Bayesian Field Theory: From likelihood fields to hyperfields

Jörg C. Lemm
WGZ-Bank, Düsseldorf

Abstract. Bayesian models are based on the combination of a likelihood model and a prior model. The nonparametric Bayesian models discussed in this chapter combine the likelihood models of density estimation, regression, and inverse quantum theory, expressed in terms of 'likelihood fields', with nonparametric prior models for such fields. Starting from Gaussian process priors, which, for example, are able to implement approximate symmetries for the likelihood fields, those are made more flexible by introducing hyperparameters and, generalized to a hyperparameter valued function, by hyperfields. In a joint Maximum A Posteriori approximation this results in coupled equations for likelihood fields and hyperfields. An example of a useful application of hyperfields is the adaption of the mean function of a Gaussian process prior.

1 Introduction

It is well known that empirical learning can not be based on observational data alone but does always also depend on *a priori* information, explicitly or implicitly. For models with a small number of parameters, the essential part of *a priori* information being in that case the restriction of the model space, additional priors for the parameters are typically easily overwhelmed by the data if the number of observations becomes large enough. The situation is completely different for nonparametric approaches, where the number of data is too small to determine all the degrees of freedom of the model. Thus, in nonparametric modeling it is essential that all available *a priori* information is included in the model. To this end prior models are useful which are formulated explicitly in terms of the likelihood of interest. Most priors used for nonparametric models in practice are some kind of smoothness priors. In the following we will discuss some possibilities to construct priors more general than smoothness priors and present techniques to adapt general prior models to the available *a priori* knowledge. Before we discuss specific nonparametric likelihood and prior models we first have to define the basic quantities in empirical learning and their probabilistic relations.

2 The Bayesian framework

2.1 The basic probabilistic model

For empirical learning it is convenient to distinguish three groups of variables:

1. observable (visible) independent variables x representing 'inputs' or 'controlled causes',
2. observable (visible) dependent variables y representing 'outputs' or 'measured effects',

3. not directly observable (hidden, latent) variables ϕ describing the possible (pure) 'states of Nature' considered in the model under study.

The joint probability can be factorized according to

$$p(x, y, \phi) = p(y|x, \phi)\, p(\phi|x)p(x). \tag{1}$$

The first factor $p(y|x, \phi)$ describes the probability (or probability density for continuous y) of finding y given the visible variables are in state x under a specific model state ('state of Nature') ϕ. For given data, $D = (x, y)$, the factor $p(y|x, \phi)$ is also known as the *likelihood* of ϕ under D.

The hidden variables ϕ parameterize the space of possible states we are using to model the probabilistic input–output relation between the x and the y. In 'parametric' models the hidden variables ϕ represent a (in many practical cases relatively small) number of parameters, while in 'nonparameteric' models the hidden variables ϕ form a function or 'field', so that expressions like $p(\phi|x)$ represent a stochastic process [8, 11, 40, 36]. For example, in nonparametric density estimation the function values of the likelihood itself, i.e., all the numbers $\phi(x, y) = p(y|x, \phi)$ for all x and y, can be considered as the primary degrees of freedom. A function $\phi(x, y)$ which determines the likelihood $p(y|x, \phi)$ will be called a *likelihood field*. Likelihood fields which are different from $\phi(x, y) = p(y|x, \phi)$ will be discussed below. Similar to nonparametric density estimation problems, we will speak of nonparametric regression if each value of the regression function $\phi(x) = \int dy\, y\, p(y|x, \phi)$ is considered as a primary degree of freedom. Another example is nonparametric inverse quantum theory where the function of interest can be a function $\phi(x) = v(x)$ describing a local quantum potential as function of a coordinate x. The primary degrees of freedom $\phi(x, y)$ or $\phi(x)$ can then be related by explicit prior information like a smoothness constraint. As the number of parameters in a parametric model can become large, for instance, in a neural network, and also nonparametric models must be discretized in some basis to perform numerical calculations, there is no sharp distinction between parametric and nonparametric models in practice. Nonparametric models as we will understand them, however, try to formulate the available *a priori* information not in terms of a parameters, which are often difficult to interpret, but explicitly in terms of the function ϕ of interest, e.g., in terms of the likelihood values $p(y|x, \phi)$ in density estimation problems, or in terms of the regression function in regression problems.

In the following we will assume scenarios with n independent (tuples of) *training data*

$$D = \{(x_i, y_i)|1 \le i \le n\} = \{(x_D, y_D)\} \tag{2}$$

which are sampled according to the likelihood $p(y_i|x_i, \phi)$ and where the x_i and y_i can be vectors. The independent variables in the data, $x_D = \{x_i|1 \le i \le n\}$, may have been either fixed in advance or sampled under some given $p(x)$. We also assume, that $p(\phi|x)$ can be written as $p(\phi|D_0)$ with D_0 being that part of the independent visible variables which determines the prior $p(\phi|D_0)$ of the hidden variables ϕ. We will call D_0 the *a priori information* on ϕ and $p(\phi|D_0)$ the *prior* for ϕ. Under those assumptions, we end up with models of the form

$$p(x, y, \phi|D_0) = p(\phi|D_0)p(y_D|x_D, \phi)\, p(x_D) = p(\phi|D_0) \prod_{i=1}^{n} p(y_i|x_i, \phi)\, p(x_i), \tag{3}$$

or, conditional on the independent variable x,

$$p(y, \phi|x, D_0) = p(y_D|x_D, \phi)\, p(\phi|D_0) = \prod_{i=1}^{n} p(y_i|x_i, \phi)\, p(\phi|D_0). \tag{4}$$

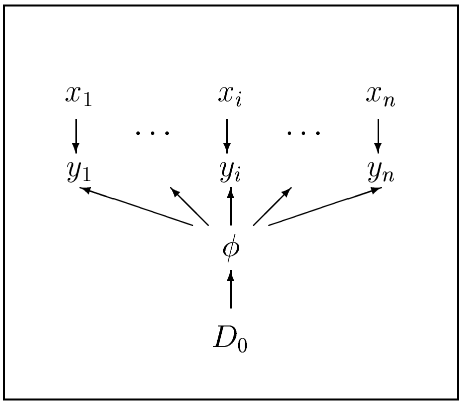

Figure 1: Graphical representation of the probabilistic model (4) with independent data, conditioned on the independent variables x_i and the unknown state ϕ.

2.2 Bayesian decision theory and predictive density

Empirical learning is used to make decisions. Hence, let us consider a set of possible actions a from which we want to choose an optimal one. In selecting an optimal action we want to use all the information we have at hand, i.e., training data D and *a priori* information D_0. In particular, we will be interested in approximation problems, like density estimation where the possible actions represent predictive densities $p(y|x,a)$ and our aim is to approximate the true state $p(y|x,\phi_{\rm true})$, by an optimal posterior predictive density $p(y|x,D,D_0)$. To define optimality we select a loss function $l(x,y,a)$ which describes the *loss* suffered in situation x if y appears and action a has been selected. Then an optimal action (e.g., an optimal prediction) minimizes, for given data D and prior D_0, the *expected risk* [1, 34]

$$r(a,D,D_0) = \int dx\, dy\, p(x)\, p(y|x,D,D_0)\, l(x,y,a). \tag{5}$$

If we denote the expectation under the joint posterior predictive density $p(x,y|D,D_0) = p(x)p(y|D,D_0)$ by $< \cdots >_{X,Y|D,D_0}$, we can write Eq. (5) as

$$r(a,D,D_0) =< l(x,y,a) >_{X,Y|D,D_0} . \tag{6}$$

We may remark that $p(x)$ is the density of x for the (test) data for which prediction is intended and not the possibly different $p(x)$ which has been used for sampling the training data D. The typical loss function for approximation problems is the log–loss

$$l(x,y,a) = -\ln p(y|x,a), \tag{7}$$

for which the expected risk becomes the (x–averaged) Kull–Back–Leibler distance between $p(y|x,a)$ and $p(y|x,D,D_0)$. As it is not difficult to show by using Jensen's inequality the Kullback–Leibler–distance is minimal for $p(y|x,a) = p(y|x,D,D_0)$. Thus, under log–loss the optimal model is the posterior predictive density

$$p(y|x,D,D_0) = \int d\phi\, p(y|x,\phi)\, p(\phi|D,D_0). \tag{8}$$

In nonparametric models where ϕ is a function the integral over ϕ stands for a functional integral. Such integrals are mathematically well defined for Gaussian processes ('generalized free Euclidean fields' or 'generalized Brownian motion' in the language of physics), but for more general processes ('(self–)interacting fields') a general mathematical definition of functional integrals is still an unsolved problem. One possible approach is to define a non–Gaussian functional integral by renormalization procedures in perturbation theory [5] or on a lattice [29], another simpler possibility is to discretize the function ϕ in some appropriate basis so the functional integral becomes a standard (but typically still extremely highdimensional) integral. The remaining highdimensional integrals can either be approximated by Monte Carlo methods [16, 14, 12, 32, 19] or by the method of steepest descents (Laplace approximation), in this context known as Maximum A Posteriori (MAP) approximation [7, 4, 34, 12]. The latter assumes that

$$p(y|x, D, D_0) \approx p(y|x, \phi^*) \tag{9}$$

for the 'MAP solution' ϕ^* which maximizes the posterior

$$\phi^* = \text{argmax}_\phi p(\phi|D, D_0). \tag{10}$$

Hence, in a MAP approximation the integration over ϕ is replaced by maximization over ϕ. Eq. (9) is a good approximation if the posterior $p(\phi|D, D_0)$ can be well approximated by a single Gaussian and the factor $p(y|x, \phi)$ in Eq. (8) is slowly varying at the stationary point ϕ^* compared to the posterior. This is often expected to be the case if the number of data is sufficiently large. In principle one can also go beyond the MAP approximation by using the MAP solution as starting point for a perturbation series, for example, graphically expressed by Feynman diagrams [33, 17].

