Variational problems in machine learning and their solution with finite elements

M. Hegland* M. Griebel†

April 2007

Abstract

Many machine learning problems deal with the estimation of conditional probabilities $p(y \mid x)$ from data $(x_1, y_i), \ldots, (x_n, y_n)$. This includes classification, regression and density estimation. Given a prior for $p(y \mid x)$ the maximum a-posteriori method estimates $p(y \mid x)$ as the most likely probability given the data. This principle can be formulated rigorously using the Cameron–Martin theory of stochastic processes and allows a variational characterisation of the estimator. The resulting nonlinear Galerkin equations are solved numerically. Convexity and total positivity lead to existence, uniqueness and error bounds. For machine learning problems dealing with large numbers of features we suggest the use of sparse grid approximations.

Contents

*Statistical Machine Learning Program, NICTA and Centre for Mathematics and its Applications, ANU, Canberra, AUSTRALIA mailto:markus.hegland@anu.edu.au

†Institut für Numerische Simulation, University of Bonn, GERMANY mailto:griebel@ins.uni-bonn.de

1 Introduction

Density estimation is one of the most versatile tools for data exploration of large and complex data sets. From the density one can determine characterising features like moments, modes or clusters. Conditional distributions are used for classification and regression but they are more general as they allow for multiple classes or regressors. This is useful when information is missing which is typical for data mining. There are three main classes of density estimators [6]: Histograms, kernel density estimators and (penalised) maximum likelihood estimators. Histograms are widely used for simple, low-dimensional explorations. Kernel density estimators are more precise and are a smoothed version of the empirical distribution. Using a kernel function ρ they take the form $p(y) = \frac{1}{nh} \sum_{i=1}^{n} \rho\left(\frac{y-y_i}{h}\right)$ where y_i are the observed data of the variable y and h is the bandwidth of the estimator. The computational effort is concentrated in the evaluation of the kernel estimator and may require $O(n)$ operations per evaluation of one value $p(y)$. The third family of estimators is based on finding a density $p(y)$ for which the likelihood $L(p|y_1, \ldots, y_n) = \prod_{i=1}^{n} p(y_i)$ of the data is large. In most cases this leads to an ill-posed problem and thus an additional regularising penalty functional is used. The maximum aposteriori estimators considered in this article are an example of a penalised maximum likelihood estimator. By penalising the logarithm of p one can avoid negative values of $p(y)$ which otherwise may occur for this method. The approach developed here is closely related to kernel methods used for classification and regression [5]. It generalises that approach in two important aspects: First, it estimates conditional densities and second, it considers infinite-dimensional function spaces.

The remainder of this article is organized as follows: In the next section, we discuss the maximum a-posteriori method and exponential function families. The main challenge for this approach is to overcome the difficulty that the prior of an infinite dimensional function class does not possess a density. In the third section we derive approximate solutions to the resulting variational problem using a Ritz method. We also show that these problems have a unique solution and we provide error bounds. In the fourth section we illustrate the approach by some applications from the area of density estimation. We discuss the numerical solution approach via Newtons method, a discretisation by finite elements for continuous problems and, for a two-dimensional continuous problem, an application of the sparse grid combination method. Finally we close with some concluding remarks.

2 The maximum a-posteriori method with exponential families

The maximum a-posteriori method (MAP) introduces a prior probability over the set of all models and interprets the likelihood as the conditional probability of the data given the model. From this, the "posterior" probability of models is defined as the conditional probability of the model given the data. The estimated model is then the maximal mode or the "most likely" model of the posterior. A difficulty occurs with this approach when the model space is infinite dimensional. In the following, we adopt the approach suggested in [4] to address this difficulty. In our particular case the model is defined by a set X, a set Y with finite measure and a function $u \in \mathbb{R}^{X\times Y}$ such that $\int_Y \exp(u(x,y))\,dy < \infty$. The conditional probability distribution is

