\, \rho(u+v). \tag{9}$$

Note that the set of admissible v is not empty for λ, a Gauss measure on $C(T)$ where the Cameron–Martin space has a reproducing kernel k_t, as for any $v = \sum_{j=1}^m c_j k_{t_j}$ one has $\lambda_v \ll \lambda$ and $\frac{d\lambda_v}{d\lambda}(w)$ is continuous for all w. Note, however, that in this definition we do not assume that λ is necessarily Gaussian, even though in the application it will be.

This definition generalizes the usual finite-dimensional definition of a mode, where λ is the Lebesgue or Haar measure; thus $d\lambda_v/d\lambda = 1$ and ρ is the ordinary density. In the case where $\rho = 1$, the condition (9) becomes $\frac{d\lambda_v}{d\lambda}(u) \leq 1$. Furthermore one has the following properties.

PROPOSITION 1. *Let u be a mode of a measure μ with $\mu(A) = \int_A \rho(v)\, d\lambda(v)$; see (7).*

1. *u does not depend on the particular choice of λ and ρ.*
2. *There exists a $\delta > 0$ such that for all admissible v one has*

$$\mu(A) \geq \mu(v + A)$$

for all measurable $A \subset \{w \mid \|w - u\|_\infty < \delta\}$.

Proof. For the first claim assume that $\kappa \ll \lambda$ is a measure such that for some $\psi \in L_1(\kappa)$ one has

$$\mu(A) = \int_A \psi(w) d\kappa(u).$$

As $\kappa \ll \lambda$ one can then show that

$$\rho(w) = \psi(w) \frac{d\kappa}{d\lambda}(w).$$

Combining this with the defining condition for the mode, one then obtains

$$\frac{d\kappa}{d\lambda}(u)\, \psi(u) \leq \frac{d\lambda_v}{d\lambda}(u) \frac{d\kappa}{d\lambda}(u+v)\, \psi(u+v).$$

As $\frac{d\kappa}{d\lambda}(u+v) = \frac{d\kappa_v}{d\lambda_v}(u)$ and

$$\frac{d\kappa_v}{d\kappa} = \frac{\frac{d\kappa_v}{d\lambda_v}\frac{d\lambda_v}{d\lambda}}{\frac{d\kappa}{d\lambda}},$$

one gets the desired result.

For the second claim observe that the function $\zeta \in L_1(\lambda)$ (the set of λ-integrable functions on $C(T)$) with

$$\zeta(w) = \rho(w) - \frac{d\lambda_v}{d\lambda}(w)\,\rho(w+v)$$

is continuous as v is admissible. By definition $\epsilon := \zeta(u) > 0$ and by continuity $\zeta^{-1}([0,\epsilon)) = \tilde{N}(u)$ defines a neighborhood of u. Thus there exists a δ such that $\{w \mid \|w-u\|_\infty < \delta\} \subset \tilde{N}(u)$. It follows that $\zeta(w) > 0$ for all w with $\|w-u\|_\infty < \delta$. Consequently, for any measurable $A \subset \{w \mid \|w-u\|_\infty < \delta\}$ one has

$$\begin{aligned} 0 \le \int_A \zeta(w)\,d\lambda(w) &= \int_A \rho(w)\,d\lambda(w) - \int_A \frac{d\lambda_v}{d\lambda}(w)\,\rho(w+v)\,d\lambda(w) \\ &= \int_A \rho(w)\,d\lambda(w) - \int_{A+v} \rho(w)\,d\lambda(w) \\ &= \mu(A) - \mu(A+v) \end{aligned}$$

from which the desired result follows. □

A direct consequence of the Cameron–Martin formula (8) is the following.

COROLLARY 1. *Any minimal point $u \in H$ of the functional*

$$J(v) = \frac{1}{2}\|v\|_H^2 - \log(\rho(v))$$

is a mode of μ.

Proof. Let u be a minimizer of J. Then $J(u) \le J(u+v)$ for all $v \in H$ in some neighborhood of u. Taking the exponential one gets

$$\rho(u) \le \exp((u,v)_H - \|v\|_H^2/2)\,\rho(u+v),$$

which by the Cameron–Martin formula shows that u is a mode. □

2.2. Density estimation with MAP. We now derive the posterior measure μ and the variational characterization of the MAP estimator. Recall that the density $f_u \ge 0$ to be estimated is of the form

$$f_u(t) = \exp(u(t) - \gamma(u)), \quad t \in T,$$

and γ is such that $\int_T f_u(t)\,dt = 1$. The likelihood of the data, given that the data points are pairwise independent, is

$$g(t_1,\dots,t_n \mid u) = \exp(u(t_1) + \cdots + u(t_n) - n\gamma(u)).$$

This likelihood is interpreted as the conditional probability of the data given the function u. The posterior measure is then defined by

$$\mu(A) = C\int_A g(t_1,\dots,t_n \mid u)\,d\lambda(u),$$

where $C = 1/\int_{\mathbb{R}^T} g(t_1, \ldots, t_n|h)\, d\lambda(u)$ and λ is the prior measure. We can now apply Corollary 1 with $\rho(u) = Cg(t_1, \ldots, t_n \mid u)$, and it follows that the mode u of the posterior measure μ minimizes J with

$$J(v) = \frac{1}{2}\|v\|_H^2 - \sum_{i=1}^{n} v(t_i) + n\gamma(v) + \log(C).$$

Notice that the minimizer of J does not depend on C. Furthermore, replace $\|\cdot\|_H$ by the equivalent $n\|\cdot\|_H$ and recall that the reproducing kernel of the original $\|\cdot\|_H$ is the covariance of the prior to get after division by n the following result.

PROPOSITION 2. *Given a Hilbert space H with norm $\|\cdot\|_H$ and let u be the minimizer of the functional*

$$j(v) = \frac{1}{2}\|v\|_H^2 - \frac{1}{n}\sum_{i=1}^{n} v(t_i) + \gamma(v)$$

with

$$\gamma(v) = \log \int_T \exp(v(t))\, dt.$$

Then $f_u(t) = \exp(u(t) - \gamma(u))$ is the MAP estimator of the density f from the data $t_1, \ldots, t_n$ and the Gaussian process prior for u with expectation zero and covariance $nk(t, s)$ where k is the reproducing kernel of H with respect to $\|\cdot\|_H$.

3. Properties of the functional $j(u)$ and the dual problem. Finding the minimum of the functional $j(u)$ is a problem of variational and convex analysis. Here, efficient techniques are available to solve such problems, at least approximately. A key role is played by the reproducing kernel Hilbert space H. In this setting, the minimization problem is well posed when a few extra assumptions hold.

3.1. Existence, uniqueness, and characterization of the minimum. The functional $j : H \to \mathbb{R}$ with

$$j(u) = \frac{1}{2}\|u\|_H^2 + \log \int_T \exp(u(t))\, dt - \frac{1}{n}\sum_{i=1}^{n} u(t_i), \quad u \in H, \tag{10}$$

has three terms: The first term is the squared norm of the space H which represents the prior and serves to regularize the problem, the second term corresponds to the log partition function in statistical mechanics, and the third term is the data term. The functional can be controlled by the first term as follows: We assume that H contains only continuous functions, i.e.,

$$H \subset C(T). \tag{11}$$

As H is a reproducing kernel Hilbert space, one has $u(t) = (k_t, u)_H$ and thus $|u(t)| \le \|k_t\|_H \|u\|_H$. We also assume that $\|k_t\|_H$ is bounded for $t \in T$ and introduce the embedding constant

$$C_H = \sup_{t \in T} \|k_t\|_H < \infty. \tag{12}$$

One then has

$$\|u\|_\infty \le C_H \|u\|_H.$$

Now we assume that T is normalized such that

$$\int_T dt = 1. \tag{13}$$

With this normalization, the estimates $\exp(-\|u\|_\infty) \le \exp(u(t)) \le \exp(\|u\|_\infty)$, the monotonicity of the logarithm, and the embedding constant, one finally obtains

$$\left|\log \int_T \exp(u(t))\, dt\right| \le \|u\|_\infty \le C_H \|u\|_H.$$

For the third term in the formula for j one applies the triangle inequality and gets

$$\left|\frac{1}{n}\sum_{i=1}^{n} u(t_i)\right| \le C_H \|u\|_H.$$

By inserting the bounds for the second and third terms into the definition of $j(u)$ one obtains

$$\frac{1}{2}(\|u\|_H - 2C_H)^2 - 2C_H^2 \le j(u) \le \frac{1}{2}(\|u\|_H + 2C_H)^2 - 2C_H^2. \tag{14}$$

Clearly, $-\infty < j(u) < \infty$ for all $u \in H$ and thus the functional j is *proper*. If for a sequence $u_1, u_2, \ldots$, the norm is unbounded, i.e., $\|u_n\|_H \to \infty$, then the values of the functional are also unbounded, $j(u_n) \to \infty$, and so j is *coercive*.

While the first and third terms of $j(u)$ are continuous, the second term is *lower semicontinuous* since

$$\log \int_T e^{u(t)+h(t)}\, dt - \log \int_T e^{u(t)}\, dt = \log \frac{\int e^{u(t)} e^{h(t)}\, dt}{\int e^{u(t)}\, dt} \le \|h\|_\infty \le C_H \|h\|_H,\ u, h \in H.$$

We now show that j is convex. The first term of j is quadratic, thus convex. For the second term we consider the Hölder inequality:

$$\int \exp(u_1(t))^\theta (\exp(u_2(t)))^{1-\theta}\, dt \le \left(\int \exp(u_1(t))\, dt\right)^\theta \left(\int \exp(u_2(t))\, dt\right)^{1-\theta}$$

for all $\theta \in [0, 1]$. By taking the logarithm and rearranging the terms one gets

$$\log \int \exp(\theta u_1(t) + (1-\theta) u_2(t))\, dt \le \theta \log \int \exp(u_1(t))\, dt + (1-\theta) \log \int \exp(u_2(t))\, dt,$$

which means that the second term is convex. As the third term is linear, one thus has that $j(u)$ is *strictly convex*.

We now have all the ingredients for the existence and uniqueness of the minimization problem.