In Gaussian regression $p(y|x, a)$ is assumed to be of Gaussian form with fixed variance and only the mean or regression function $a(x) = < y >_{Y|x,a} = \int dy\, y\, p(y|x, a)$ is adapted. In linear (parametric) Gaussian regression the regression function is chosen as a linear funtion in x, i.e., $< y >_{Y|x,a} = a_0 + a_1 x$, while in nonparametric Gaussian regression for each x the value $< y >_{Y|x,a}$ is treated as a single degree of freedom so that the form of the regression function is not restricted. For Gaussian regression log–loss is equivalent to a squared error loss and it is straightforward to check that the regression function which minimizes the expected risk is the posterior regression function, i.e., the expectation of y under the posterior predictive density

$$a^*(x) = < y >_{Y|x,a^*} = < y >_{Y|x,D,D_0} = \int dy\, y\, p(y|x, D, D_0). \tag{11}$$

A typical loss function for classification problems (where a discrete dependent variable y represents the class variable) is the 0–1–loss

$$l(y, x, a) = -\delta_{y,a(x)} \tag{12}$$

where $\delta_{y,a(x)}$ stands for the Kronecker-δ which is equal to one if $y = a(x)$ and zero otherwise. In that case it turns out that the optimal $a(x)$ is a mode function $\max_y p(y|x, D, D_0)$ of the predictive density. Table 1 lists the basic functions which appear in Bayesian models.

2.3 Bayes' theorem: From prior and likelihood to the posterior

The short discussion of Bayesian decision theory in 2.2 showed that for density estimation, regression, as well as for classification learning can be based on the posterior predictive density $p(y|x, D, D_0)$. Using $p(A) = \int dB\, p(A, B)$ for $A = \phi$ and $B = y$, $p(y|x, \phi, D, D_0) =$

$p(\phi\|D_0)$	prior
$p(\phi\|D, D_0)$	posterior
$p(y\|x, \phi)$	likelihood, predictive density for a pure state
$p(y\|x, D_0)$	prior predictive density
$p(y\|x, D, D_0)$	posterior predictive density

Table 1: The basic probabilities (or densities, stochastic processes, respectively).

$p(y|x,\phi)$, and $p(\phi|x,D,D_0) = p(\phi|D,D_0)$ the posterior predictive density can be expressed as

$$p(y|x,D,D_0) = \int d\phi\, p(y|x,\phi)p(\phi|D,D_0), \tag{13}$$

where the density (or the stochastic process in nonparametric models) $p(\phi|D,D_0)$ is known as the *posterior*.

The posterior, which we are interested in, is linked to the prior and likelihood, which are assumed to be given, by Bayes' theorem:

$$p(H|D) = \frac{p(D|H)p(H)}{p(D)}. \tag{14}$$

Because $p(D)$ can not be zero for observed data, Eq. (14) is a direct consequence of the definition of conditional probabilities,

$$p(H,D) = p(H|D)p(D) = p(D|H)p(H). \tag{15}$$

In a Bayesian context D represents the available data and H stands for the hidden variables (hypotheses, theories, models) which in the nonparametric approaches we will study are described by functions or likelihood fields, like $\phi(x)$ in regression or $\phi(x,y)$ in general density estimation problems. The terms in Eq. (14) are commonly referred to as

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}, \tag{16}$$

where the evidence $p(D)$ is independent of H. Conditional on the independent training data $x_D = \{x_i|1 \le i \le n\}$, denoting the *a priori* information explicitly by D_0, and writing ϕ for H Bayes' theorem (14) reads

$$p(\phi|y_D,x_D,D_0) = \frac{p(y_D|x_D,\phi)p(\phi|D_0)}{p(y_D|x_D,D_0)}. \tag{17}$$

According to Eq. (17) two ingredients are needed to calculate the posterior: a likelihood model describing the measurement process from which the data are obtained and a prior model implementing the available *a priori* information.

3 Likelihood models

Likelihood or data models define the different problem classes. In the following we will introduce the nonparametric likelihood models of density estimation, regression and inverse quantum theory.

3.1 Log–probabilities, energies, and density estimation

In practice it is often more convenient to work with log–probabilities (or log–densities, log–processes, respectively)

$$L = \ln p \tag{18}$$

instead of probabilities p. Indeed, for log–probabilities no non–negativity constraint is necessary because $p = \exp L$ is non–negative for arbitrary real L. Another advantage of log–probabilities is that products of probabilities become sums for log–probabilities,

$$p(A,B) = p(A)p(B) \Rightarrow L(A,B) = L(A) + L(B). \tag{19}$$

In limits where the number of probability factors approaches infinity the corresponding sums of log–probability terms become integrals, e.g., in the log–prior, which are more common objects than the infinite products which appear for probabilities. For example, according to Bayes' theorem (17) the *log–posterior*

$$L(\phi|D,D_0) = \ln p(\phi|D,D_0) \tag{20}$$

becomes up to a ϕ–independent constant the sum of the *log–prior*

$$L(\phi|D_0) = \ln p(\phi|D_0), \tag{21}$$

which in nonparametric approaches typically consists of an integral over x and/or y, and the *log–likelihood*

$$L(y_D|x_D,\phi) = \sum_i L(y_i|x_i,\phi) = \ln \prod_i p(y_i|x_i,\phi), \tag{22}$$

which, in contrast to the log–prior, is a finite sum over data points. We remark that the log–lilekihood can be related to the *averaged conditional Kerridge inaccuracy* K_c,

$$L(y_D|x_D,\phi) = \sum_i L(y_i|x_i,\phi) = -nK_c[p_{\rm emp}(x',y'),p(y'|x',\phi)], \tag{23}$$

in short $L = \sum_i L_i = -nK_c$, where the averaged conditional Kerridge inaccuracy

$$K_c[p_{\rm emp}(x',y'),p(y'|x',\phi)] = -\int dx'\, p_{\rm emp}(x')K[p_{\rm emp}(y'|x'),p(y'|x',\phi)] \tag{24}$$

is the empirical expectation of the corresponding Kerridge inaccuracy of the conditional densities, obtained by integrating over x with $p_{\rm emp}(x')$,

$$K[p_{\rm emp}(y'|x'),p(y'|x',\phi)] = -\int dy'\, p_{\rm emp}(y'|x') \ln p(y'|x',\phi), \tag{25}$$

where the empirical densities for data $D = \{(x_i, y_i)|1 \le i \le n\}$ are given by

$$p_{\text{emp}}(y|x_i) = \frac{p_{\text{emp}}(x_i, y)}{p_{\text{emp}}(x_i)}, \tag{26}$$

$$p_{\text{emp}}(x, y) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i)\delta(y - y_i), \tag{27}$$

$$p_{\text{emp}}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i), \tag{28}$$

with $x_i \in x_D$ so that $p_{\text{emp}}(y|x_i)$ is defined. Denoting the likelihood vector by $P(\phi)$ with components $P(x, y; \phi) = p(y|x, \phi)$ and defining the data vector $N(x, y) = np_{\text{emp}}(x, y)$ we can also write for the likelihood, alternatively to Eq. (23),

$$L(y_D|x_D, \phi) = -\int dx\, dy\, \ln p(y|x, \phi) N(x, y) = -\langle \ln P(\phi)|\, N\rangle \tag{29}$$

using the bra–ket notation with respect to x and y, i.e., $\langle f|\, g\rangle = \int dx\, dy f(x, y)g(x, y)$.