$$p(y \mid x) = \frac{\exp(u(x,y))}{\int_Y \exp(u(x,y))\,dy}. \tag{1}$$

Let μ denote a probability measure over the set $\mathbb{R}^{X\times Y}$. We assume here that the set of functions u for which $\exp(u(x,\cdot))$ is in $L_1(Y)$ is a μ-measurable set. One defines for an appropriate function h the translated measure as $\mu_h(A) = \mu(h + A)$ where $h + A$ is the set of all functions $u + h$ where $u \in A$. If μ_h is absolutely continuous with respect to μ one can introduce the Radon-Nikodym derivative

$$r_h = \frac{d\mu_h}{d\mu}$$

which is an element of $L_1(\mu)$. Using the composition properties of shifts and the Radon-Nikodym derivative one can then introduce a function ρ such that $r_h(u) = \rho(u+h)/\rho(u)$ and where $\rho(0) = 1$. It turns out that for finite dimensional function spaces this coincides with the density up to a normalisation factor. A mode of a distribution is then characterised as a maximum of ρ. Note, however, that ρ is in general only defined on a subset of the set of all functions so that we can only really consider modes in that subset. But it turns out that this subset is flexible enough for our applications. Note that while ρ generalises the concept of a density it is different from the concept of an "improper probability density" as sometimes used in Bayesian statistics. There the improper density is typically a density of a non-finite measure whereas in our case ρ is not a density at all but describes a finite measure.

The prior we consider here is a Gaussian prior with zero expectation. This can be characterised by the covariance operator and leads to the Cameron-Martin space [1] which is a (reproducing kernel) Hilbert space H with norm $\|\cdot\|_H$. One can show that the function ρ in this case is $\rho(u) = \exp(-\|u\|_H^2)$.

One can also determine the function ρ for the a-posteriori distribution using the conditional likelihood and the exponential family from equation (1). One gets in this case

$$\rho(u) = \exp\left(-\|u\|_H^2 + \sum_{i=1}^n u(x_i, y_i) - \sum_{i=1}^n \log\left(\int_Y \exp\left(u\left(x_i, y\right)\right)\, dy\right)\right).$$

More details on the derivation can be found in a recent paper of the first author [4]. The MAP estimator u is then the maximum of this function or, equivalently, $u = \operatorname{argmin}_{v\in H} J(v)$ where

$$J(v) = \|v\|_H^2 + \sum_{i=1}^n \log(\int_Y \exp(v(x_i, y))\, dy) - \sum_{i=1}^n v(x_i, y_i).$$

3 The variational problem and the Ritz method

The maximum a-posteriori method leads to a variational characterisation of the function u. In particular, let the *log partition function* at point $x \in X$ be defined as

$$\phi(v, x) = \log\left(\int_Y \exp(v(x, y))\, dy\right),$$

and let a nonlinear functional F and a linear functional b be defined by

$$F(v) = \|v\|_H^2 + \sum_{i=1}^n \phi(v, x_i) \quad \text{and} \quad \langle b, v\rangle = \sum_{i=1}^n v(x_i, y_i).$$

It follows then that $J(v) = F(v) - \langle b, v\rangle$. If F' denotes the Gateaux derivative of F then the minimizer u of J satisfies $F'(u) = b$, or $\langle F'(u), v\rangle = \langle b, v\rangle$ for all $v \in H$.

For the numerical solution of this variational problem we consider the Ritz method. Let $V_n \subset H$ for $n = 0, 1, 2, \ldots$ be a sequence of linear finite-dimensional spaces. The (Rayleigh–)Ritz approximant in space V_n is then

$$u_n = \operatorname{argmin}_{v\in V_n} J(v) \tag{2}$$

and it satisfies the Ritz–Galerkin equations

$$\langle F'(u_n), v\rangle = \langle b, v\rangle, \quad v \in V_n. \tag{3}$$

In order to show well-posedness of the occurring problems and the convergence of u_n to the exact solution u we first establish some properties of F.