PROPOSITION 3. *The functional $j(u)$ has exactly one minimum $u \in H$.*

Proof. One can apply proposition 1.2 from the book by Ekeland and Témam [15, p. 35] which states that if $j : H \to \mathbb{R}$ is strictly convex, lower semicontinuous, proper, and coercive, then it has exactly one minimum. ∎

In order to further characterize the solution, one uses the Gâteaux derivative of j at point u in direction v defined by

$$\langle \nabla j(u), v\rangle = \lim_{\tau \to 0} \frac{j(u + \tau v) - j(u)}{\tau}. \tag{15}$$

The Gâteaux derivative of the first part of $j(u)$ can be seen to be $(u,v)_H$, and the derivative of the last part is $\frac{1}{n}\sum_{i=1}^n v(t_i)$. The Gâteaux derivative of the second part is

$$\lim_{\tau\to 0}\frac{1}{\tau}\log\frac{\int_T e^{u(t)}e^{\tau v(t)}\,d\tau}{\int_T e^{u(t)}\,dt}.$$

In order to obtain the derivative, we use the Taylor remainder formula

$$e^{\tau v(t)} = 1 + \tau v(t) + e^{\tau v(t)\eta(t)}\frac{\tau^2 v(t)^2}{2}$$

for some $\eta(t)\in[0,1]$. Inserting this into the integral gives

$$\log\left(\frac{\int_T e^{u(t)}e^{\tau v(t)}\,d\tau}{\int_T e^{u(t)}\,dt}\right) = \log\left(1+\tau\frac{\int_T e^{u(t)}v(t)\,dt}{\int_T e^{u(t)}\,dt}+R(\tau)\right),$$

where

$$|R(\tau)| \le \frac{\tau^2\|v\|_\infty^2}{2}e^{\tau\|v\|_\infty}.$$

Next we use the Taylor expansion for the logarithm $\log(1+z) = z - \frac{z^2}{2(1+\zeta)^2}$ for appropriate ζ and get the Gâteaux derivative of the second term of $j(u)$ as

$$\frac{\int_T e^{u(t)}v(t)\,dt}{\int_T e^{u(t)}\,dt}.$$

It follows that the Gâteaux derivative of $j(u)$ in direction v is

$$\langle\nabla j(u), v\rangle = (u,v)_H + \frac{\int_T e^{u(t)}v(t)\,dt}{\int_T e^{u(t)}\,dt} - \frac{1}{n}\sum_{i=1}^n v(t_i). \tag{16}$$

The Gâteaux derivative is clearly a linear functional of v, it has three parts which are all bounded, and thus the functional is continuous. The continuity of the second part follows from the triangular inequality, and one gets

$$\left|\frac{\int_T e^{u(t)}v(t)\,dt}{\int_T e^{u(t)}\,dt}\right| \le \|v\|_\infty \le C_H\|v\|_H.$$

Combining the bounds for the three parts gives the estimate

$$\|\nabla j(u)\| \le (1+2C_H)\|u\|_H$$

for the operator norm of the Gâteaux derivative. The characterization of the minimum is now given by the following.

PROPOSITION 4. *Under the conditions of the previous proposition, u minimizes j if and only if*

$$(u,v)_H + \frac{\int_T e^{u(t)}v(t)\,dt}{\int_T e^{u(t)}\,dt} - \frac{1}{n}\sum_{i=1}^n v(t_i) = 0, \quad \textit{for all } v\in H. \tag{17}$$

Proof. The result follows by application of proposition 2.1 in [15, pp. 36/37] since the Gâteaux derivative is continuous. ∎

3.2. Duality. We now consider another approach to deal with the nonlinearity which this time is exact. In general, one would like to separate components one and three in j from the second one, and this can be done using Fenchel's results [15]. Let $j_0 : H \to \mathbb{R}$ be the functional of the approximated uniform distribution used previously. Furthermore, let j_1 be the nonlinear functional with domain $\operatorname{dom} j_1 \subset L_2(T)$ (the domain of a functional is defined as the set of arguments for which the values of the functional are finite) defined by

$$j_1(u) = \log \int_T \exp(u(t))\, dt.$$

Now let $E : H \to L_2(T)$ be the (continuous) embedding so that one gets

$$j(u) = j_0(u) + j_1(Eu).$$

Remember that the dual of a functional ϕ is defined as

$$\phi^*(u^*) = \sup_{u \in H} \left(\langle u^*, u \rangle - \phi(u) \right),$$

where u^* is in the dual space of the space of u. The dual of j_0 is then defined for $u^* \in H$ as

$$j_0^*(u^*) = \sup_{u \in H} \left((u^*, u)_H - \frac{1}{2}\|u\|_H^2 + \frac{1}{n} \sum_{i=1} u(t_i) \right)$$

since H is dual to itself. Using the reproducing kernel k_t (and the reproducing property $u(t) = (k_t, u)_H$), one sees that

$$j_0^*(u^*) = \sup_{u \in H} \left(-\frac{1}{2} \left\| u - \frac{1}{n} \sum_{i=1}^n k_{t_i} - u^* \right\|_H^2 + \frac{1}{2} \left\| \frac{1}{n} \sum_{i=1}^n k_{t_i} + u^* \right\|_H^2 \right),$$

and it follows that at the supremum the first term becomes zero as it is nonpositive and so

$$j_0^*(u^*) = \frac{1}{2} \left\| \frac{1}{n} \sum_{i=1}^n k_{t_i} + u^* \right\|_H^2 = \frac{1}{2}\|u^*\|_H^2 + \frac{1}{n} \sum_{i=1}^n u^*(t_i) + \frac{1}{2} \left\| \frac{1}{n} \sum_{i=1}^n k_{t_i} \right\|_H^2.$$

The determination of the dual of j_1 is simplified considerably by choosing the domain to be a subset of $L_2(T)$ instead of H as in the case of j_0. In order to distinguish between $L_2(T)$ and H, we let $z = Eu$ denote the value of the function u as an element of $L_2(T)$. One then has

$$j_1^*(z^*) = \sup_{z \in L_2(T)} \left(\int_T z^*(t) z(t)\, dt - \log \int_T e^{z(t)}\, dt \right),$$

which maximizes

$$\phi(z; z^*) = \int_T z^*(t) z(t)\, dt - \log \left(\int_T e^{z(t)}\, dt \right).$$

This functional has a Gâteaux derivative with respect to z given by

$$\nabla_z \phi(z; z^*)(t) = z^*(t) - \frac{e^{(z(t))}}{\int_T e^{(z(s))}\, ds}.$$

If z^* is a maximizer of the functional $\phi(z; z^*)$, the Gâteaux derivative has to be zero. As the second term is positive and is subtracted, this can thus happen only if $z^*(t) > 0$. As the integral (over t) of the second term is one, it follows that one also has $\int_T z^*(t) = 1$. In summary, j_1^* can be well defined and finite only if z^* is a probability density. What remains is the determination of $z(t)$ as a function of $z^*(t)$ in the case where $\phi(z; z^*)$ is maximized. With $\gamma(z) = \log \int_T \exp(z(t))\, dt$ one has for the maximum

$$z(t) = \log z^*(t) + \gamma(z).$$

As for the maximizing z one has $j_1^*(z^*) = \phi(z; z^*)$ and one gets

$$\begin{aligned} j_1^*(z^*) &= \int_T z^*(t)(\log(z^*(t)) + \gamma(z))\, dt - \log\left(e^{\gamma(z)} \int_T z^*(t)\, dt\right) \\ &= \int_T z^*(t) \log(z^*(t))\, dt + \gamma(z) \int_T z^*(t)\, dt - \gamma(z) - \log \int_T z^*(t)\, dt. \end{aligned}$$

Recall that the supremum of ϕ is finite only if z^* is a probability distribution, in particular, if $\int_T z^*(t)\, dt = 1$. In this case the last three terms of j_1^* vanish, and one thus gets

$$j_1^*(z^*) = \begin{cases} \int_T z^*(t) \log z^*(t)\, dt, & \text{if } z^*(t) > 0 \text{ a.e. and } \int_T z^*(t)\, dt = 1, \\ \infty, & \text{else.} \end{cases} \tag{18}$$

The functional j_1^* thus turns out to be just the entropy of the probability distribution z^*.

By the Cauchy–Schwartz inequality and the embedding property one gets

$$|(z^*, Eu)| = \left| \int_T z^*(t) u(t)\, dt \right| \le \|u\|_\infty \int_T z^*(t)\, dt \le C_H \|u\|_H \int_T z^*(t)\, dt,$$

and it follows that (z^*, Eu) is a continuous functional with respect to u for every probability density z^*. By the Riesz representation theorem there exists a $E^* z^* \in H$ such that

$$(z^*, Eu) = (E^* z^*, u)_H,$$

where E^* is the dual of E. Using the representation theorem again, one obtains the values of the function $E^* z^*$ as

$$E^* z^*(t) = (E^* z^*, k_t)_H = (z^*, E k_t) = \int_T z^*(s) k_t(s)\, ds.$$

Now Fenchel duality theory (see, e.g., [15, pp. 59ff]) tells us that

$$\min_{u \in H} j_0(u) + j_1(Eu) = - \min_{z \in L_2(T)} j_0^*(-E^* z^*) + j_1^*(z^*).$$

In order for the Fenchel duality result to hold, one requires a condition on the functionals j_i to hold, the so-called *constraint qualification*. In this case we can show that

$$E \operatorname{dom} j_0 \cap \operatorname{dom} j_1 \neq \emptyset.$$

In fact, the domain of j_0 is H and, on the set $EH \subset L_2(T)$, the functional j_1 is bounded as we have seen previously by

$$j_1(Eu) = \log \int_T e^{u(t)}\, dt \le C_H \|u\|_H, \quad u \in H.$$

The Fenchel duality theory also provides a way to compute the solution of the primal problem if the solution of the dual problem is known. In full generality, one has for the subdifferential $\partial j_0(u)$ of j_0 at u the inclusion

$$-E^* z^* \in \partial j_0(u)$$

if z^* is the solution to the dual problem and u is the solution to the primal problem. In the case considered here j_0 is differentiable and so $-E^*z^*$ is just the gradient; i.e., the subdifferential inclusion property can be seen to simplify to

$$-E^* z^* = u + \frac{1}{n}\sum_{i=1}^n k_{t_i}.$$

Consequently,

$$u(t) = \int_T k(t,s) z^*(s)\, ds - \frac{1}{n}\sum_{i=1}^n k(t,t_i), \tag{19}$$

which means that u is the difference of the expectation of the kernel with respect to z^* and the empirical expectation. The determination of u is thus in terms of the complexity as expensive as the evaluation of $\frac{1}{n}\sum_{i=1}^n k(t,t_i)$, i.e., has the same complexity as a kernel density estimator once the solution z^* of the dual problem is known. Such an approach may be computationally advantageous in the case of a moderate number of data points n. It also provides connections to maximum entropy methods. For the next sections, however, (19) is important as it establishes the regularity of the solution.

The dual problem consists of finding the minimum of

$$j_0^*(-E^*z^*) + j_1^*(z^*) = \frac{1}{2}\left\| \frac{1}{n}\sum_{i=1}^n k_{t_i} - \int_T k_t z^*(t)\, dt \right\|_H^2 + \int_T z^*(t) \log z^*(t)\, dt.$$

Note that the first term in this sum is just one-half the norm of u squared, and one thus has that the dual problem consists of minimizing

$$\Phi(u, z^*) = \frac{1}{2}\|u\|_H^2 + \int_T z^*(t) \log z^*(t)\, dt$$

as a function of u and z^* with the linear constraint

$$u(t) + \int_T k(t,s) z^*(s)\, ds = \frac{1}{n}\sum_{i=1}^n k(t_i, t). \tag{20}$$

This *augmented problem* is known in machine learning and least squares problems; see, e.g., [16]. The dual problem balances thus the H-norm of u with the entropy of z^*.

If one takes the Gâteaux derivative (using similar arguments as above) as for the primal problem, one now gets an integral equation for z^* as

$$\log z^*(t) + \int_T k(t,s) z^*(s)\, ds = \frac{1}{n} \sum_{i=1}^{n} k(t_i, t) - 1.$$

The solution of the dual problem requires thus the solution of a nonlinear Fredholm integral equation of the first kind. Computational approaches based on these equations are very popular in the machine learning community. These are summarized under the term kernel methods and are related to the radial basis function approaches in interpolation and smoothing. An advantage of this approach is that it does not require the determination of the H-norm but relies solely on the reproducing kernels. The dual "integral equation" approach is also mentioned in [4]. In the following we will instead discuss the numerical solution of the primal problem which consists of minimizing the functional $j(u)$ defined in (10).