Besides being non–negative, probabilities (or densities) have to be normalized to one. Like non–negativity, the normalization condition can be implemented in the parameterization of a probability. Indeed, if we parameterize a probability as follows

$$p(x) = \frac{\exp\left(-\beta E(x)\right)}{Z}, \tag{30}$$

where the normalization factor or *partition sum* Z is given by

$$Z = \int dx\, \exp\left(-\beta E(x)\right), \tag{31}$$

then $p(x)$ is normalized and non–negative for arbitrary real functions $E(x)$. In analogy to statistical physics, we will call the function E *energy*. Clearly, the random variable x in Eq. (30) can be replaced by any other random variable. Hence energies, for example, appear as *likelihood energy* $E(y|x, \phi)$, *prior energy* $E(\phi|D_0)$, or *posterior energy* $E(\phi|D, D_0)$. The factor β, which in statistical physics plays the role of an *inverse temperature*, can be useful for several technical purposes, for example, when calculating the moments of the random variable $E(x)$. We will set $\beta = 1$ in the following.

If searching for a maximum of $p(x)$ the x–independent normalization factor Z is irrelevant. That means, when maximizing a probability, e.g., a posterior $p(\phi|D, D_0)$, then when working with energies like $E(\phi|D, D_0)$ we do not have to implement non–negativity and normalization constraints explicitly and we do not have to calculate the normalization factor Z. This is especially useful if Z is an integral over a high dimensional ϕ–space.

Summarizing, the terms in Eq. (30) can be referred to as

$$\text{probability} = \frac{1}{\text{partition sum}} \exp\left(-\frac{\text{energy}}{\text{temperature}}\right). \tag{32}$$

Table 2 lists some possible choices for the likelihood field $\phi(x, y)$ in density estimation and the corresponding constraints the field has to obey.

field $\phi(x,y)$	likelihood $p(y\|x,\phi) =$	constraints	
likelihood	$\phi(x,y)$	norm	non–negativity
unnormalized likelihood	$\frac{\phi(x,y)}{\int dy\, \phi(x,y)}$	—	non–negativity
log–likelihood	$\exp\left(\phi(x,y)\right)$	norm	—
likelihood energy	$\frac{\exp(-\beta\phi(x,y))}{\int dy\, \exp(-\beta\phi(x,y))}$	—	—
distribution function	$\frac{d\phi(x,y)}{dy}$	boundary	monotony

Table 2: Some choices of likelihood fields and the corresponding constraints.

3.2 Regression

In regression problems the only unknown parameter of the likelihood , i.e., the conditional density $p(y|x,\phi)$ is assumed to be the mean or regression function

$$\phi(x) = \int dy\, y\, p(y|x,\phi). \tag{33}$$

A common case is Gaussian regression where the likelihood is chosen as a Gaussian with fixed variance

$$p(y|x,\phi) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|y-\phi(x)|^2}{2\sigma^2}\right). \tag{34}$$

An example of such a Gaussian likelihood is shown in Fig. 2.

Regression models, however, can not only be formulated with a Gaussian likelihood as in Eq. (34) but also with different likelihood models. For example, in classical inverse problems one assumes a likelihood of the form

$$p(y|x,\phi) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|y-(\mathbf{F}\phi)(x)|^2}{2\sigma^2}\right). \tag{35}$$

In that case data $D = \{(x_i, y_i)|1 \le i \le n\}$ are sampled for $(\mathbf{F}\phi)(x)$ and not for $\phi(x)$. The operator $\mathbf{F}$ often represents an 'instrument or apparatus function'. If it is linear, then

$$(\mathbf{F}\phi)(x) = \int dx'\, \mathbf{F}(x,x')\phi(x'). \tag{36}$$

A regression model with a Poisson likelihood

$$p(y|x,\phi) = \frac{[\phi(x)]^y}{y!}\, e^{-\phi(x)}, \quad y \in N,\ \phi > 0 \tag{37}$$

can be useful for counting events (y counts with mean $\phi(x)$ at x). Similarly , a binomial likelihood

$$p(y|x,n;\phi) = \binom{n(x)}{y}[\phi(x)]^y[1-\phi(x)]^{n(x)-y}, \quad y \in N,\ y \le n \in N,\ 0 \le \phi \le 1, \tag{38}$$

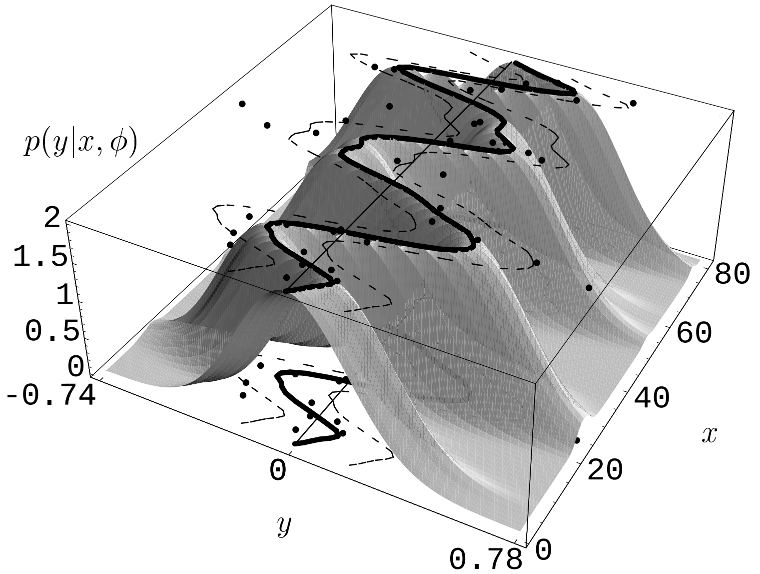

Figure 2: Example of a Gaussian likelihood $p(y|x,\phi)$ as in Eq. (34) where the regression function $\phi(x) = \int dy\, y\, p(y|x,\phi)$ (thick line) is sampled from a prior process $p(\phi|D_0)$ which favors smooth and approximately periodic functions. (The corresponding prior covariance is shown in Fig. 4.) The black dots represent data points $D = \{(x_i, y_i)|1 \le i \le n\}$ sampled from the given likelihood, the dashed line shows the $\pm\sigma$-range around the mean function $\phi(x)$.

is appropriate if at x from $n(x)$ binary events a number of y 'ones' and $n(x) - y$ 'zeros' have been observed, $\phi(x)$ being the probability of 'one' at x. Finally we may remark that support vector machines for regression are formally equivalent to an regression model with ϵ–insensitive likelihood energy $E_\epsilon(y,\phi(x)) = \Theta(|y-\phi(x)|-\epsilon)(|y-\phi(x)|-\epsilon)$, $\Theta(x)$ denoting the step function and ϵ a constant [37, 41, 35].

3.3 Inverse quantum theory

As an example of a specific application of nonparametric Bayesian methods we will give a short introduction to the Bayesian approach to inverse quantum theory. In inverse quantum theory a quantum system is determined by measurements. A typical example, is the reconstruction of a quantum potential (which determines the force acting on a particle) from a finite number of position measurements. Such problems are based on the specific likelihood model of quantum theory: Measuring observable $\hat{X}$ for a quantum system in a state described by density operator ρ, e.g., depending on an unknown potential ϕ, the probability to obtain value y is given by [39, 28]

$$p(y|\hat{X},\phi) = \mathrm{Tr}(\Pi_{\hat{X},y}\,\rho(\phi)), \tag{39}$$

where Tr stands for the trace and $\Pi_{\hat{X},y}$ is the projector on the space of eigenfunctions of operator $\hat{X}$ with eigenvalue y. For a system in a state described by wave function ψ the density operator is given by

$$\rho = \frac{|\psi\rangle\,\langle\psi|}{\langle\psi|\,\psi\rangle}, \tag{40}$$