Recall that $p(y \mid x) = \exp(u(x,y))/\int_Y \exp(u(x,z))\,dz$. One can easily show that the Gateaux derivative $\phi_u(u,x)$ is then

$$\langle \phi_u(u,x),\, v\rangle = \int_Y p(y \mid x) v(x,y)\,dy$$

and from this one sees that the Gateaux derivative F' satisfies:

$$\langle F'(u),\, v\rangle = 2(u,v)_H + \sum_{i=1}^{n} \int_Y p(y \mid x_i) v(x_i,y)\,dy, \quad v \in H.$$

The second Gateaux derivative is a linear map $\phi_{u,u}(u,x) : H \to H^*$ and one gets

$$\begin{aligned}\langle \phi_{uu}(u,x)v,\, w\rangle = \quad & \int_Y p(y \mid x) v(x,y) w(x,y)\,dy \\ & - \int_Y p(y \mid x) v(x,y) dy \int_Y p(y \mid x) w(x,y)\,dy.\end{aligned}$$

It follows that ϕ_u is the expectation operator and ϕ_{uu} thus gives the covariance of random variables $v, w \in H$, or, with $E(v \mid x) = \langle \phi_u(u,x),\, v\rangle$ one has $\langle \phi_{uu}(u,x)v,\, w\rangle = E((v - E(v \mid x))(w - E(w \mid x)) \mid x)$. For the second derivative F'' one gets

$$\langle F''(u)v,\, w\rangle = 2(v,w)_H + \sum_{i=1}^{n} E((v - E(v \mid x_i))(w - E(w \mid x_i)) \mid x_i).$$

From this, the following Lemma immediately follows:

Lemma 1. *$F''(u)$ is positive definite and F is strictly convex.*

Furthermore, one also has the following lemma:

Lemma 2. *Let Y have finite measure and let $H \subset C[X \times Y]$ be continuously embedded. Then $F(u)$ is coercive, i.e., $F(u)/\|u\|_H \to \infty$ as $\|u\|_H \to \infty$.*

Proof. One has $|\phi(u,x)| \le \|u\|_\infty |Y| \le C\|u\|_H |Y|$ and so the second term in F is at most of $O(\|u\|_H)$ and is thus dominated by the first term which is $\|u\|_H^2$. The coercivity follows directly from this observation. □

Furthermore, one sees that

$$\langle \phi_v(v,x) - \phi_w(w,x), v - w\rangle = \int_0^1 \langle \phi_{uu}((1-t)w + tv)(v-w), v - w\rangle\,dt \ge 0$$

as ϕ_{uu} is positive semidefinite. From this it follows that $\langle F'(v) - F'(w), v - w\rangle \ge 2\|v-w\|_H^2$ and, consequently, F' is uniformly monotone. One can then apply Theorem 42A from a book by Zeidler [7] and gets:

Theorem 1. *Let $V_1 \subset V_2 \subset \cdots \subset H$ be a sequence of subspaces such that for every $v \in H$ the projections $P_i v \in V_i$ converge to v. Then the variational problems in H and V_i have unique solutions u and u_i respectively and $u_i \to u$. Furthermore, there exists a constant $C > 0$ independent of u and u_i such that*

$$\|u - u_i\|_H^2 \leq \min_{v \in V_i} J(v) - \min_{v \in H} J(v) \leq C\|u - u_i\|_H^2.$$

Now, the remaining problem is to solve the unconstrained optimisation problems (2) for the u_i. This can be done with Newton's method. Using three examples, the associated numerical approach is explained in more detail in the next section.

4 Examples and numerical approach

Three simple examples shall illustrate the numerical approach for the case of density estimation, i.e., where X has only one value. In the first example Y has two discrete values whereas in the second and third example we consider a continuous one- and two-dimensional interval, respectively. Here, as an application, we choose the distribution of the observations of the eruption lengths and the intervals between observations for the Old Faithful geyser in the US Yellowstone park, see Silverman's book on density estimation [6] for more information about the data which we obtained from the R package. The code was implemented in Python.