4. A Newton–Galerkin method. After having considered properties of the optimization problem we now focus on numerical techniques for the determination of the minimum which is characterized by the (nonlinear) Galerkin equations:

$$\langle \nabla j(u), v \rangle = (u, v)_H + \frac{\int_T e^{u(t)} v(t)\, dt}{\int_T e^{u(t)}\, dt} - \frac{1}{n} \sum_{i=1}^{n} v(t_i) = 0, \quad v \in H.$$

The Gâteaux derivative $\nabla j(u)$ is a continuous functional and, by the Riesz representation theorem, there is a $F(u) \in H$ such that

$$\langle \nabla j(u), v \rangle = (F(u), v)_H, \quad u, v \in H.$$

One has the explicit representation

$$F(u) := u + \frac{\int_T e^{u(t)} k_t\, dt}{\int_T e^{u(t)}\, dt} - \frac{1}{n} \sum_{i=1} k_{t_i}$$

where the integral

$$\int_T e^{u(t)} k_t\, dt$$

is defined in a weak sense using the continuous linear functional

$$v \to \int_T e^{u(t)} v(t)\, dt.$$

Continuity of this functional on $L_2(T)$ and thus on H is established as every $u \in H$ is a continuous function. The integral $\int_T e^{u(t)} k_t\, dt$ is then defined using the Riesz representation theorem for this functional.

With this the optimization problem is reduced to finding the solution of

$$F(u) = 0.$$

As we will use Newton's method, we first need to determine the Fréchet derivative DF of F.

THEOREM 1. *The nonlinear operator $F(u)$ is Fréchet differentiable for all $u \in H$ and the derivative $DF(u)$ is defined by*

$$DF(u)\,v = v + \frac{\int_T e^{u(t)} k_t v(t)\,dt}{\int_T e^{u(t)}\,dt} - \frac{\int_T e^{u(t)} v(t)\,dt \int_T e^{u(t)} k_t\,dt}{\left(\int_T e^{u(t)}\,dt\right)^2}, \quad v \in H,$$

where the integral $\int_T e^{u(t)} k_t v(t)\,dt \in H$ is defined in the weak sense introduced above.

Proof. We need to show that

$$\lim_{v \to 0} \frac{\|F(u+v) - F(u) - DF(u)v\|_H}{\|v\|_H} = 0. \tag{21}$$

With $\psi(v) := \|F(u+v) - F(u) - DF(u)v\|_H$ one gets from the triangle inequality

$$\begin{aligned}
\psi(v) &= \left\| \frac{\int_T e^{u+v} k_t\,dt}{\int_T e^{u+v}\,dt} - \frac{\int_T e^u k_t\,dt}{\int_T e^u\,dt} - \frac{\int_T e^u k_t v\,dt}{\int_T e^u\,dt} + \frac{\int_T e^u v\,dt \int_T e^u k_t\,dt}{\left(\int_T e^u\,dt\right)^2} \right\|_H \\
&\le \left\| \frac{\int_T e^{u+v} k_t\,dt}{\int_T e^{u+v}\,dt} - \frac{\int_T e^u k_t\,dt}{\int_T e^{u+v}\,dt} - \frac{\int_T e^u k_t v\,dt}{\int_T e^{u+v}\,dt} \right\|_H \\
&\quad + \left\| \frac{\int_T e^u k_t\,dt}{\int_T e^{u+v}\,dt} - \frac{\int_T e^u k_t\,dt}{\int_T e^u\,dt} - \frac{\int_T e^u k_t\,dt \int_T e^u v\,dt}{\left(\int_T e^u\,dt\right)^2} \right\|_H \\
&\quad + \left\| \frac{\int_T e^u k_t v\,dt}{\int e^{u+v}\,dt} - \frac{\int_T e^u k_t v\,dt}{\int_T e^u\,dt} \right\|_H,
\end{aligned}$$

where u and v denote $u(t)$ and $v(t)$, respectively.

Now we proceed to bound these three terms. For the first term we have the estimate

$$\frac{\left\|\int_T \left(e^{u+v} - e^u - ve^u\right) k_t\,dt\right\|_H}{\left|\int_T e^{u+v}\,dt\right|} \le \frac{\|v\|_\infty^2 e^{\|u\|_\infty + \|v\|_\infty} C_H}{2\,e^{-\|u\|_\infty - \|v\|_\infty}}$$

by the Taylor reminder theorem, using bounds on the integrand and the reproducing kernel property. It then follows that the first term is of order $O(\|v\|_\infty^2) = O(\|v\|_H^2)$ for small v. Using the quotient rule of differentiation and a bound for the difference $\int_T e^{u+v}\,dt - \int_T e^u\,dt - \int_T e^u v\,dt$, one sees that the second term is also $O(\|v\|_H^2)$. The third term can be seen to be a product of two $O(\|v\|_H)$ terms and is therefore $O(\|v\|_H^2)$ as well. The sum of these terms is then $O(\|v\|_H^2)$ and indeed (21) holds. It follows that $DF(u)$ is the Fréchet derivative of $F(u)$. ∎

To establish Newton's method we introduce the bilinear form corresponding to $DF(u)$:

$$a(v_1, v_2; u) := (v_1, DF(u)v_2)_H = (v_1, v_2)_H + \frac{\int_T e^u v_1 v_2\,dt}{\int_T e^u\,dt} - \frac{\int_T e^u v_1\,dt \int_T e^u v_2\,dt}{\left(\int_T e^u\right)^2},$$

where the dependencies on t of u, v_1, and v_2 are omitted to shorten the notation. Clearly this defines a local energy norm $v \to a(v, v; u)$. Using this bilinear form, a Newton step which consists of determining $\Delta u \in H$ from the linear operator equations

$$DF(u)\Delta u = -F(u)$$

may be restated in weak form as

$$a(v, \Delta u; u) = -(v, F(u))_H \quad \text{for all } v \in H.$$

Interestingly, the bilinear form has a statistical interpretation. With

$$f(t; u) = e^{u(t)} \Big/ \int_T e^{u(s)}\, ds,$$

one can see that the second and third terms of the bilinear form a can be interpreted as the covariance of v_1 and v_2 with respect to f and one has

$$a(v_1, v_2; u) = (v_1, v_2)_H + \int_T f(t; u)\,(v_1(t) - E(v_1; u))\,(v_2(t) - E(v_2; u))\, dt,$$

where

$$E(v; u) := \int_T f(t; u) v(t)\, dt$$

is the expectation of v with respect to $f(t; u)$. From this we get the following.

PROPOSITION 5. *The local energy norm is uniformly (in u) equivalent to the H-norm, and one has*

$$\|v\|_H^2 \le a(v, v; u) \le (1 + C_H^2)\|v\|_H^2, \quad v \in H.$$

Furthermore, one has the bound

$$|a(v_1, v_2; u)| \le (1 + 2C_H^2)\|v_1\|_H \|v_2\|_H,$$

and it follows that the family of local bilinear forms $a(\cdot, \cdot; u)$ are uniformly (in u) H-elliptic.

Proof. The lower bound follows directly from

$$a(v, v; u) = \|v\|_H^2 + \int_T f(t; u)\,(v(t) - E(v; u))^2\, dt.$$

For the upper bound one expands the square in the integral to get

$$a(v, v; u) = \|v\|_H^2 + E(v^2; u) - E(v; u)^2 \le \|v\|_H^2 + E(v^2; u) \le \|v\|_H^2 + \|v\|_\infty^2$$

and the bound follows by $\|v\|_\infty \le C_H \|v\|_H$. Similarly, for the boundedness one has

$$a(v_1, v_2; u) = (v_1, v_2)_H + E(v_1 v_2; u) - E(v_1; u) E(v_2; u)$$

and thus

$$|a(v_1, v_2; u)| \le |(v_1, v_2)_H| + 2\|v_1\|_\infty \|v_2\|_\infty \le (1 + 2C_H^2)\|v_1\|_H \|v_2\|_H. \qquad \square$$

From this one gets directly the following application of Céa's lemma.

COROLLARY 2 (Céa). *Let $V_h \subset H$ be a finite-dimensional subspace, $\Delta u \in H$ satisfy*

$$a(v, \Delta u; u) = -(v, F(u))_H, \quad v \in H, \tag{22}$$

and $\Delta u_h \in V_h$ *satisfy*

$$a(v, \Delta u_h; u) = -(v, F(u))_H, \quad v \in V_h.$$

Then

$$\|\Delta u - \Delta u_h\|_H \leq (1 + 2C_H^2)\|\Delta u - v\|_H, \quad v \in V_h.$$

In practice, the linear equations defining Δu_h are solved iteratively with a certain accuracy. Note that at this stage V_h merely denotes an abstract finite-dimensional subspace of H. It will be specified later to be a space of piecewise multilinear functions.

We have now all the tools to establish the numerical solution procedure which is a *Newton–Galerkin iteration with damping*. In particular, the method is based on an iteration

$$u^{k+1} = u^k + \lambda_k \delta u^k,$$

where the correction $\delta u^k \in V_h$ satisfies

$$a(v, \delta u^k; u^k) = -(v, F(u^k))_H, \quad v \in V_h. \tag{23}$$

An important point is the choice of the damping parameter. The damping parameter λ_k in particular has to ensure that the method is globally convergent. In practice we found that the Armijo choice (see, e.g., [17]) worked best. Deuflhard and Weiser [18] provide conditions under which the Newton–Galerkin method is a descent method and we have here the following.

THEOREM 2. *Let* u^k *and* δu^k *be given by the Newton–Galerkin method defined above. If the damping parameter* λ_k *satisfies*

$$0 \leq \lambda_k \leq \lambda_K^{max} \frac{4}{1 + \sqrt{1 + 8h_k/3}},$$

where $h_k = 2C_H^3\sqrt{\epsilon_k}$ *and* $\epsilon_k = a(\delta u^k, \delta u^k; u^k)$*, then*

$$j(u^{k+1}) \leq j(u^k) - t_k(\lambda_k)\epsilon_k,$$

where $t_k(\lambda_k) = \lambda - \lambda_k^2/2 - h_k\lambda_k^3/6$.

If in addition $\delta\lambda_k \leq \lambda_k^{max} - \delta$ *for sufficiently small* $\delta > 0$*, then the sequence* u^k *converges to the minimum in* V_h.

Proof. This theorem is a direct consequence of theorem 2.2 in [18]. We thus only need to show that the conditions of the theorem are satisfied.

First we note that the "Galerkin condition"

$$(\delta u^k, r^k)_H = 0$$

holds where $r^k = DF(u^k)\,\delta u^k - F(u^k)$ as we are using a Galerkin method to solve the inner equations (23).