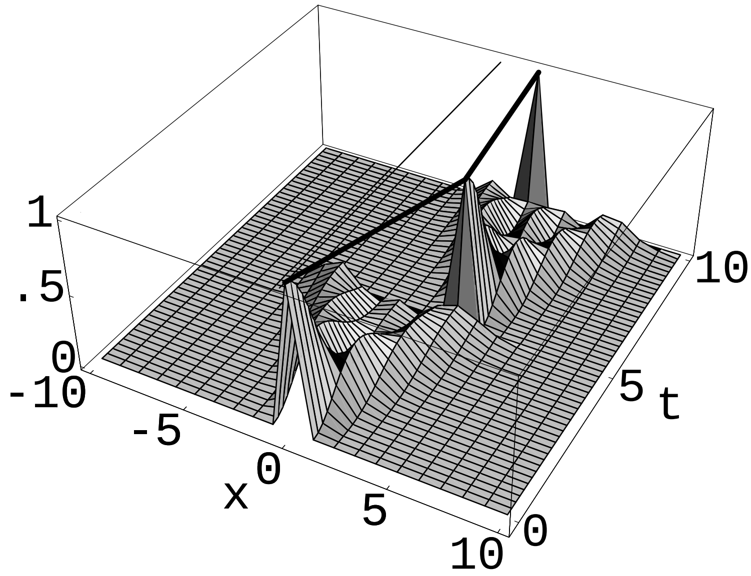

Figure 3: Quantum time series data: Time evolution of a particle observed at times t_1 = 5 and t_2 = 10 at positions x_i in a given potential $v_{\rm true}(x)$ unknown to the learner. (The potential $v_{\rm true}(x)$ is shown on the right hand side of Fig. 7.) The figure shows the likelihood $p(x(t_i)|\hat{X}, x(t_{i-1}), v_{\rm true})$ of Eq. (42), where the dependent variable y is represented by the measured coordinates $x(t_i)$, the potential function is chosen as field $\phi(x)$, and the Hamiltonian in (43) is the sum of a kinetic energy term and the local potential. In time–dependent inverse quantum theory the aim is to reconstruct the potential by combining the likelihood of such times series data with a (nonparametric) prior on $v(x)$.

for a system at (physical) temperature $1/\beta$ the density operator of the corresponding (canonical) ensemble is of the form (setting Boltzmann's constant equal to 1) [23]

$$\rho = \frac{\exp\left(-\beta H(\phi)\right)}{\mathrm{Tr}\exp\left(-\beta H(\phi)\right)}, \tag{41}$$

where H stands for the Hamiltonian of the system which depends on the field ϕ (and which may represent, for example, a potential). Similarly, the likelihood for quantum time series data $y(t_i)$ for an observable $\hat{X}$ measured at times t_i, $1 \le i \le n$, reads [22]

$$p(y(t_i)|\hat{X}, y(t_{i-1}), \phi) = |\left\langle y_i \left| U_i \right| y_{i-1} \right\rangle|^2 \tag{42}$$

where $|y_i\rangle$ stands for the eigenfunction of $\hat{X}$ with eigenvalue $y(t_i)$ and (setting $\hbar$ = 1)

$$U_i = \exp\left(-i(t_i - t_{i-1})H\right) \tag{43}$$

is the time evolution operator of a quantum system with time–independent Hamiltonian H. An example of the likelihood of a particle moving in a quantum potential is shown in Fig. 3.

In nonparametric Bayesian inverse quantum theory the likelihood model of quantum mechanics is combined with a prior model for the field ϕ. Being interested in the reconstruction of a quantum potential, the field $\phi(x)$ may represent the diagonal elements $v(x)$ of a local potential (where this time x stands not for an independent variable but for the coordinates of a quantum particle). The prior model for ϕ then implements available *a priori* information on $v(x)$, like smoothness or an approximate periodicity, the possibility of certain distortions or a specific fractal structure. Such nonparametric prior models, useful for density estimation, regression, and inverse quantum theory, will be discussed in the next section.

4 Prior models

For nonparametric models, where the number of degrees of freedom (represented by the field values $\phi(x)$ or $\phi(x,y)$) is large compared to the number n of data, learning requires the combination of the likelihood model with an appropriate prior model.

4.1 Gaussian prior factors and approximate symmetries

The simplest but already very flexible nonparametric prior models are Gaussian process priors [40, 42, 32, 18]

$$p(\phi|D_0) = \left(\det \frac{\mathbf{K}}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}\langle \phi - t|\mathbf{K}|\phi - t\rangle} \propto e^{-E(\phi|D_0)}, \tag{44}$$

where

$$\langle \phi - t\,|\mathbf{K}|\,\phi - t\rangle = \int dx\,dx'\,dy\,dy'\;(\phi(x,y) - t(x,y))\,\mathbf{K}(x,y;x',y')\,(\phi(x',y') - t(x',y')) \tag{45}$$

with $\mathbf{K}$ being the real symmetric positive semi–definite inverse covariance operator of the Gaussian process and $t(x,y)$ (or $t(x)$ for regression) the mean function of the process. As the $\phi(x,y)$ with maximal prior is given by the mean function, $t(x,y)$ will also be called a reference function or 'prior template' for the field ϕ.

A prior of the form (44) implements the *a priori* knowledge that the field ϕ is expected to be similar to the mean function t measured in a distance defined by the inverse covariance $\mathbf{K}$. Similarly, the *a priori* knowledge that the field ϕ is expected to be similar to a reference field t_1 OR to a reference field t_2, (measured in a distance being defined the by two inverse covariances $\mathbf{K}_1$ and $\mathbf{K}_2$) can be implemented by using a sum of two Gaussian prior components of the form (44). Generalized to more than two components this yields a *Gaussian mixture prior* $p(\phi|D_0) = \sum_k a_k p_k(\phi)$, with $\sum_k a_k = 1$, mean functions t_k and inverse covariances $\mathbf{K}_k$. In particular if the $\mathbf{K}_k$ are all equal, then such mixture models are not much more difficult to solve than models with a single Gaussian prior [20, 21]. An application of a mixture of Gaussian priors is presented in Fig. 6.

We remark that a product of Gaussian likelihood factors (34) for regression problems can be expressed in a form similar to that of the prior (44): The sum of mean squared error terms appearing in the exponent of the likelihood product can be written [20]

$$\sum_{i=1}^{n} \frac{|y_i - \phi(x_i)|^2}{2\sigma^2} = \frac{1}{2}\,\langle \phi - t_D\,|\mathbf{K}_D|\,\phi - t_D\rangle + \frac{n}{2}V^D \tag{46}$$

with ϕ–independent data variance V_D, data mean vector $t_D = \mathbf{K}_D^{-1}\sum_{i=1}^n \mathbf{K}_i t_i$, inverse data covariance $\mathbf{K}_D = \sum_{i=1}^n \mathbf{K}_i$, where $t_i(x) = y_i$ and $\mathbf{K}_i(x,x') = \delta(x-x')\delta(x-x_i)/\sigma^2$.

The inverse prior covariance $\mathbf{K}$ can be chosen to implement approximate invariance of the field ϕ under an operation $\mathbf{S}$. Indeed, if we choose in (44)

$$\mathbf{K} = (\mathbf{I} - \mathbf{S})^T(\mathbf{I} - \mathbf{S}), \tag{47}$$

$\mathbf{I}$ denoting the identity and $\mathbf{S}^T$ the transpose of $\mathbf{S}$, then fields which fulfill $\phi = \mathbf{S}\phi$, i.e., which are invariant under $\mathbf{S}$, maximize the prior (44). Hence, a Gaussian prior (44) with an inverse

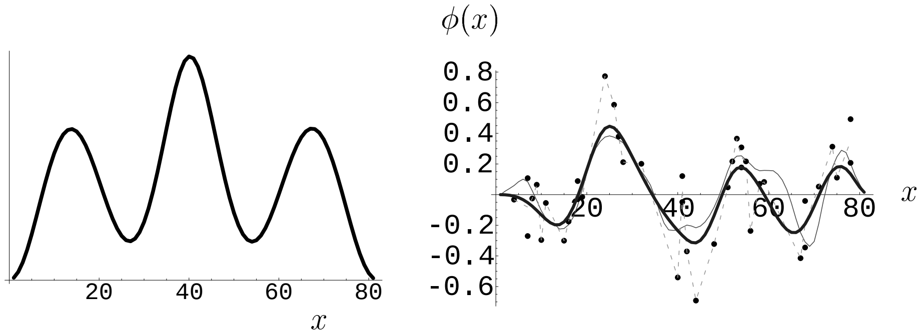

Figure 4: Left hand side: Prior covariance $\mathbf{K}^{-1}(x, x_0)$ (shown for fixed $x_0 = 40$ on mesh with 80 points) used to generate the sample data in Fig. 2. Right hand side: Regression function (thick line) reconstructed from the data points shown in Fig. 2 and using the prior covariance shown on the left. Also shown are: the data points (dots), the true regression function (thin line), and a piecewise linear interpolation (dashed line).

covariance of the form (47) generates functions which are approximately invariant under $\mathbf{S}$. For example, if $\mathbf{S}$ is chosen as the translation of ϕ by θ units in x–direction, i.e., $\mathbf{S}^x_\theta \phi(x, y) = \phi(x - \theta, y)$, then the invariant fields are periodic in x and $\mathbf{K} = (\mathbf{I} - \mathbf{S}^x_\theta)^T(\mathbf{I} - \mathbf{S}^x_\theta)$ implements approximate periodicity in x–direction. A regression function sampled from a prior which favors smooth and approximate periodic functions is shown in Fig. 2.