Example 1

First we consider simulated data with $Y = \{0, 1\}$. Then u is a vector $u = (u_0, u_1)^T$ and the log partition function is $\phi(u) = \log(e^{u_0} + e^{u_1})$. As a prior we assume that the u_i are i.i.d. and normally distributed with expectation zero and variance σ^2. The scalar product of the Cameron-Martin space is then $(u, v)_H = \alpha(u_0 v_0 + u_1 v_1)$ where, to simplify notation, we introduce $\alpha = 1/(2\sigma^2)$. The reproducing kernel k_y (defined by $u(y) = (k_y, u)_H$) is then $k_y = \alpha^{-1}(1, 0)$ and $\alpha^{-1}(0, 1)$ for $y = 0$ and 1, respectively.

The data consists of n records $(y_0, y_1, \ldots, y_n)$ where $y_i \in \{0, 1\}$. Let n_0 and n_1 be the number of records with $y_i = 0$ and $y_i = 1$, respectively. The total number of records is then $n = n_0 + n_1$. The functional F is then $F(u) = \alpha(u_0^2 + u_1^2) + n \log(e^{u_0} + e^{u_1})$ and it follows that the derivative is $F'(u) = 2\alpha(u_0, u_1) + n(e^{u_0}, e^{u_1})/g(u)$ where $g(u) = e^{u_0} + e^{u_1}$. The "data part" b in the nonlinear functional $J(u) = F(u) - \langle b, u \rangle$ is $b = (n_0, n_1)$. The

minimizer of J satisfies the Euler equations

$$\begin{aligned} 2\alpha u_0 + ne^{u_0}/g(u) &= n_0, \\ 2\alpha u_1 + ne^{u_1}/g(u) &= n_1. \end{aligned}$$

The solution of the Euler equations is obtained using Newton's method

$$u^{(k+1)} = u^{(k)} - F''(u^{(k)})^{-1}(F'(u^{(k)}) - b)^T$$

where the Hessian of F is

$$F''(u) = 2\alpha \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + n \begin{bmatrix} p_0(1-p_0) & -p_0p_1 \\ -p_0p_1 & p_1(1-p_1) \end{bmatrix}$$

and $(p_0, p_1) = (e^{u_0}, e^{u_1})/g(u)$. In our example we have 100 observations where 80 are in class 0. Furthermore, we choose $\alpha = 0.0001$. (For a practical situation, one would determine α from the data using cross validation. See [6] for further details.)

The Euclidean norms $\|r_k\|$ of the residuals $r_k = F'(u^{(k)})$ for iterations $k = 0, \ldots, 4$ are 42, 0.21, $5.9 \cdot 10^{-4}$, $4.6 \cdot 10^{-9}$, $1.5 \cdot 10^{-14}$ which confirms the quadratic convergence of the Newton iteration.

Example 2

In the second case we consider the distribution of the eruption lengths (which are less than 10 Minutes) of the Old Faithful geyser so that $Y = [0, 10]$. The prior for u is a Gaussian process with expectation zero. We choose further

$$(u, v)_H = \alpha \int_0^{10} u'(y)v'(y)\, dy + \beta \int_0^{10} u(y)v(y)\, dy$$

and get from this the covariance operator of the prior as the reproducing kernel. Note that this is just the weak form which is associated to the differential operator $L = -\alpha d^2/dx^2 + \beta I$. The log partition function is $\phi(u) = \log \int_0^{10} \exp(u(x))dx$ and the chosen parameters are $\alpha = 0.1$ and $\beta = 0.001$. The data set has $n = 272$ observed values.