Now we show the special affine conjugate Lipschitz condition. As $a(v, v; u) \geq \|v\|_H$ one has

$$\|DF(u)^{-1}\| = \sup_{v \in H} \frac{\|v\|_H^2}{a(v, v; u)} \leq 1. \tag{24}$$

By definition,

$$\|DF(u_2) - DF(u_1)\| = \sup_{v_1,v_2 \in H} \frac{a(v_1, v_2; u_2) - a(v_1, v_2; u_1)}{(v_1, v_2)_H}.$$

An application of the triangle inequality for integration results in

$$|a(v_1, v_2; u_2) - a(v_1, v_2; u_1)| \leq \int_T |f(t; u_2) - f(t; u_1)| v_1(t) v_2(t)\, dt.$$

The right-hand side corresponds to a positive semidefinite continuous operator, and one thus, by the Cauchy–Schwarz inequality, gets

$$\|DF(u_2) - DF(u_1)\| \leq \sup_{v \in H} \frac{\int_T |f(t; u_2) - f(t; u_1)| v(t)^2\, dt}{\|v\|_H^2}.$$

By the mean value theorem one has

$$f(t; u_2) - f(t; u_1) = \frac{d}{d\theta} f(t; u_1 + \theta(u_2 - u_1))$$

for some $\theta \in [0, 1]$. Application of the quotient rule of differentiation gives

$$\frac{d}{d\theta} f(t; u_1 + \theta(u_2 - u_1)) = \left(u_2(t) - u_1(t) - \int_T (u_2(s) - u_1(s)) f(s; u_1 + \theta(u_2 - u_1))\, ds \right) f(t; u_1 + \theta(u_2 - u_1))$$

from which the following bound is derived:

$$\left| \frac{d}{d\theta} f(t; u_1 + \theta(u_2 - u_1)) \right| \leq 2\|u_2 - u_1\|_\infty f(t; u_1 + \theta(u_2 - u_1)).$$

Using $\|u_2 - u_1\|_\infty \leq C_H \|u_2 - u_1\|_H$, this then leads to

$$\|DF(u_2) - DF(u_1)\| \leq \sup_{v \in H} \frac{\int_T |f(t; u_2) - f(t; u_1)| v(t)^2\, dt}{\|v\|_H^2} \leq 2C_H^3 \|u_2 - u_1\|_H. \tag{25}$$

In contrast to most other nonlinear problems we have the explicit bounds (24) and (25). As a consequence one can then get the more general Lipschitz conditions:

$$\|DF(z)^{-1}(DF(x) - DF(y))\| \leq 2C_H^3 \|y - x\|_H$$

and

$$\|DF(x)^{-1}(DF(y) - DF(x))(y - x)\|_H \leq 2C_H^3 \|y - x\|_H^2.$$

The affine invariant counterparts of these bounds follow easily from the equivalence of the local energy norms and the H-norm.

For the convergence we first observe that $DF(u)$ is uniformly positive definite (as the energy norm is equivalent to the H-norm) and furthermore the level sets defined by $j(u) \leq$ const are closed and bounded. The convergence then follows from theorem 2.4 of [18]. ∎

A quadratic convergence result is also available in [19, 20]. The proof is similar to the previous ones. To simplify, we introduce the following notation. First the local energy norm shall be denoted as

$$\|v\|_u := a(v, v; u)^{1/2}.$$

We assume that u^k is obtained by an ordinary Newton–Galerkin method, i.e.,

$$u^{k+1} = u^k + \delta u^k,$$

where $\delta u^k \in V_h$ solves the Galerkin equations

$$a(v, \delta u^k; u^k) = -(v, F(u^k))_H, \quad v \in V_h,$$

and thus one gets for the residual $r^k = DF(u^k)\delta u^k + F(u^k)$ the Galerkin condition

$$(\delta u^k, r^k)_H = 0.$$

We will also use the approximation error e^k defined by $DF(u^k)e^k = r^k$. One then has the following.

THEOREM 3. *Let $h_k = 2C_H^3\|\delta u^k\|_{u^k}$ and u^0 such that $h_0 \le 2/(1+\rho)$ for some $\rho > 0$. If the accuracy*

$$\delta_k = \frac{\|e^k\|_{u^k}}{\|\delta u^k\|_{u^k}}$$

of the Galerkin approximation is such that

$$\delta_k \le \frac{\rho h_k}{h_k + \sqrt{4 + h_k^2}},$$

then the Newton–Galerkin iterates u^k converge quadratically to the minimizer of j such that

$$\|\delta u^{k+1}\|_H \le (1+\rho)C_H^3(1 + C_H^2)\|\delta u^k\|_H^2.$$

Proof. In order to apply the theorem from [19] one needs to (i) show the conjugate Lipschitz condition for collinear v_1, v_2, and v_3:

$$\|DF(v_3)^{-1}(DF(v_1) - DF(v_2))v\|_{v_3} \le 2C_H^3\|v_1 - v_2\|_{v_2}\|v\|_{v_2}$$

(which follows directly from (25)), (ii) verify the closedness and boundedness of the set $\{v \mid j(v) \le j(u^0)\}$, and (iii) invoke the equivalence of the H-norm with the local energy norms. ∎

In short, the theorem states that if the approximation error of the Galerkin method is sufficiently small, then the inexact Newton method converges like the exact Newton method.

5. Approximation spaces, approximation properties, and discretization. So far, we were not specific what finite-dimensional subspace V_h we use for the discretization. In the following we first consider function approximation in Hilbert spaces with an orthogonal system and use this abstract approach for error estimates in various norms provided that specific mixed regularity assumptions are valid. Here, we focus on pure approximation results and discretization error estimates. Later, in

the numerical experiments, we will use the space of piecewise multilinear functions on uniform grids for discretization.

Let us equip the Hilbert space H with an orthonormal system $\{\phi_i, i \in \mathbb{N}\}$. Any function $u \in H$ is then represented as

$$u(x) = \sum_{i \geq 1} c_i \phi_i(x) \quad \text{with coefficients } c_i = (\phi_i, u)_H. \tag{26}$$

In the multivariate case we employ a product construction; i.e., we use the coordinates $x = (x_1, \ldots, x_d)$, multi-indices $i = (i_1, \ldots, i_d)$, and the product system $\{\phi_i(x) = \prod_{j=1}^d \phi_{i_j}(x_{i_j})\}$. We sum in (26) on $i \in \mathbb{N}^d$.

In the following, we consider the spaces H^t_{mix} of dominating mixed derivatives defined on $T = [0,1]^d$ as

$$H^t_{mix}(T) = \left\{ u(x) = \sum_i c_i \phi_i : ||u||_{H^t_{mix}} = \left(\sum_i \left(\prod_{j=1}^d i_j^{2t} \right) |c_i|^2 \right)^{1/2} < \infty \right\}.$$

Note that $H^t_{mix}(T) = H^t([0,1]) \times \cdots \times H^t([0,1])$; i.e., $H^t_{mix}(T)$ is just the product of Sobolev spaces H^t on the one-dimensional domains $[0,1]$. We here consider $H^t_{mix}(T)$ since in later applications we just choose $H^t_{mix}(T)$ as the reproducing kernel Hilbert space H. This is due to the choice of the corresponding reproducing kernel $k_t = K(t, \cdot)$ as a product of one-dimensional kernels which allows a straightforward extension of our approach to the general d-dimensional case.[5] In the definition of H^t_{mix}, we directly see how the decay of the coefficients in different coordinate directions enters multiplicatively. It also holds that if $u \in H^t_{mix}$, then

$$|c_i| \leq C \cdot \frac{1}{\prod_{j=1}^d i_j^t} \quad \text{for all } c_i.$$

For further details on the Sobolev spaces H^t_{mix} of dominating mixed derivatives see, e.g., [21, 22, 23].

Now, let us define the approximation spaces[6]

$$V_m = \text{span}\{\phi_i : |i|_\infty \leq m\} = \left\{ u(x) = \sum_{|i|_\infty \leq m} c_i \phi_i(x) \right\}. \tag{27}$$

Their dimension is clearly

$$|V_m| = m^d,$$

and we see an exponential dependence on the dimension d which resembles the curse of dimensionality. In this respect, V_h resembles a conventional discretization on, e.g., a "full grid."

[5] Note here that a choice of $H = H^t(T)$ as standard Sobolev space has to cope with the Sobolev embedding; i.e., t then depends on d with $t = \lceil \frac{d+1}{2} \rceil$, whereas for our choice $H = H^t_{mix}(T)$ this is not the case.

[6] We use here the notation V_m for the approximation space V_h, $h = 1/m$, to allow for general values of m and general, not necessarily dyadic, function systems $\{\phi_i\}$. Later on, we will switch to uniform meshes and a doubling of the degrees of freedom from level to level and will use $m = 2^l - 1, l \in \mathbb{N}$, the mesh size $h = 1/m$, and the notation V_h again.

We then have the following theorem for the approximation error in the H^s_{mix}-norm.

THEOREM 4. *Let $s \le t$. Furthermore let $u = \sum_{i\in\mathbb{N}^d} c_i\phi_i(x) \in H^t_{mix}$ and let $u_m(x) = \sum_{|i|_\infty\le m} c_i\phi_i(x) \in V_m$. For the approximation error, we then have the estimate*

$$\inf_{u_m\in V_m} ||u-u_m||_{H^s_{mix}} \le (m+1)^{-(t-s)}||u||_{H^t_{mix}}.$$

Proof. Plug $u(x) = \sum_{i\in\mathbb{N}} c_i\phi_i$ and $u_m = \sum_{|i|_\infty\le m} c_i\phi_i$ into $||u-u_m||_{H^s_{mix}}$. This directly gives

$$\begin{aligned}
||u-u_m||^2_{H^s_{mix}} &= \left|\left|\sum_{i\in\mathbb{N}} c_i\phi_i - \sum_{|i|_\infty\le m} c_i\phi_i\right|\right|^2_{H^s_{mix}} = \left|\left|\sum_{|i|_\infty> m} c_i\phi_i\right|\right|^2_{H^s_{mix}} \\
&= \sum_{|i|_\infty>m}\left(\prod_{j=1}^d i_j^{2s}\right)|c_i|^2 = \sum_{|i|_\infty>m}\frac{\prod_{j=1}^d i_j^{2s}\prod_{j=1}^d i_j^{2t}}{\prod_{j=1}^d i_j^{2t}}|c_i|^2 \\
&\le \max_{|i|_\infty>m}\frac{\prod_{j=1}^d i_j^{2s}}{\prod_{j=1}^d i_j^{2t}}\sum_{|i|_\infty>m}\prod_{j=1}^d i_j^{2t}|c_i|^2 \\
&\le \max_{|i|_\infty>m}\frac{\prod_{j=1}^d i_j^{2s}}{\prod_{j=1}^d i_j^{2t}}||u||_{H^t_{mix}}. \qquad (28)
\end{aligned}$$

With $\prod_{j=1}^d i_j^{2s}/\prod_{j=1}^d i_j^{2t} = \prod_{j=1}^d i_j^{2s-2t}$, the maximum is attained at $(m+1)^{2(s-t)}$ for $s \le t$. ∎

In later applications we will measure the error in the H^s_{mix}-norm for $s = 0$ using a H^t_{mix}-regularity of u with $t = 3/2-\varepsilon, \varepsilon > 0$. This predicts an approximation rate of $3/2-2\varepsilon$.

But note that Theorem 4 merely states an approximation rate and not a convergence rate of the Galerkin discretization yet. To derive an estimate for that, we would need a nonlinear Céa lemma which is not available. To this end, we can at least resort to the local energy norm $a(v,v;u)$ within our Newton–Galerkin method. After convergence this also gives us the energy norm in the solution. Recall from Proposition 5 that the local energy norm is uniformly (in u) equivalent to the H-norm and we have $\|v\|^2_H \le a(v,v;u) \le (1+C_H^2)\|v\|^2_H, v \in H$, which gives us a local version of Céa's lemma. If u_m now denotes the Galerkin solution of the (local) discrete problem (22), we can estimate the discretization error in V_m for our choice $H = H^s_{mix}$ in the energy norm as

$$\begin{aligned}
a(u-u_m,u-u_m,w) &\le (1+C_H^2)\|u-u_m\|^2_H = (1+C_H^2)\|u-u_m\|^2_{H^s_{mix}} \\
&\le (1+C_H^2)(m+1)^{-(t-s)}\|u\|_{H^t_{mix}} \quad \forall w \in H, \qquad (29)
\end{aligned}$$

and we get, using, for example, piecewise multilinear functions in the construction of the orthogonal system $\{\phi_i\}$, with $s = 1$ (kernel k for H which involves the weak form of a second order differential operator in the $\|\cdot\|_H$-norm) and $t = 3/2-\varepsilon$ the rate $1/2-\varepsilon$ for the discretization error in V_m with respect to the energy error.