Similarly, approximate invariance under infinitesimal transformations s, generating a Lie group $\mathbf{S}(\theta) = \exp(\theta \mathbf{s})$ with parameter θ, can be implemented by inverse covariances

$$\mathbf{K} = \mathbf{s}^T \mathbf{s}. \tag{48}$$

A typical example is approximate invariance under infinitesimal translations ('smoothness'), say with respect to a K–dimensional vector x with components x_k, for which $\mathbf{s}_k = \partial/\partial x_k$ so that $\mathbf{K} = \sum_{k=1}^K \mathbf{s}_k^T \mathbf{s}_k$ becomes (under approriate boundary conditions) the negative Laplacian $\mathbf{K} = -\Delta = -\sum_k \partial^2/\partial x_k^2$ (being of the form of kinetic energy terms in Euclidean field theory). In statistics one typically uses inverse covariances which also include higher order derivatives and which result in 'smoother', i.e., more times differentiable fields ϕ^* as MAP solutions. One such example is the Radial Basis Function (RBF) inverse prior covariance [15]

$$\mathbf{K} = \lambda \sum_{k=0}^{\infty} \frac{1}{k!} \left(-\frac{\sigma_0^2 \Delta}{2} \right)^k = \lambda \exp\left(-\frac{\sigma_0^2 \Delta}{2} \right), \tag{49}$$

with parameters λ and σ_0.

For density estimation with likelihood fields where the non–negativity and normalization constraints are not fulfilled automatically, those hard constraints have to be added to the Gaussian prior of Eq. (44), resulting in

$$p(\phi|D_0, \text{norm, non–neg.}) \propto \prod_x \delta(\int dy\, p(y|x,\phi) - 1) \prod_{xy} \Theta\Big(p(y|x,\phi)\Big) p(\phi|D_0) \tag{50}$$

where the δ–functions ensure normalization for all x and the step functions Θ non–negativity of the likelihood for all x and y. The δ–function can be transformed into a Lagrange multiplier term for the log–prior, yielding

$$\begin{aligned} E(\phi) &= -\langle \ln P(\phi) |\, N \rangle + \frac{1}{2} \langle \phi - t \,|\mathbf{K}|\, \phi - t \rangle + \langle P(\phi) |\, \Lambda_X \rangle \qquad (51) \\ &= -\langle \ln P(\phi) |\, N \rangle + \frac{1}{2} \langle \phi \,|\mathbf{K}|\, \phi \rangle - \langle J |\, \phi \rangle + \langle P(\phi) |\, \Lambda_X \rangle + c, \end{aligned}$$

with a *Lagrange multiplier function* $\Lambda_X(x,y) = \Lambda_X(x)$ which is determined by the normalization condition. A similar Lagrange multiplier term can be introduced for the non–negativity condition, but calculating the posterior predictive density in MAP approximation (see Eq. (10)) non–negativity terms are in many practical cases not necessary. Indeed, because the likelihood is by definition larger than zero at all data points, the field ϕ^* with maximal posterior can typically also not become smaller than zero between data points if we assume some common smoothness prior. Technically speaking, the non–negativity constraint is for smoothness priors typically not active at the stationary point ϕ^*. In the language of physics the term $J = \mathbf{K}t$ in Eq. (51) represents an *external field* coupling to $\phi(x,y)$, similar, for example, to a magnetic field. A non–zero field J leads to a non–zero expectation of ϕ in the no–data case. The ϕ–independent c stands for the term $\frac{1}{2}\langle t\,|\mathbf{K}|\,t\rangle$, which is for invertible $\mathbf{K}$ equal to $\frac{1}{2}\langle J\,|\mathbf{K}^{-1}|\,J\rangle$, and can be skipped when minimizing $E(\phi)$ with respect to ϕ.

A MAP solution ϕ^*, which maximizes the posterior $p(\phi|D,D_0)$ or, equivalently, minimizes the energy $E(\phi)$ can be found by setting the functional (Fréchet) derivative of the functional $E(\phi)$ in (51) with respect to the function $\phi(x,y)$ to zero, i.e., by solving the stationarity equations $\delta E(\phi)/\delta\phi(x,y) = 0$ for all x and y. The stationarity equations for ϕ resulting from (51) can easily be expressed in vector notation

$$0 = \mathbf{P}'(\phi)\mathbf{P}^{-1}(\phi)N - \mathbf{P}'(\phi)\Lambda_X - \mathbf{K}\,(\phi - t)\,, \tag{52}$$

where the Lagrange multiplier function follows from the normalization constraints over y for all x yielding, if $\mathbf{P}\mathbf{P}'^{-1}$ is invertible and $\Lambda_X \neq 0$,

$$\Lambda_X = \mathbf{I}_X\left(N - \mathbf{P}\mathbf{P}'^{-1}\mathbf{K}\,(\phi - t)\right), \tag{53}$$

with identity on X, $\mathbf{I}_X(x,y;x',y') = \delta(x-x')$ and diagonal likelihood operator

$$\mathbf{P}(x,y;x',y';\phi) = \delta(x-x')\delta(y-y')P(x,y;\phi), \tag{54}$$

with Jacobian

$$\mathbf{P}'(x,y;x',y';\phi) = \frac{\delta P(x',y';\phi)}{\delta\phi(x,y)}. \tag{55}$$

Choosing the regression function as field the stationarity equation (52) is linear for (nonparametric, nonlinear) regression problems provided the likelihood model and the prior model are Gaussian. In general, however, the stationarity equation (52) is nonlinear and has to be solved by iteration.

4.2 Hyperparameters and hyperfields

4.2.1 Hyperparameters and the boosting of parametric models

Hyperparameters are parameters of the prior [1, 2, 12, 6, 27], like the 'regularization factor' λ and the width σ_0 in Eq. (49). Introducing hyperparameters into a model means decomposing the prior according to

$$p(\phi|D_0) = \int d\theta\; p(\phi|\theta,D_0)p(\theta|D_0) \tag{56}$$

denoting hyperparameters collectively by θ. In many practical cases the integral over θ in Eq. (56) can not be calculated analytically; like the integral over ϕ in Eq. (8) it then has

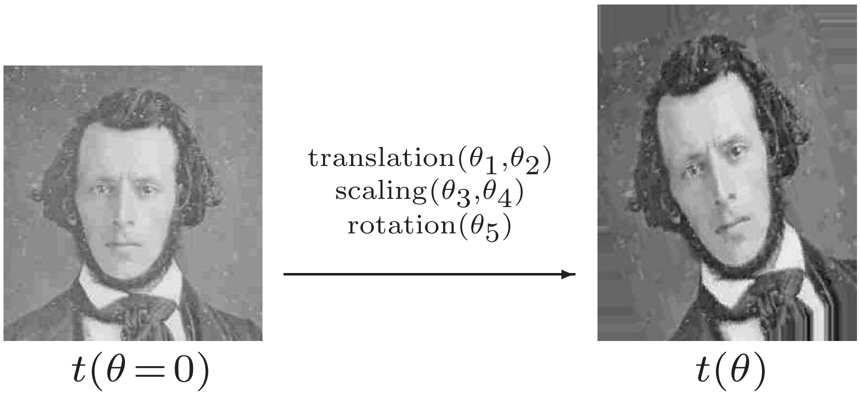

Figure 5: Example of a mean function $t(x, \theta)$ with a five–dimensional hyperparameter vector (and a two–dimensional x) used for a Gaussian prior as in Eq. (44) with prior energy $\frac{1}{2}\langle\phi - t\,|\mathbf{K}|\,\phi - t\rangle$. Such (graylevel) mean functions $t(x, \theta)$ have been used in image completion tasks (see Fig. 6) where the (graylevel) function $\phi(x)$ represents an image which has to be reconstructed from pixels sampled with Gaussian noise from the original image [20, 21]. The five dimensional hyperparameter vector θ which includes two translation, one rotation, and two scaling parameters makes the mean function t an adaptable template for the image ϕ. In a joint MAP approximation the function $\phi(x)$ and the hyperparameters θ are optimized simultaneously.

to be calculated numerically, e.g., by Monte Carlo methods, or in Maximum A Posteriori Approximation. In a joint MAP approximation the integral over ϕ in Eq. (8) and the integral over θ in Eq. (56) are calculated simultaneously in Maximum A Posteriori Approximation, resulting in coupled stationarity equations for ϕ and θ.