We use piecewise linear functions on a regular grid and employ natural boundary conditions. The log partition function is approximated by $\phi^h(u) = \log(h*(0.5\exp(u(0)) + \sum_{i=1}^{m} \exp(u(ih)) + 0.5\exp(u(10))))$ where $h = 1/(m-1)$. This corresponds to a Galerkin approximation with quadrature rule approximation of the integral. For the vector $\vec{u} = (u(0), u(h), \ldots, u(10))$ of values at the grid points one gets

$$\alpha A\vec{u} + \beta B\vec{u} + nhW\vec{p} = \vec{b}$$

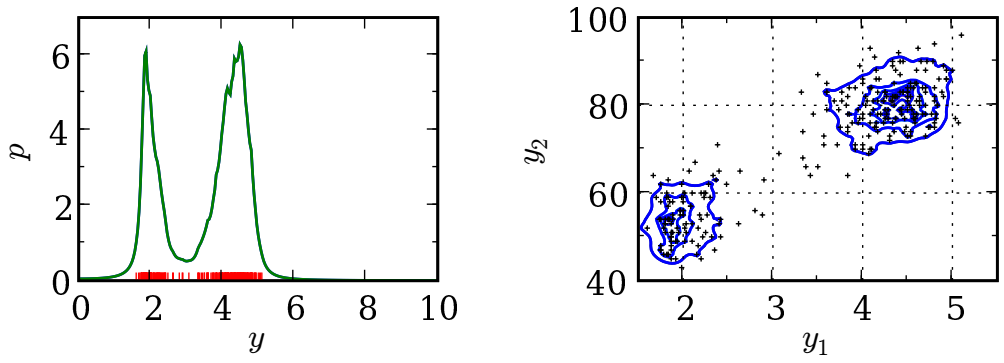

Figure 1: Old Faithful geyser: a) Distribution of eruption times, b) Joint distribution of eruption times and waiting times.

where A and B are the usual stiffness and mass matrices, $p_i = \exp(u(ih))/g(u)$, $g(u) = h * (0.5 \exp(u(0)) + \sum_{i=1}^{m} \exp(u(ih)) + 0.5 \exp(u(10)))$, W is the identity except that the first and the last diagonal elements are replaced by 0.5 and $b_i = \sum_{j=1}^{n} H(y^{(j)}/h - i)$ where H is the standard hat function. As in the previous example we use Newton's method and can again confirm its quadratic convergence. In order to assess the error of the Ritz approximation we compare the solution against the solution p^{4097} for $m = 4097$. One then gets $\|p^m - p^{4097}\|_1 = 1.10267,\ 0.92190,\ 0.78281,\ 0.31284,\ 0.10981,\ 0.04327,\ 0.01925,\ 0.00720,\ 0.00282,\ 0.00107$. The convergence appears to be between linear and quadratic in h. The less than quadratic convergence may relate to the fact that the solution is not C^2 but we will need to investigate this further. Figure 1 a) shows the density using our technique with $m = 257$ grid points.

Example 3

We now consider the case of the joint distribution of waiting times and eruption times of the Old Faithful geyser. There one has $Y[1.5, 5.5] \times [40, 100]$. The solution u is determined by a finite element method as in the previous example. Here, we focus on two sub-cases: First, a discretization on a uniform two-dimensional grid and, second, a discretization on a sparse grid [2] by means of the combination technique [3]. The scalar product of the Cameron–Martin space is

$$\|u\|_H^2 = \gamma \int_Y u_{y_1,y_2}^2 \, dy_1 \, dy_2 + \alpha \int_Y \left(u_{y_1}^2 + u_{y_2}^2\right) \, dy_1 \, dy_2 + \beta \int_Y u^2 \, dy_1 \, dy_2$$

which relates to the weak form of a second order differential operator. Again, we used normal boundary conditions.

Discretization on uniform two-dimensional grid

First, we employed a discretization with (bi-)linear hat functions on a uniform grid on Y. The resulting contour plot for a 129×129 grid with $\alpha = 2$, $\beta = 1000$ and $\gamma = 0.0001$ can be seen in Figure 1 b). As before, we found that Newton's method showed quadratic convergence. Comparing the results with the ones for a 257×257 grid one gets for $m = 17, 33, 65, 129$: $\|p^{257\times257} - p^{m\times m}\|_1 = 0.067,\ 0.033,\ 0.0142,\ 0.0054$ which is consistent with the results from the previous example.