If we want to predict the rates of the discretization error with respect to the L_2-norm, we have to resort to the well-known Aubin–Nitsche lemma. For the specific

pair of spaces L_2 and H^t_{mix} with associated norms $\|\cdot\|_{L_2}$ and $\|\cdot\|_{H^t_{mix}}$ it reads as follows; compare also [24].

LEMMA 1. *There holds for the Galerkin solution u_m in $V_m \subset H^s_{mix}$*

$$\|u-u_m\|_{L_2} \le C\cdot\|u-u_m\|_{H^s_{mix}} \sup_{g\in L_2}\left\{\frac{1}{\|g\|_{L_2}}\inf_{v\in V_m}\|\phi_g - v\|_{H^s_{mix}}\right\},$$

where, to each $g \in L_2$, the unique weak solution $\phi_g \in H^s_{mix}$ of the equation

$$a(w,\phi_g) = (g,w) \quad \textit{for } w \in H^s_{mix}$$

is assigned. From this we directly get for $s = 1$

$$\begin{aligned}\|u-u_m\|_{L_2} &\le C\cdot\|u-u_m\|_{H^1_{mix}} \sup_{g\in L_2}\left\{\frac{1}{\|g\|_{L_2}}\inf_{v\in V_m}\|\phi_g - v\|_{H^1_{mix}}\right\}\\ &\le C\cdot\|u-u_m\|_{H^1_{mix}} \sup_{g\in L_2}\left\{\frac{1}{\|g\|_{L_2}}\tilde{C}\cdot(m+1)^{-1}\|g\|_{L_2}\right\}\\ &\le \hat{C}\cdot(m+1)^{-(t-1)}\|u\|_{H^t_{mix}}(m+1)^{-1} = \hat{C}\cdot(m+1)^{-t}\|u\|_{H^t_{mix}}, \qquad (30)\end{aligned}$$

which results in the discretization error rate $3/2-\varepsilon$ for $t = 3/2-\varepsilon$. Note here that H^t_{mix} is continuously embedded in L_2; i.e., the prerequisite of the Aubin–Nitsche lemma also holds for the spaces of bounded mixed derivatives.

Using $\|\cdot\|_{L_1} \lesssim \|\cdot\|_{L_2}$ we finally obtain estimates for the discretization error also in the L_1-norm.

If we want to switch now to function systems with dyadic refinement, we may proceed as follows: We first consider the case $d = 1$. In [25] it has been shown that for $s < 2$ the norm defined by

$$|u|_s^2 = \|P_0 u\|_{L_2}^2 + \sum_{k=1}^{\infty} 2^{-ks}\|P_k u - P_{k-1}u\|_{L_2}^2$$

defines a norm on the (usual) Sobolev spaces $H^s[0,1]$ where P_k is the L_2-orthogonal projection onto V_h for $h = 2^{-k}$. The reasoning uses Jackson and Bernstein inequalities and is based on the Strang–Fix condition for the hat function; see section 5 of [25]. For example, for the case of homogeneous Dirichlet boundary conditions, we can now choose any L_2-orthogonal basis ψ_i (with obvious modifications for other boundary conditions) which satisfies $V_h = \operatorname{span}\{\psi_1,\dots,\psi_{2^k-1}\}$ and the norm $\|\cdot\|_{H^s}$ can then be shown to be equivalent to the norm $|\cdot|_s$ defined above. Taking tensor products then gives the analogous result for $H^s_{\mathrm{mix}}([0,1]^d)$. The corresponding rates in Theorem 4, inequality (29), and inequality (30) then relate, of course, to base h instead of to base $(m+1)^{-1}$. A similar reasoning allows one to use here also more general wavelet-like bases and frames [26] instead of L_2-orthogonal function systems. If we then employ, for example, prewavelets based on piecewise multilinear functions [27], we may use the fact that they span in our full grid case the same space as the standard multilinear "hat" functions on the full grid and our results hold for this case as well (albeit with different constants in the estimates).

Note at this point that instead of the full grid with index set $|i|_\infty \le m$ also other subspace constructions and associated index sets may be chosen in the definition of the approximation space (27). Examples are sparse grid spaces/hyperbolic crosses ($|i|_1 \le m+d-1$) and generalized sparse grids with more general index sets Λ.

For the approximation error estimate the maximum in (28) must then be taken over $|i|_1 > m + d - 1$ or $\mathbb{N}/\Lambda$, respectively. This way the dimension of the approximation space and thus the involved cost is often substantially reduced without compromising the rate of the approximation error and the curse of dimensionality can be broken, at least to some extent. Estimates for the discretization error in the $\|\cdot\|_H$-norm then follow the same direction as above; estimates for the L_2-norm are gained with the help of the Aubin–Nitsche lemma again.[7] For further details on optimized approximation spaces and their associated complexities, see [21, 22, 23]. In this article we will not follow this direction but stick to the simpler "full grid" case and restrict ourselves to the dimensions $d = 1, 2, 3$ for reasons of simplicity.

6. Examples. Numerical analysis is concerned with the trade-off of computational resources against the accuracy of the computed result. In this section we consider the effect of the finite-dimensional approximation on the accuracy. We cover several "synthetic" and real data sets in the one-, two-, and three-dimensional case.

Density estimation considers the reconstruction of a probability density $f(t)$ from given data $t_1, \ldots, t_n$ which presumably were drawn from this density and are pairwise independent. Except in very simple cases, the computed probability density $f_h(t)$ contains errors which reflect the limited information available and the limited computational resources used. Roughly, one may thus distinguish between two error components, namely, the statistical error and the numerical error. Here we define the statistical error as the error which is due to the limited data but also the error which is due to the statistical estimation procedure (here the MAP method, and in particular the choice of the prior). While the analysis of the statistical errors is beyond the scope of this investigation, it is important to have some idea of the statistical error as it makes little sense to make the numerical error very small compared to the statistical error. A thorough discussion of statistical aspects can be found in the statistical literature, for example, in the books by Scott [9], Silverman [28], and Tapia and Thompson [29].

6.1. Error measurement. For the analysis of the error of the approximation of probability densities a variety of error measures have been used which include several norms like the L_1- and L_2-norms and divergences like the Kullback–Leibler divergence [28, 9]. The choice of a particular error does, of course, have an effect on how well a computational procedure performs and thus has to be done carefully. This choice does depend very much on the nature of the application of the estimate.

In many cases, one uses the probability density to estimate the probability of events; i.e., for a given set $A \subset T$ one would like to estimate the integral

$$P(A) = \int_A f(t)\,dt.$$

Many statistical tests are based on such computations. The error which one gets if one uses $f_h(t)$ instead of f is then

$$|P(A) - P_h(A)| = \left| \int_A (f(t) - f_h(t))\,dt \right| \leq \int_T |f(t) - f_h(t)|\,dt,$$

and it is thus natural to measure the error $f(t) - f_h(t)$ in the L_1-norm as this gives an upper bound on any probability estimation error. In the case where $\int_T dt = 1$ one

[7] Note that due to $H = H^s_{mix}$ this is straightforward. The use of the conventional Sobolev space $H = H^s$ and a sparse grids space for V_m would cause problems due to the necessary regularity assumption on the solution of the dual problem and thus of the regularity of its right-hand side.

gets from the Cauchy–Schwartz inequality the bound

$$\int_T |f(t) - f_h(t)|\,dt \leq \sqrt{\int_T (f(t) - f_h(t))^2\,dt},$$

and so the L_2-error is also frequently used. Note, however, that this bound may give too pessimistic results in the case where the f contains narrow high peaks. In the statistical literature, the square of the L_2-error is also called integrated squared error (ISE). This error depends on the actual data samples. An error measure which does not depend on the data is the expectation of the ISE which is also called mean integrated squared error (MISE) defined by

$$E\left[\int_T |f(t) - f_h(t)|^2\,dt\right].$$

In the method considered here, approximations u_h of u are obtained. Recall that the approximation f_h of the density f is then obtained by the exponential formula as

$$f_h = \frac{\exp u_h}{\int_\Omega \exp(u_h(t))\,dt}.$$

The previous error bounds given were in terms of norms of $u_h - u$. It turns out that for a method for which u_h converges to u in the $\|\cdot\|_H$-norm one automatically gets convergence of the probability densities.

PROPOSITION 6. *Let $u_h \to u$ converge in H for $h \to 0$. Then there exists for each $h_0 > 0$ a constant $C_0 > 0$ such that*

$$\|f_h - f\|_{L_p} \leq C \cdot \|u_h - u\|_{L_p}, \quad 0 < h < h_0.$$

Proof. Let $v = u_h - u$ and

$$\phi(\theta, t) = \frac{\exp(u(t) + \theta v(t))}{\int_\Omega \exp\left(u(s) + \theta v(s)\right)\,ds}.$$

In particular, $\phi(0,t) = f(t)$ and $\phi(1,t) = f_h(t)$. The derivative with respect to θ is

$$\frac{d\phi}{d\theta}(\theta, t) = \phi(\theta, t)\left(v(t) - \int_\Omega \phi(\theta, s)v(s)\,dt\right).$$

It follows by the intermediate value theorem that the pth power of the L_p-norm of the error of the density is

$$\|f_h - f\|_{L_p}^p = \int_\Omega |\phi'(\tau(t), t)|^p\,dt = \int_\Omega |\phi(\tau(t), t)|^p \left|v(t) - \int_\Omega \phi(\tau(t), s)v(s)\,ds\right|^p\,dt.$$

As $u_h \to u$ in H one has pointwise convergence by the reproducing property, and consequently $\phi(\tau(t), t)$ is uniformly bounded independently of h. Thus the first term of the integrand is bounded. The second term can be bounded using the triangle inequality and a second application of the boundedness of ϕ to complete the proof. ∎

As a consequence, it suffices to provide estimates for the errors of the u. We have seen that the size of the error of u_h does control the size of the error of the probability density f_h. The converse, however, is not true. This reflects the simple fact that u_h

is (up to normalization) the logarithm of f_{hz}, the logarithm does have a singularity at zero, and consequently, u_h is not a continuous function of f_h.

Related to this observation is a second point: For the well posedness of the problem a careful choice of the reproducing kernel of H, i.e., the covariance of the prior, has to be made. In practice, the H-norm involves a parameter, $\alpha > 0$ (see the following examples), which has to be chosen to be sufficiently large enough. This relates to problems where f has values which are very close to zero. As in the Newton solver the local mass matrices are generated by some approximation of f, and the case where this approximation is close to zero will lead to a mass matrix which is singular because substantial regions where f is zero will lead to zero submatrices of the mass matrix. In this case the stiffness matrix of the local energy $a(v_1, v_2; u)$—which is provided by the regularization or prior—guarantees the well posedness. A possibility to deal with this issue is to consider a mixture of a uniform density and the empirical density given by the data. After the mixture is determined one may "subtract" the uniform density to retrieve an estimate of the original density. This approach, however, also has its problems—in particular, it can return negative estimates—and we will not consider it any further here. In the experiments we will always assume that $\alpha > 0$ is not too small.