Hyperparameters can be used to adapt the mean function or the covariance of a Gaussian process prior. An example of the latter is the Automatic Relevance Detection by MacKay and Neal [31, 26, 25, 20]. Hyperparameters for the mean functions of a Gaussian process have been used, for example, in image completion tasks (see Fig. 5 and Fig. 6). The adaption of the mean function is technically simpler then changing the covariance because changing the mean of a Gaussian process does not change the normalization constant but changing the covariance normally does. Solving the coupled stationarity equations for θ and ϕ in a joint MAP approximation for adapting the mean function of a Gaussian prior can be interpreted as a *nonparametric boosting of a parametric model* [20] where during iteration first a parametric model $t(x, y : \theta)$ is optimized with respect to the (hyper)parameters θ and then the optimal parametric solution $t(x, y; \theta^*)$ is used as a mean function for the Gaussian prior for ϕ. Fig. 7 shows as an example the reconstruction of a quantum potential $v(x)$ using as reference potential a parametric solution $t(x; \theta^*)$ with a θ^* which is obtained in maximum likelihood approximation (i.e., in a MAP approximation with uniform (hyper)prior for θ).

4.2.2 Local hyperfields and filtered differences

Under *local hyperfields* $\theta(x)$ we understand a collection of hyperparameters indexed by the visible variables x [20, 24]. It is straightforward to consider local hyperfields $\theta(x, y)$ depending on both, x and y, but for the sake of simplicity we restrict to fields $\theta(x)$, having in mind, for example, regression problems or the reconstruction of a quantum potential $v(x)$. Similarly, nonlocal hyperfields can be introduced depending, for example, on more than one x–value. Local hyperfields are useful to adapt the mean function or the inverse covariance of a Gaussian process locally. To show this, we decompose a real–symmetric, positive (semi–)definite

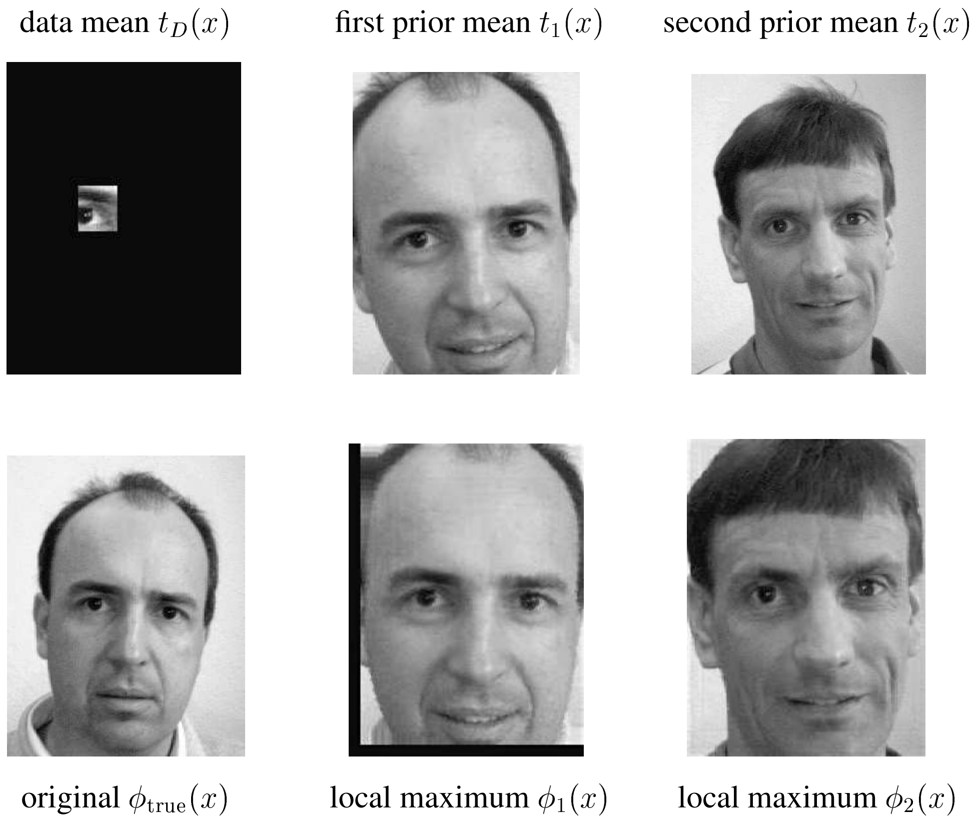

Figure 6: Image completion with a Gaussian mixture prior as an example of a regression problem [20, 21]. The data (first row on the left, for the definition of t_D see 46) have been sampled from the original image $\phi_{\text{true}}(x)$ (second row on the left) with Gaussian noise. The prior consists of a sum of two Gaussian priors $p(\phi|D_0) = 0.5p_1(\phi) + 0.5p_2(\phi)$ both components with a negative Laplacian covariance but with two different mean functions (image templates) $t_1(x)$ and $t_2(x)$ as shown in the first row. Both mean functions t_i have been made flexible using the five dimensional hyperparameter vector introduced in Fig. 5. Shown is a situation where the mixture prior and also the posterior possesses two local maxima ('low temperature case'). The two regression functions $\phi_1(x)$ and $\phi_2(x)$ representing the local maxima of the posterior are shown in the second row, optimally scaled, shifted and rotated. The global maximum, which is $\phi_1(x)$, is the MAP reconstruction of the true regression function ϕ_{true}.

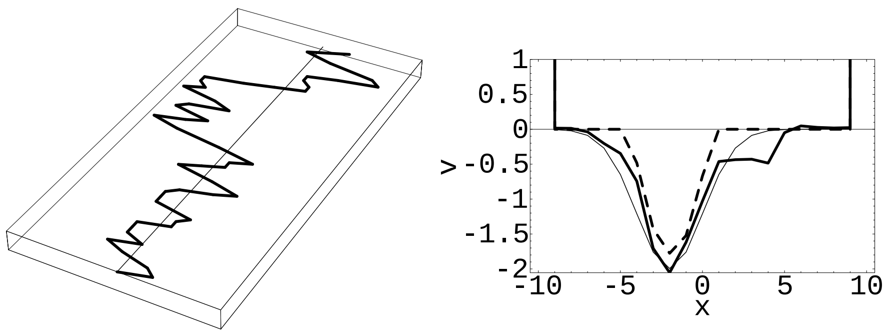

Figure 7: Left hand side: Shown is a sample path of a quantum particle, obtained from 50 coordinate measurements at fixed time intervals analogous to Fig. 3. Right hand side: An example of the numerical reconstruction of a quantum potential from the 50 coordinate measurements with likelihood as in (42) and a Gaussian prior on $v(x)$ similar to Eq. (49). (For more details see [22].) Shown are the true potential $v_{\text{true}}(x)$ (thin line), the best parametric approximation used as reference potential $t(x)$ (dashed line), and the reconstructed potential $\phi(x) = v(x)$ (thick line).

inverse covariance $\mathbf{K}(x, x')$ (defining for example a regression problem) into square roots,

$$\mathbf{K} = \mathbf{W}^T \mathbf{W}, \tag{57}$$

i.e.,

$$\mathbf{K}(x, x') = \int dx''\, \mathbf{W}^T(x, x'') \mathbf{W}(x'', x'). \tag{58}$$

We will call the real square roots $\mathbf{W}$ *filter operators* and define the corresponding *filtered differences*

$$\omega(x) = \int dx'\, \mathbf{W}(x, x')[\phi(x') - t(x')]. \tag{59}$$

Using Eq. (59) the negative logarithm of a Gaussian prior like in Eq. (44) with mean function $t(x)$ and inverse covariance $\mathbf{K}(x, x')$ becomes, up to a ϕ–independent normalization term,

$$\begin{aligned} \frac{1}{2} \langle \phi - t \,|\mathbf{K}|\, \phi - t \rangle &= \frac{1}{2} \int dx\, dx'\, dx''\, [\phi(x) - t(x)] \mathbf{W}^T(x, x') \\ &\quad \times\ \mathbf{W}(x', x'')[\phi(x'') - t(x'')] \\ &= \frac{1}{2} \int dx\, |\omega(x)|^2. \end{aligned} \tag{60}$$

At this point we can introduce a local hyperfield by replacing the filtered difference $\omega(x)$ by a local mixture of two alternative filtered differences

$$\omega(x; \theta) = [1 - \theta(x)]\, \omega_1(x) + \theta(x)\, \omega_2(x), \tag{61}$$

with the real local hyperfield $\theta(x) \in [0, 1]$ contolling the mixture of the two ω_i. Such a hyperfield $\theta(x)$ can adapt the prior by selecting locally the best mixture of the two filtered differences ω_i. Similarly, additional hyperfields can be introduced to mix more then two filtered differences.