Sparse grid discretization and combination technique

Finally, we used the combination technique to obtain a sparse grid approximation to the solution. This way, only $O(m \log m)$ degrees of freedom are involved instead of $O(m^2)$ on a uniform grid. Such a sparse grid approach can cope with the curse of dimensionality, at least to some extent, when it comes to higher dimensional problems. For a thorough discussion of sparse grids and the combination technique see the papers by Griebel and collaborators [2, 3]. The basic idea is that the sparse grid solution is put together from the solutions obtained on certain uniform grids albeit with different mesh sizes (m_1, m_2) for different coodinate directions similar in spirit to a multivariate extrapolation method.

The combination technique approximation u^{CT} is of the form

$$u^{\mathrm{CT}}(y_1, y_2) = \sum_{(m_1,m_2)\in I^{\mathrm{fine}}} u^{(m_1,m_2)}(y_1, y_2) - \sum_{(m_1,m_2)\in I^{\mathrm{inter}}} u^{(m_1,m_2)}(y_1, y_2).$$

The sets I^{fine} and I^{inter} are the sets of grid sizes (m_1, m_2) of the finest grids generating the sparse grid and their intersections, respectively, for details see the paper where the combination technique was introduced [3]. In Table 1 we give the L_1 error norms together with the finest regular grids which span the associated sparse grids.

We clearly see from this result the effectiveness of the sparse grid approximation. The error on a sparse grid nearly behaves like that on a uniform full grid whereas the degrees of freedom are substantially reduced.

Table 1: L_1 errors of combination technique approximants.

finest grids	L_1 error
(257,3),(129,5),(65,9),(33,17),(17,33),(9,65),(5,129),(3,257)	0.114
(257,5),(129,9),(65,17),(33,33),(17,65),(9,129),(5,257)	0.058
(257,9),(129,17),(65,33),(33,65),(17,129),(9,257)	0.020
(257,9),(129,17),(65,33),(33,65),(17,129),(9,257)	0.0075
(257,17),(129,33),(65,65),(33,129),(17,257)	0.0029
(257,65),(129,129),(65,257)	0.00104

5 Concluding remarks

In this article we discussed the maximum aposteriori method with exponential families for machine learning problems and their formulation as variational problems. The Ritz method then allows for numerical approximations using finite element subspaces. We have shown uniqueness and error bounds of the corresponding solutions. For the case of density estimation, we presented the basic numerical approach to the solution of the associated non-linear problems.

For the two-dimensional example with continuous Y we applied the combination method, which works on a sparse grid, as well as a discretization on a uniform grid. This opens the way to an efficient treatment of higher dimensional problems. Here, however, additional numerical experiments still have to be carried out.

In addition to the presented examples, further studies for simulated data were done and the approach was compared with other existing techniques for density estimation including a histogram method, a kernel density estimator and another maximum likelihood approach. It could be shown that in all cases the approach discussed here outperformed the traditional methods especially for multimodal problems.

References

[1] V. I. Bogachev. *Gaussian measures*, volume 62 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1998.

[2] H.-J. Bungartz and M. Griebel. *Acta Numerica*, volume 13, chapter Sparse Grids, pages 1–123. Cambridge University Press, 2004.

[3] M. Griebel, M. Schneider, and C. Zenger. A combination technique for the solution of sparse grid problems. In *Iterative methods in linear algebra (Brussels, 1991)*, pages 263–281. North–Holland, Amsterdam, 1992.

[4] M. Hegland. Approximate maximum a posteriori with Gaussian process priors. *Constructive Approximation*, April 2007. Electronic pre-publication in Online First, DOI 10.1007/s00365-006-0661-4, URL http://www.springerlink.com/content/985521n52h7310x2/.

[5] B. Schölkopf and A. Smola. *Learning with Kernels, Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, Cambridge, MA, 2002.

[6] B. W. Silverman. *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986.

[7] E. Zeidler. *Nonlinear functional analysis and its applications. III*. Springer–Verlag, New York, 1985. Variational methods and optimization, Translated from the German by Leo F. Boron.