6.2. Experiments. The following numerical experiments shall confirm and illustrate the error bounds for norms of $u_h - u$ obtained in the previous sections. Computational experiments were performed for one-, two-, and three-dimensional problems. The numerical method was implemented in Python using scipy, weave, and numpy for performance and pylab and open-dx for visualization. The linear systems were solved using GMRES with restart and an ILU preconditioner implemented in the PyTrilinos package. The nonlinear problem was solved with an inexact Newton method using Armijo stepsize control. Typically this required less than 10 Newton steps to get full accuracy. Furthermore, the mass matrix required numerical integration (which was done by piecewise Gauss quadrature). The code was run on a variety of computer servers, workstations, and laptops all using the Linux operating system.

The main focus of our analysis and experiments has been on numerical errors or bias. Statistical errors have not been considered; in particular, we have not investigated how more data would have improved the estimators. This has been investigated thoroughly in the statistical literature (see, e.g., the book by Scott [9]), and in conjunction with the current method this has been considered in the thesis [30]. We have also not considered how to choose the regularization parameter, and for this point we again refer to the statistical literature and [30] which furthermore contains a comparison of the numerical MAP approach discussed here with other numerical procedures, including histograms and kernel density estimators.

6.2.1. One-dimensional density. In one dimension, the domain is $T = [0, 1]$. Here we set

$$k(t, s) = k(s, t) = \frac{\sinh(\beta(1 - t)) \sinh(\beta s)}{\beta \sinh(\beta)}, \quad s \leq t.$$

By the addition theorem of sinh one can see that $\partial k(t, s)/\partial t$ has a jump of one at $t = s$. It follows that

$$-\frac{\partial^2 k(t, s)}{\partial t^2} + \beta^2 k(t, s) = \delta(t - s), \tag{31}$$

where δ is the Dirac distribution. As $k(t, s)$ is piecewise polynomial and continuous, it is in H^1 (see, for example, the textbook [24] by Braess), and furthermore

$k(0,s) = k(1,s) = 0$. It follows that $k(t,s)$ is the Green's function characterized by the differential equation (31), and consequently, it is the reproducing kernel of the norm defined by $\sqrt{\int_0^1 (u'(t))^2 + \beta^2 u(t)^2)\,dt}$. The structure parameter $\beta > 0$ plays the same role as the width in Gaussian kernels and, more generally, in kernel density estimators. It has to be specified and for our experiments we choose $\beta = 1$. We now set

$$k_1(t,s) = \frac{n}{\alpha} k(s,t).$$

This is the kernel of the Sobolev space $H = H_0^1[0,1]$ with the norm

$$\|u\|_H = \sqrt{\frac{\alpha}{n}\left(\int_0^1 (u'(t))^2\,dt + \beta^2 \int_0^1 u(t)^2\,dt\right)}.$$

That k_1, is indeed, a reproducing kernel follows from the fact that k is the reproducing kernel as shown above. It then follows that the functional j takes the form

$$j(u) = \frac{\alpha}{n}\int_0^1 (u'(t))^2\,dt + \frac{\alpha\beta^2}{n}\int_0^1 u(t)^2\,dt + \log\int_0^1 e^{u(t)}\,dt - \frac{1}{n}\sum_{i=1}^n u(t_i).$$

As the kernel $k(s,t)$ is continuous and piecewise C^2, it follows from the duality theory (see (20)) that the minimizer of j is continuous and piecewise C^2 as well. Then, $u \in H^{\frac{3}{2}-\epsilon}$ for any $\epsilon > 0$. As a consequence of the theory developed in the previous sections one sees that the piecewise linear approximation $u_h \in V_h$ is then of order $O(h^{3/2-\epsilon})$ in terms of the L_2-norm.

Now we study two special densities in more detail. First we consider the reconstruction of a normal density—which has been truncated to the interval $[0,1]$ and renormalized—and then the estimation of the density of a widely used data set. Since we now know the exact solutions, we choose the norms $\|u_h - u_{2h}\|_{L_q}$ as a substitute for the error norms. This is well justified because the errors have been shown to be of order $O(h^r)$ for some $r > 1$. For simplicity we will call these substitutes "errors" in the following.

In the first example the effect of the randomness of data is eliminated by replacing the term $\frac{1}{n}\sum_{i=1}^n u(t_i)$ in the expression for $j(u)$ by the limit for $n \to \infty$ which is $\int_0^1 f(t)u(t)\,dt$. We refer to this example as the "exact data" case. Using a similar duality argument as in (20) one can show that for exact data $u \in H^2[0,1]$, and so the approximation in the space of piecewise linear functions V_h has a L_2-error of the order $O(h^2)$. This is confirmed by the results in Table 1 which contains the L_1- and L_2-errors of u_h together with the ratios of the errors between the levels. As an example we have chosen the normal distribution with expectation $1/2$ and variance 0.05 truncated to the interval $[0,1]$ (and normalized to have integral one over $[0,1]$). Here we choose the regularization parameter[8] $\alpha = 0.0002$ and the kernel shape parameter $\beta = 1$. The approximated probability for level $l = 6$ (i.e., $h = 1/64$) and the exact normal distribution are plotted in Figure 1.

As a measure for the error or u_h we consider $\|u_h - u_{2h}\|_{L_p}$ for $p = 1, 2$. One can verify from Table 1 that the approximation is $O(h^2)$ accurate.

[8] Of course, in practice α and possibly β would be determined by a statistical procedure like cross validation.

TABLE 1
L_q-norms $\|u_h - u_{2h}\|_{L_q}$ for $q = 1, 2$, grid sizes $h = 2^{-l}$, and ratios $\|u_h - u_{2h}\|_{L_q} / \|u_{h/2} - u_h\|_{L_q}$ for normal distribution, exact data, smoothing parameter $\alpha = 0.0002$, and structure parameter $\beta = 1$.

level	L_1-error	ratio	L_2-error	ratio
2	0.5	–	0.62	–
3	0.2	2.47	0.27	2.31
4	$3.59 \cdot 10^{-2}$	5.62	$5.40 \cdot 10^{-2}$	4.96
5	$9.44 \cdot 10^{-3}$	3.80	$1.36 \cdot 10^{-2}$	3.99
6	$2.31 \cdot 10^{-3}$	4.09	$3.35 \cdot 10^{-3}$	4.05
7	$5.73 \cdot 10^{-4}$	4.03	$8.34 \cdot 10^{-4}$	4.02
8	$1.43 \cdot 10^{-4}$	4.01	$2.08 \cdot 10^{-4}$	4.01
9	$3.57 \cdot 10^{-5}$	4.00	$5.20 \cdot 10^{-5}$	4.00
10	$8.92 \cdot 10^{-6}$	4.00	$1.30 \cdot 10^{-5}$	4.00
11	$2.23 \cdot 10^{-6}$	4.00	$3.25 \cdot 10^{-6}$	4.00
12	$5.57 \cdot 10^{-7}$	4.00	$8.13 \cdot 10^{-7}$	4.00

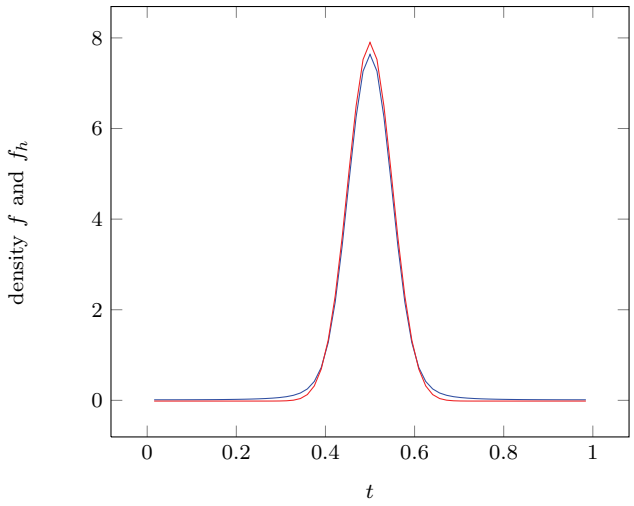

FIG. 1. *Gaussian density with variance* 0.05 *(interpolated to grid with size h) and its approximation using "exact data" and grid size $h = 1/64$.*

Now we consider the same Gaussian density with mean 1/2 and variance 0.05 and draw a random sample $t_1, \ldots, t_n$ with $n = 16$ to determine u. For our estimator we choose again $\beta = 1$ and $\alpha = 0.002$. The resulting errors $\|u_h - u_{2h}\|_{L_q}$ are given in Table 2. The average error reduction obtained through doubling the grid size is in theory $\sqrt{8} \approx 2.83$. From the results of Table 2 one gets on average over all the levels a reduction factor of 3.4 for the L_1-norm and of 2.7 for the L_2-norm. Note that the observed factor depends on the actual sample size and it increases with the sample size. For example with 100,000 data points one gets average rates of 3.9 and 3.8 for the L_1- and L_2-norms, respectively. This is to be expected as in the limit for an infinite number of points we know that the reduction factor must be 4.0 in theory. A slight improvement of the reduction rate is also obtained when choosing a larger α; for example, in the above example we got factors 3.7 for the L_1-norm and 2.9 for the L_2-norm when choosing $\alpha = 0.01$ (instead of the $\alpha = 0.002$ used earlier). Of course, both, i.e., larger amounts of data and a larger α, will lead to smoother estimates; see Figure 2 for the actual estimates, the effect of α, and the number of data points.

TABLE 2
L_q-norms $\|u_h - u_{2h}\|_{L_q}$ for $q = 1, 2$, grid sizes $h = 2^{-l}$, and ratios $\|u_h - u_{2h}\|_{L_q} / \|u_{h/2} - u_h\|_{L_q}$ for normal distribution, 16 samples, smoothing parameter $\alpha = 0.002$, and structure parameter $\beta = 1$.

level	L_1-error	ratio	L_2-error	ratio
2	0.38	–	0.41	–
3	$7.36 \cdot 10^{-2}$	5.15	$8.92 \cdot 10^{-2}$	4.62
4	$4.32 \cdot 10^{-2}$	1.71	$6.44 \cdot 10^{-2}$	1.39
5	$1.36 \cdot 10^{-2}$	3.17	$2.35 \cdot 10^{-2}$	2.74
6	$5.43 \cdot 10^{-3}$	2.51	$1.00 \cdot 10^{-2}$	2.34
7	$1.78 \cdot 10^{-3}$	3.05	$4.17 \cdot 10^{-3}$	2.40
8	$4.29 \cdot 10^{-4}$	4.16	$1.39 \cdot 10^{-3}$	3.00
9	$1.19 \cdot 10^{-4}$	3.61	$5.21 \cdot 10^{-4}$	2.66
10	$2.72 \cdot 10^{-5}$	4.37	$1.53 \cdot 10^{-4}$	3.40
11	$9.57 \cdot 10^{-6}$	2.84	$8.08 \cdot 10^{-5}$	1.90
12	$1.70 \cdot 10^{-6}$	5.63	$2.07 \cdot 10^{-5}$	3.90