One possibility to transform an unrestricted real hyperfield $-\infty \le \widetilde{\theta}(x) \le \infty$ into a bounded real hyperfield $\theta(x) \in [0, 1]$ is given by

$$\theta(x) = \sigma(\widetilde{\theta}(x) - \vartheta), \tag{62}$$

with a threshold parameter ϑ and the sigmoidal transformation

$$\sigma(x) = \frac{1}{1 + \exp\left(-2\nu x\right)} = \frac{1}{2}(\tanh(\nu x) + 1). \tag{63}$$

In the limit $\nu \to \infty$ where the sigmoid in Eq. (63) approaches a step function we obtain a binary local hyperfield $\theta(x) \in \{0, 1\}$. In contrast to a 'soft mixing' with real functions with $0 \le \theta(x) \le 1$, a binary local hyperfield implements a 'hard switching' between alternative filtered differences ω_i.

For a prior depending on hyperfields

$$p(\phi|\theta) \propto \exp\left(-E(\phi|\theta)\right) \tag{64}$$

the prior energy $E(\phi|\theta)$ can be written

$$\begin{aligned} E(\phi|\theta) &= \frac{1}{2} \int dx |\omega(x; \theta)|^2 + \ln Z_\phi(\theta) \\ &= \frac{1}{2} \int dx \Big| [1 - \theta(x)]\omega_1(x) + \theta(x)\omega_2(x) \Big|^2 + \ln Z_\phi(\theta). \end{aligned} \tag{65}$$

The normalization factor

$$Z_\phi(\theta) = \int d\phi \, \exp\left(-\frac{1}{2}\int dx |\omega(x;\theta)|^2\right) \tag{66}$$

can depend on θ, if the filters $\mathbf{W}_i$ of the ω_i differ. If depending on θ the normalization factor $Z_\phi(\theta)$ has to be included when integrating over θ or solving for the optimal θ in Maximum A Posteriori Approximation. For binary θ mixed terms proportional to $\omega_1\omega_2$ vanish in Eq. (65), because for binary θ we have $\theta(1-\theta) = 0$. In that case we can write

$$E(\phi|\theta) = \frac{1}{2}\int dx \left([1-\theta(x)]|\omega_1(x)|^2 + \theta(x)|\omega_2(x)|^2\right) + \ln Z_\phi(\theta). \tag{67}$$

4.2.3 Mixing reference functions

A filtered difference $\omega(x;\theta)$ as in Eq. (61) can be obtained, for instance, by local mixing or switching between two alternative reference functions $t_1(x')$ and $t_2(x')$

$$t_x(x';\theta) = [1-\theta(x)]\, t_1(x') + \theta(x)\, t_2(x'), \tag{68}$$

where the local reference functions $t_x(x';\theta)$ are functions of x' and x: To obtain a filtered difference $\omega(x;\theta)$ at position x, one needs the reference function t_x for all x' for which the corresponding $\mathbf{W}(x,x')$ is nonzero,

$$\omega(x;\theta) = \int dx'\, \mathbf{W}(x,x')[\phi(x') - t_x(x';\theta)]. \tag{69}$$

Thus, for every local filtered difference the whole template function $t_x(x';\theta)$, rather than individual function values $t(x,\theta)$, have to be adapted.

In contrast to the local reference functions $t_x(x')$ used in Eq. (68) one global reference function $t(x')$ can be adapted locally using

$$t(x';\theta) = [1-\theta(x')]\, t_1(x') + \theta(x')\, t_2(x'). \tag{70}$$

On the other hand, working with different reference functions $t_{1,x}(x')$, $t_{2,x}(x')$ for different x generalizes Eq. (68) to

$$t_x(x';\theta) = [1-\theta(x)]\, t_{1,x}(x') + \theta(x)\, t_{2,x}(x'). \tag{71}$$

It is possible to use a nonlocal prior (i.e., a prior with $\mathbf{K} \neq \mathbf{I}$, for example, when choosing as $\mathbf{K}$ a differential operator because smooth functions are expected) and still avoid local template functions $t_x(x',\theta)$. This can be achieved by taking the product of a Gaussian prior with $\mathbf{K} = \mathbf{I}$ and a second Gaussian prior with a nondiagonal inverse covariance $\mathbf{K}_2$ and a zero (or fixed) reference function, yielding for the combined prior energy

$$\begin{aligned} E(\phi|\theta) &= \frac{1}{2}\left\langle \phi - \tilde{t}(\theta) \middle| \phi - \tilde{t}(\theta) \right\rangle + \frac{1}{2}\left\langle \phi \,|\mathbf{K}_2|\, \phi \right\rangle & (72)\\ &= \frac{1}{2}\Big(\left\langle \phi - t(\theta)\,|\mathbf{K}|\, \phi - t(\theta)\right\rangle + \left\langle \tilde{t}(\theta) \middle| \mathbf{I} - \mathbf{K}^{-1} \middle| \tilde{t}(\theta) \right\rangle \Big), & (73) \end{aligned}$$

The second term in Eq. (73) is independent of ϕ, the effective template $t(\theta)$ is given by

$$t(\theta) = \mathbf{K}^{-1}\tilde{t}(\theta), \tag{74}$$

and the effective inverse covariance $\mathbf{K}$ by

$$\mathbf{K} = \mathbf{I} + \mathbf{K}_2. \tag{75}$$

Choosing as inverse prior covariance $\mathbf{K}$ a differential operator, the effective $t(\theta)$ becomes a smoothed version of $\tilde{t}(\theta)$.

4.2.4 Mixing filter operators

Similar to Eq. (68) a mixed filtered difference $\omega(x;\theta)$ can be obtained by mixing locally two alternative filter operators $\mathbf{W}_1(x,x')$ and $\mathbf{W}_2(x,x')$

$$\mathbf{W}(x,x';\theta) = [1-\theta(x)]\mathbf{W}_1(x,x') + \theta(x)\mathbf{W}_2(x,x'). \tag{76}$$

Introducing

$$\mathbf{K}_x(\theta) = W_x(\theta)W_x^T(\theta) \tag{77}$$

with vector $W_x(\theta) = \mathbf{W}(x,\cdot\,;\theta)$, we obtain from Eq. (76) for binary $\theta(x)$ as inverse covariance

$$\begin{aligned} \mathbf{K}(\theta) &= \int dx\, \mathbf{K}_x(\theta) = \int dx\, W_x(\theta)W_x^T(\theta) \\ &= \int dx\, \left([1-\theta(x)]W_{1,x}W_{1,x}^T + \theta(x)W_{2,x}W_{2,x}^T\right). \end{aligned} \tag{78}$$

4.3 Hyperpriors for hyperfields

Working with hyperfields typically requires non–uniform *hyperpriors* $p(\theta)$. Indeed, allowing completely unrestricted functions t and operators $\mathbf{W}$ just eliminates the corresponding prior term. In nonparametric hyperfield models, where functions like $\theta(x)$ are not restricted to some parametric family, hyperpriors for hyperfields are stochastic processes, like priors for the functions $\phi(x)$ or $\phi(x,y)$. Such hyperpriors can, for instance, favor smooth hyperfields $\theta(x)$ analogous to a smoothness prior for a function ϕ. In analogy to (Euclidean) field theory in physics the part $E(\phi|\theta)$ may be interpreted as the 'interaction' between the fields ϕ and θ while the hyperprior describes a 'free' hyperfield including possible 'self–interactions' of the hyperfield.

As a smoothness prior for a real $\theta(x)$, for example, one can use a Laplacian prior with *hyperprior energy*

$$E(\theta) = -\frac{\tau}{2}\left\langle \theta\left|\Delta\right|\theta\right\rangle. \tag{79}$$

where τ is the analogue of a regularization constant. Parameters of the hyperprior like τ in Eq. (80) can be treated as higher level hyperparameters.

Another kind of 'smoothness' is implemented by the non–Gaussian hyperprior

$$p(\theta) \propto \exp\left(-\frac{\tau}{2}\int dx\, C_\theta(x)\right), \tag{80}$$

where

$$C_\theta(x) = \sigma\left(\left(\frac{\partial\theta}{\partial x}\right)^2 - \vartheta_\theta\right) \tag{81}$$

with a sigmoid function $\sigma(x)$ like in (63) and some constant τ. For $\nu \to \infty$, where the sigmoid approaches a step function, $C_\theta(x)$ becomes 0 at locations where the square of the first derivative is smaller than a certain threshold $0 \leq \vartheta_\theta < \infty$, and 1 otherwise. Similarly, one can penalize the number $N_d(\theta)$ of discontinuities, where $(\partial\theta/\partial x)^2 = \infty$, choosing

$$p(\theta) \propto \exp\left(-\frac{\tau}{2}N_d(\theta)\right). \tag{82}$$

In the case of a binary field this means counting the number of times the field changes its value.