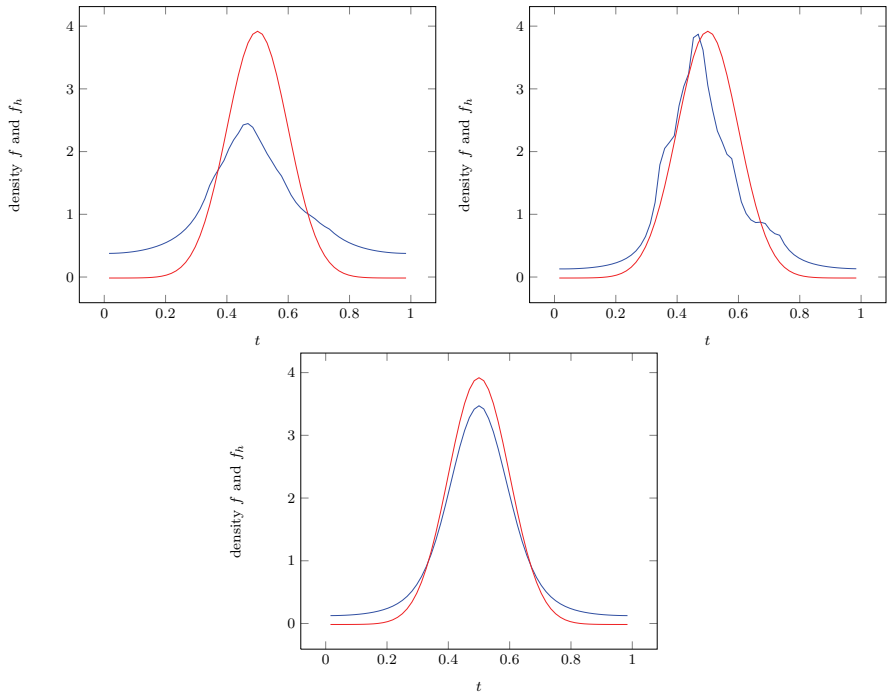

FIG. 2. *Gaussian density with variance* 0.05 *(interpolated to grid with size h) and its approximation using grid size $h = 1/64$. Top left: 16 data points and $\alpha = 0.01$; top right: 16 data points and $\alpha = 0.002$; bottom: 100,000 data points, $\alpha = 0.002$, and structure parameter $\beta = 1$.*

So far we have considered a very simple synthetic example. Now we will look at the estimation of the density for a well-known data set, the "Old Faithful" data. This data set contains 272 observations of the eruption times of the Old Faithful Geyser in the United States. It is arguably one of the most widely used data sets to illustrate the performance of density estimators [28, 9] and contains multiple modes; see [31]. In Table 3 one clearly sees the $O(h^{3/2})$ convergence behavior: The reduction rate was on average 3.1 and 2.6 for the L_1-norm and the L_2-norm, respectively. The resulting density again for $h = 1/64$ can be seen in Figure 3.

TABLE 3
L_q-norms $\|u_h - u_{2h}\|_{L_q}$ for $q = 1, 2$, grid sizes $h = 2^{-l}$, and ratios $\|u_h - u_{2h}\|_{L_q} / \|u_{h/2} - u_h\|_{L_q}$ for Old Faithful data set with 272 samples, smoothing parameter $\alpha = 0.0005$, and structure parameter $\beta = 1$.

level	L_1-error	ratio	L_2-error	ratio
2	1	–	1.12	–
3	0.38	2.62	0.43	2.61
4	$7.23 \cdot 10^{-2}$	5.26	$8.71 \cdot 10^{-2}$	4.95
5	$2.14 \cdot 10^{-2}$	3.37	$3.28 \cdot 10^{-2}$	2.66
6	$7.62 \cdot 10^{-3}$	2.81	$1.16 \cdot 10^{-2}$	2.82
7	$2.47 \cdot 10^{-3}$	3.09	$3.98 \cdot 10^{-3}$	2.92
8	$1.07 \cdot 10^{-3}$	2.30	$2.14 \cdot 10^{-3}$	1.86
9	$3.52 \cdot 10^{-4}$	3.04	$7.72 \cdot 10^{-4}$	2.77
10	$1.10 \cdot 10^{-4}$	3.21	$3.39 \cdot 10^{-4}$	2.28
11	$3.00 \cdot 10^{-5}$	3.67	$1.14 \cdot 10^{-4}$	2.98
12	$7.96 \cdot 10^{-6}$	3.76	$4.54 \cdot 10^{-5}$	2.51

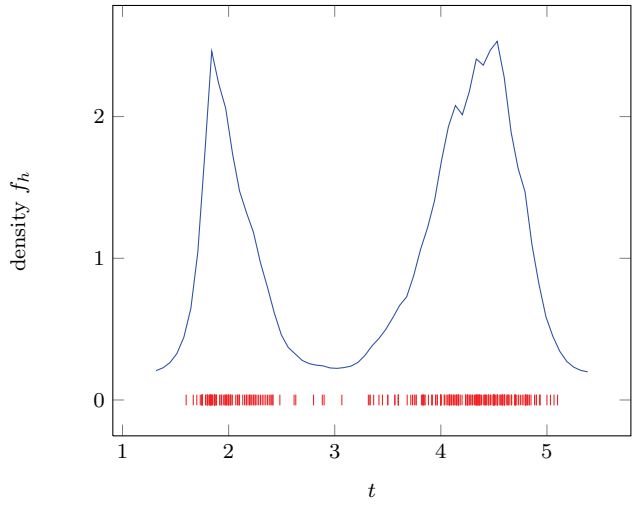

FIG. 3. *Old Faithful data with grid size $h = 1/64$, $\alpha = 0.0005$, and structure parameter $\beta = 1$.*

6.2.2. Two-dimensional density. For two dimensions, we use the product kernel

$$k_2(t, s; q, r) = \frac{n}{\alpha} k(t, q) k(s, r),$$

which corresponds to a prior that will favor independent variables if no information is available. This makes sense as the dependence structure should originate from the data (unless it is known a priori, of course). The domain is $T = [0, 1]^2$, and one can see that the kernel defines as H the mixed Sobolev space $H = H_0^1[0, 1] \times H_0^1[0, 1]$. The norm which is also derived from this kernel is

$$\|u\|_H = \sqrt{\frac{\alpha}{n} \left(\int_0^1 \int_0^1 u_{t,s}(t, s)^2 + \beta^2 u_t(t, s)^2 + \beta^2 u_s(t, s)^2 + \beta^4 u(t, s)^2 \, dt \, ds \right)}.$$

As in the one-dimensional case, other norms are possible, and one might wish to choose a smoother kernel. However, this kernel has been selected here because it

TABLE 4
L_q-norms $\|u_h - u_{2h}\|_{L_q}$ for $q = 1, 2$, grid sizes $h = 2^{-l}$, and ratios $\|u_h - u_{2h}\|_{L_q} / \|u_{h/2} - u_h\|_{L_q}$ for normal distribution, 16 samples, smoothing parameter $\alpha = 0.0002$, and structure parameter $\beta = 1$.

level	L_1-error	ratio	$L2$-error	ratio
2	0.26	–	0.31	–
3	$9.16 \cdot 10^{-2}$	2.82	0.13	2.31
4	$2.19 \cdot 10^{-2}$	4.18	$3.56 \cdot 10^{-2}$	3.71
5	$7.97 \cdot 10^{-3}$	2.75	$1.51 \cdot 10^{-2}$	2.36
6	$3.13 \cdot 10^{-3}$	2.55	$6.09 \cdot 10^{-3}$	2.47
7	$9.80 \cdot 10^{-4}$	3.19	$2.18 \cdot 10^{-3}$	2.80
8	$3.50 \cdot 10^{-4}$	2.80	$8.18 \cdot 10^{-4}$	2.66

TABLE 5
L_q-norms $\|u_h - u_{2h}\|_{L_q}$ for $q = 1, 2$, grid sizes $h = 2^{-l}$, and ratios $\|u_h - u_{2h}\|_{L_q} / \|u_{h/2} - u_h\|_{L_q}$ for two-dimensional lipid data set, $\alpha = 0.00005$, and structure parameter $\beta = 1$.

level	L_1-error	ratio	$L2$-error	ratio
2	1.33	–	1.56	–
3	0.22	6.14	0.28	5.55
4	$3.14 \cdot 10^{-2}$	6.90	$5.05 \cdot 10^{-2}$	5.57
5	$1.14 \cdot 10^{-2}$	2.76	$1.97 \cdot 10^{-2}$	2.57
6	$3.72 \cdot 10^{-3}$	3.06	$7.96 \cdot 10^{-3}$	2.47
7	$1.30 \cdot 10^{-3}$	2.85	$2.56 \cdot 10^{-3}$	3.11
8	$5.48 \cdot 10^{-4}$	2.38	$1.06 \cdot 10^{-3}$	2.42

clearly exhibits numerical errors in the computations. For a kernel which relates to H_0^2, see [30]. In our case, it holds that $k_2 \in H^{3/2-\epsilon}[0,1] \times H^{3/2-\epsilon}[0,1]$ for all $\epsilon > 0$, and it follows as in the one-dimensional case that the solution u has the same regularity as the kernel. Consequently, one expects to see an $O(h^{3/2})$ error. This can be observed in Table 4 for the case of a (truncated) normal distribution with variance 0.1. Here we have chosen only 50 data points and a relatively small $\alpha = 0.0002$. We found that the asymptotics (in the number of data points) which leads to an approximate $O(h^2)$ error starts earlier in two dimensions. The average reduction factor of the error is here 3.0 and 2.7 for the L_1- and L_2-norms, respectively.

As the next example we consider the lipid data set from [9]. It contains observations of the cholesterol and triglyceride values of 371 male patients. The density estimation should provide insight into the structure of the population; in particular, it should show whether these values are related to heart disease. The errors are given in Table 5. They again show clearly a convergence rate of substantially more than $O(h)$ but slightly less than $O(h^2)$. The observed reduction factor of the error is fluctuating and appears to be asymptotically decreasing. An average value, computed as the geometric mean of the reduction factors over the full range, is 3.7 for the L_1-norm and 3.4 for the L_2-norm. Note here that a larger than usual reduction of the error occurred right at the beginning for the coarsest grids, i.e., in the preasymptotic region.

6.2.3. Three-dimensional density. For three-dimensional density estimation we considered the landsat data set from [9]. This example was earlier discussed in [32] where the authors point out some of the statistical difficulties for density estimation in three and more dimensions. The data which are described in detail in [32, 9] are based on remote sensing data measured over North Dakota in 1977 and contain the following three variables: the time of maximum greenness, the ripening period of the crop, and the value of the maximum greenness. The variables were obtained by fitting

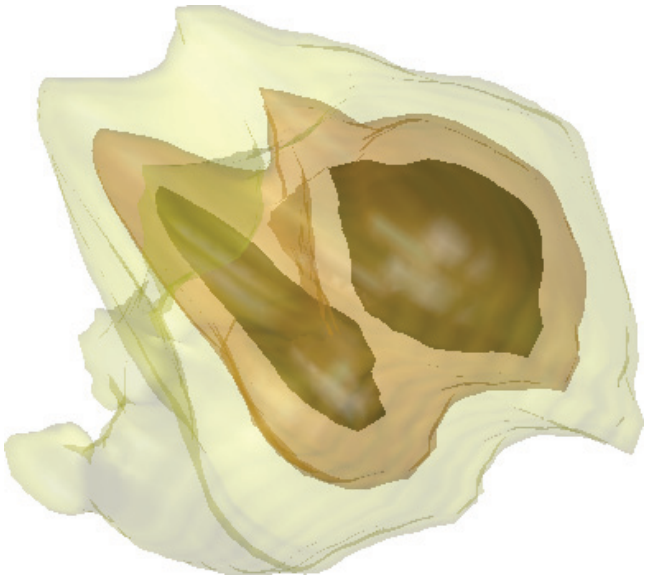

FIG. 4. *Three-dimensional (rotated) view of density $f_h, h = 2^{-7}$, landsat data set, and isosurfaces for three different values.*

TABLE 6

L_q-norms $\|u_h - u_{2h}\|_{L_q}$ for $q = 1, 2$, grid sizes $h = 2^{-l}$, and ratios $\|u_h - u_{2h}\|_{L_q} / \|u_{h/2} - u_h\|_{L_q}$ for three-dimensional landsat data set, $\alpha = 0.00001$, and structure parameter $\beta = 1$.

level	L_1-error	ratio	$L2$-error	ratio
2	0.29	–	0.42	–
3	$7.13 \cdot 10^{-2}$	4.05	0.13	3.10
4	$1.57 \cdot 10^{-2}$	4.54	$3.32 \cdot 10^{-2}$	4.05
5	$3.79 \cdot 10^{-3}$	4.15	$8.64 \cdot 10^{-3}$	3.84
6	$1.03 \cdot 10^{-3}$	3.67	$2.68 \cdot 10^{-3}$	3.22

a growth model to time-spatial observations of reflectance intensities. It is difficult to see any structure from lower-dimensional projections of the data; however, one can clearly distinguish two main clusters (which correspond to different crops) in the three-dimensional density. See Figure 4 which was obtained with our code and the open-dx visualization software.

The landsat data set contains 22932 observations of three variables of which we selected a subregion with 22513 observations which contains "most of the action" but was substantially smaller than the original region. As prior we have chosen a tensor product of the one-dimensional priors. The kernel is then

$$k_3(t_1, t_2, t_3; s_1, s_2, s_3) = \frac{n}{\alpha} k(t_1, s_1) k(t_2, s_2) k(t_3, s_3)$$

and the H-norm is

$$\|u\|_H = \sqrt{\frac{\alpha}{n} \left(\iiint_{[0,1]^3} \left(u_{t_1 t_2 t_3}^2 + \beta^2 \sum_{i<j} u_{t_i t_j}^2 + \beta^4 \sum_{i=1}^{3} u_{t_i}^2 + \beta^6 u^2 \right) dt_1 dt_2 dt_3 \right)}.$$

The associated space turns out to be the mixed space $H_0[0,1] \times H_0[0,1] \times H_0[0,1]$.

Table 6 shows the L_1- and L_2-errors for the case of $\alpha = 0.0002$. Similar errors were also obtained for the case $\alpha = 0.000001$. Here we observe on average a decrease by 4.1 of the L_1-error and by 3.5 of the L_2-error. This high convergence rate is remarkable; however, one commonly observes that the preasymptotic region extends for higher-dimensional data over several orders of magnitudes of h. Because of complexity reasons we are presently limited to six levels of refinement in three dimensions and can not reach the asymptotic range here.

7. Concluding remarks. In this article, we derived a variational problem characterizing the density estimator defined by the maximum a posteriori method with Gaussian process priors. We demonstrated that MAP estimators are penalized maximum likelihood estimators using the Cameron–Martin theory of stochastic processes. We then showed that this problem is well posed and can be solved with Newton's method. Furthermore, we proposed a Newton–Galerkin approach for its solution and discussed the computational performance of this approach for discretizations on regular "full grids" for reasons of simplicity.

The MAP density estimators belong to the class of penalized maximum likelihood estimators. Other big classes include the histograms and kernel methods; see the books by Scott [9], Silverman [28], and Tapia and Thompson [29] and a rich literature for a further statistical discussion of comparative merits of these methods. Arguably, kernel density estimators are among the most popular approaches due to their simplicity of implementation. However, a naive implementation of kernel methods—especially in higher dimensions—pose a complexity problem: It will require visiting all data points for the evaluation at any point which is extremely costly. This problem can be circumvented by a further discretization [28], but we are not aware of any studies of the associated errors in the literature. In contrast to that, we introduced here a new class of methods where the discretization is an integral component of the approximation and we were able to use standard numerical estimation techniques to get error bounds. Similar methods are well known to be highly efficient in solving partial differential equations. We have shown that they can be adopted with the same efficiency to density estimation. A difference to the solution of partial differential equations is the occurrence of point evaluations on the right-hand side of the equations which is akin to computing Green's functions. This is why reproducing kernels now play an important role.

So far, the curse of dimension limits us to three dimensions. To this end, instead of the "full grid" discretization other subspace constructions may also be chosen like sparse grid spaces/hyperbolic crosses and generalized sparse grids. This way the dimension of the approximation space, and thus the involved cost, is often substantially reduced without compromising the rate of the approximation error and the curse of dimensionality can be broken, at least to some extent. In the future we plan to investigate several techniques for the solution of high-dimensional problems including Opticom [33, 34] and other variants of the sparse grid combination technique [35, 16].

The extension of the current methods to other types of priors would require the availability of the equivalent of the Cameron–Martin formula for the Radon–Nikodým derivative of the translated measure with respect to the original measure (defined by the prior) and will have to be the subject of future work.

Note finally that, while we discussed the new method for density estimation, our approach can also be used for conditional density estimation; see [7]. In this case one is given a sequence of data pairs $(t_1, y_1), \ldots, (t_n, y_n)$. The values of y_k are random and the values of t_k can be either random or fixed. The conditional probability distribution

is defined by a conditional density $f(y \mid t)$ which shall be found from the data. From this conditional density one can then obtain Bayesian (optimal) estimators of the value of y given t which leads to various *classification* methods in the case where y is discrete and *regression* techniques when y is continuous.

REFERENCES

[1] E. Parzen, *An approach to time series analysis*, Ann. Math. Statist., 32 (1961), pp. 951–989.

[2] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley and Sons, New York, 1973.

[3] I. Good and R. Gaskins, *Nonparametric roughness penalties for probability densities*, Biometrika, 58 (1971), pp. 255–277.

[4] T. Leonard, *Density estimation, stochastic processes and prior information*, J. Roy. Statist. Soc. Ser. B, 40 (1978), pp. 113–146.

[5] R. H. Cameron and W. T. Martin, *Transformations of Wiener integrals under translations*, Ann. of Math. (2), 45 (1944), pp. 386–396.

[6] V. I. Bogachev, *Gaussian measures*, Mathematical Surveys and Monographs 62, American Mathematical Society, Providence, RI (1998).

[7] M. Hegland, *Approximate maximum a posteriori with Gaussian process priors*, Constr. Approx., 26 (2007), pp. 205–224.

[8] M. G. Bulmer, *Principles of Statistics*, 2nd ed., Dover, New York, 1979.

[9] D. W. Scott, *Multivariate density estimation*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley and Sons, New York, 1992.

[10] R. Fisher, *Two new properties of mathematical likelihood*, Proc. Roy. Soc. A, 144 (1934), pp. 285–307.

[11] Y. Yamasaki, *Measures on infinite-dimensional spaces*, Series in Pure Mathematics 5, World Scientific, Singapore, 1985.

[12] G. Wahba, *Spline models for observational data*, CBMS-NSF Regional Conference Series in Applied Mathematics 59, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

[13] M. Seeger, *Gaussian processes for machine learning*, Internat. J. Neural Syst., 14 (2004), pp. 69–106.

[14] R. J. Adler, *The geometry of random fields*, Wiley Series in Probability and Mathematical Statistics, John Wiley and Sons, New York, 1981.

[15] I. Ekeland and R. Témam, *Convex analysis and variational problems*, English ed., Classics in Applied Mathematics 28, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999.

[16] J. Garcke, *Maschinelles Lernen durch Funktionsrekonstruktion mit verallgemeinerten dünnen Gittern*, Doktorarbeit, Institut für Numerische Simulation, Universität Bonn, 2004.

[17] C. T. Kelley, *Solving nonlinear equations with Newton's method*, Fundamentals of Algorithms, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2003.

[18] P. Deuflhard and M. Weiser, *Global inexact Newton multilevel FEM for nonlinear elliptic problems*, Multigrid Methods V (Stuttgart, 1996), Lect. Notes in Comput. Sci. Engrg. 3, Springer, Berlin, 1998, pp. 71–89.

[19] P. Deuflhard, *Newton methods for nonlinear problems*, Affine invariance and adaptive algorithms Springer Series in Computational Mathematics 35, Springer-Verlag, Berlin, 2004.

[20] P. Deuflhard and M. Weiser, *Local inexact Newton multilevel FEM for nonlinear elliptic problems*, Computational science for the 21st century, M. O. Bristeau, G. Etgen, W. Fitzgibbon, J. L. Lions, J. Periaux, and M. Wheeler, eds., John Wiley and Sons, New York 1997, pp. 129–138.

[21] M. Griebel and S. Knapek, *Optimized tensor-product approximation spaces*, Constr. Approx., 16 (2000), pp. 525–540.

[22] S. Knapek, *Approximation und Kompression mit Tensorprodukt-Multiskalenräumen*, Doktorarbeit, Universität Bonn, 2000.

[23] M. Griebel and S. Knapek, *Optimized general sparse grid approximation spaces for operator equations*, Math. Comput. (2008), submitted. Also available as SFB611 preprint 402.

[24] D. Braess, *Finite Elements*, 3rd ed., Cambridge University Press, Cambridge, 2007. Translated from the German by Larry L. Schumaker.

[25] R. A. DeVore and B. J. Lucier, *Wavelets*, Acta Numer., Cambridge University Press, Cambridge, 1992, pp. 1–56.

[26] P. Oswald, *Multilevel Finite Element Approximation*, Teubner Skripten zur Numerik, Teubner, Stuttgart, 1994.
[27] M. Griebel and P. Oswald, *Tensor product type subspace splitting and multilevel iterative methods for anisotropic problems*, Adv. Comput. Math., 4 (1995), pp. 171–206.
[28] B. W. Silverman, *Density estimation for statistics and data analysis*, Monographs on Statistics and Applied Probability, Chapman & Hall, London, 1986.
[29] R. A. Tapia and J. R. Thompson, *Nonparametric probability density estimation*, Johns Hopkins Series in the Mathematical Sciences, Vol. 1, Johns Hopkins University Press, Baltimore, MD, 1978.
[30] P. Hahnen, *Nichtlineare numerische Verfahren zur multivariaten Dichteschätzung*. Master's thesis, Institut für numerische Simulation, Universität Bonn, 2006.
[31] A. Azzalini and A. W. Bowman, *A look at some data on the Old Faithful Geyser*, Appl. Statist., 39 (1990), pp. 357–365.
[32] D. Scott and J. Thompson, *Probability density estimation in higher dimensions*, Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface, Amsterdam, North-Holland, 1983, pp. 173–179.
[33] J. Garcke, *Regression with the optimised combination technique*, in Proceedings of the 23rd ICML '06, W. Cohen and A. Moore, eds., ACM Press, New York, 2006, pp. 321–328.
[34] M. Hegland, J. Garcke, and V. Challis, *The combination technique and some generalisations*, Linear Algebra Appl., 420 (2007), pp. 249–275.
[35] J. Garcke, M. Griebel, and M. Thess, *Data mining with sparse grids*, Comput. 67 (2001), pp. 225–253.