Eq. (81) can be generalized to

$$C_\theta(x) = \sigma\left(|\omega_\theta(x)|^2 - \vartheta_\theta\right), \tag{83}$$

with filtered difference

$$\omega_\theta(x) = \int dx'\, \mathbf{W}_\theta(x,x')[\theta(x') - t_\theta(x')], \tag{84}$$

like in Eq. (59). The reference function $t_\theta(x')$ in Eq. (84) gives the expected form for the hyperfield $\theta(x)$, while the filter operator $\mathbf{W}_\theta$ defines the measure which is used to measure the distance of hyperfields $\theta(x)$ from the reference $t_\theta(x')$.

4.4 Auxiliary fields

When working with hyperfields one introduces additional degrees of freedom which influence the prior for ϕ. Integrating over the hyperfields $\theta(x)$ to obtain the predictive density in a full Bayesian approach would leave us with a prior $p(\phi)$ which is non–Gaussian in ϕ, even if all conditional priors $p(\phi|\theta)$ are Gaussian in ϕ. Similarly, when solving the problem in Maximum A Posteriori Approximation, the stationarity equation for ϕ (linear for Gaussian $p(\phi|\theta)$ for given θ) and the nonlinear stationarity equation for θ are coupled. Eliminating θ from the set of coupled equations leaves us with a nonlinear equation in ϕ. Hence, instead of introducing hyperfields as additional degrees of freedom one may try as well to directly formulate a non–Gaussian prior for ϕ.

One possibility to obtain a non–Gaussian prior from a Gaussian prior is to use as prior energy instead of the square ω^2 a non–quadratic function $\psi(\omega)$ of the filtered difference, corresponding to a prior

$$p(\phi|D_0) \propto \exp\left(-\int dx\, \psi(\omega(x))\right) \tag{85}$$

where for density estimation problems x can be replaced by the pair (x, y). Typical choices are 'cup' functions with flat tails for which one large step is cheaper than many small ones (see Fig. 8). Such non–Gaussian priors are, for example, used to deal with discontinuities in images [14, 3, 30, 13, 45, 44] or in the identification of 'outliers' like the separation of background and signal in experimental spectra [38, 9, 10].

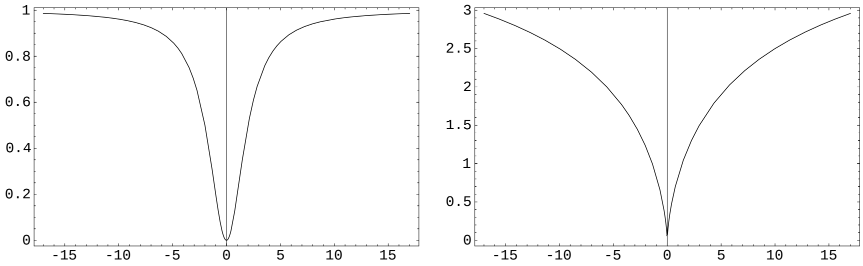

Figure 8: Two 'cup' functions of the form $\psi(x) = a(1.0 - 1/(1 + (|x - x_0|/b)^\gamma))$ [44] as example of a non–quadratic function $\psi(x)$ in Eq. (85). Left hand side: Winkler's cup function [43] ($a = 5$, $b = 10$, $\gamma = 0.7$, $x_0 = 0$). Right hand side: Function with cusp ($a = 1$, $b = 3$, $\gamma = 2$, $x_0 = 0$).

Another possibility to construct non–Gaussian priors is the introduction of *auxiliary fields* $B(x;\phi)$, or more general $B(x,y;\phi)$, whose function values are not independent variables but are defined as functionals of ϕ. (To simplify notation we will denote $B(x;\phi)$ by $B(x)$ or $B(\phi)$, depending on the context.) Similar to hyperfields $\theta(x)$ auxiliary fields can be used to select locally the best adapted filtered difference from a set of alternative ω_i, each of the form of Eq. (59). As an example in analogy to Eq. (83) consider an auxiliary field

$$B(x) = \sigma\left(u(x) - \vartheta\right), \tag{86}$$

with

$$u(x) = |\omega_1(x)|^2 - |\omega_2(x)|^2, \tag{87}$$

ϑ representing a threshold, and $\sigma(x)$ a sigmoidal function like in (63). The ω_i are filtered differences defined in terms of ϕ analogous to Eq. (59). In contrast to a hyperfield the auxiliary field $B(x)$ in Eq. (86) does not introduce new degrees of freedom because the ω_i in Eq. (87) are defined in terms of ϕ. Note that if $\omega_i(x)$ is nonlocal with respect to $\phi(x)$ then also $B(x)$ is nonlocal, meaning that a value $B(x)$ depends on more than one $\phi(x)$–value. For a negative Laplacian prior in one–dimension, i.e.,

$$\mathbf{K}(x,x') = -\delta(x,x')\frac{\partial^2}{\partial x^2} \tag{88}$$

(the δ-function is usually skipped from the notation) Eq. (86) becomes (for appropriate boundary conditions)

$$B(x) = \sigma\left(\left|\frac{\partial(\phi - t_1)}{\partial x}\right|^2 - \left|\frac{\partial(\phi - t_2)}{\partial x}\right|^2 - \vartheta\right). \tag{89}$$

While auxiliary fields $B(x)$ are directly determined by ϕ, hyperfields $\theta(x)$ are indirectly related to ϕ, for instance through the stationarity equations of a Maximum A Posteriori Approximation or through integration over $\theta(x)$ in a full Bayesian approach.

Auxiliary fields can be used similarly to hyperfields. They can help to adapt reference functions t or filter operators $\mathbf{W}$. For instance, to switch between two filtered differences one can use a binary $B(x)$

$$|\omega(x;B)|^2 = [1 - B(x)]|\omega_1(x)|^2 + B(x)|\omega_2(x)|^2, \tag{90}$$

yielding a prior energy of a form similar to Eq. (67)

$$E(\phi) = \frac{1}{2}\int dx\,\left([1 - B(x)]|\omega_1(x)|^2 + B(x)|\omega_2(x)|^2\right). \tag{91}$$

Resembling the role of hyperpriors $p(\theta)$, additional prior factors

$$p(B(\phi)) \propto \exp(-E_B(\phi)), \tag{92}$$

depending on ϕ only over $B(\phi)$, can be introduced. For example, if $N_d(B)$ counts the number of discontinuities of $B(x)$, the number of switchings is restricted by choosing

$$E_B(\phi) = \frac{\tau}{2} N_d(B). \tag{93}$$

For real $B(x)$ one can use terms of the form

$$E_B(\phi) = \frac{\tau}{2} \int dx |\omega_B(x)|^2 \tag{94}$$

where

$$\omega_B(x) = \int dx' \, \mathbf{W}_B(x, x')[B(x') - t_B(x')] \tag{95}$$

is a filtered difference of B with some filter operator $\mathbf{W}_B$ and template t_B. Non–quadratic energies as in (83) become, now written for B,

$$E_B(\phi) = \frac{\tau}{2} \int dx \, C_B(x), \tag{96}$$

with

$$C_B(x) = \sigma \left(|\omega_B(x)|^2 - \vartheta_B \right). \tag{97}$$

The normalization factor

$$Z = \int d\phi \, \exp\left(-E(\phi) - E_B(\phi)\right) \tag{98}$$

for a prior

$$p(\phi) \propto \exp\left(-E(\phi) - E_B(\phi)\right), \tag{99}$$

is by definition independent of ϕ and can thus be skipped for calculations in Maximum A Posteriori Approximation.

We have now encountered two methods for constructing prior models which may result in quite similar looking expressions. For instance, combining prior energy (91) with (93) for a binary auxiliary field (86) results in a prior

$$p(\phi) \propto \exp\left(-\frac{1}{2} \int dx \, \left([1 - B(x)] |\omega_1(x)|^2 + B(x) |\omega_2(x)|^2 \right) - \frac{\tau}{2} N_d(B) \right). \tag{100}$$

A similar looking prior model with a binary hyperfield can be obtained by combining a conditional Gaussian prior (67) with the hyperprior (82)

$$p(\phi, \theta) = p(\phi|\theta) p(\theta) \propto \tag{101}$$

$$\exp\left(-\frac{1}{2} \int dx \, \left[(1 - \theta(x)) |\omega_1(x)|^2 + \theta(x) |\omega_2(x)|^2 \right] - \frac{\tau}{2} N_d(\theta) - \ln Z_\phi(\theta) \right).$$

We can now compare the two models: Working with hyperfields means working with conditional priors $p(\phi|\theta)$, so that normalization factors which are in general θ–dependent.

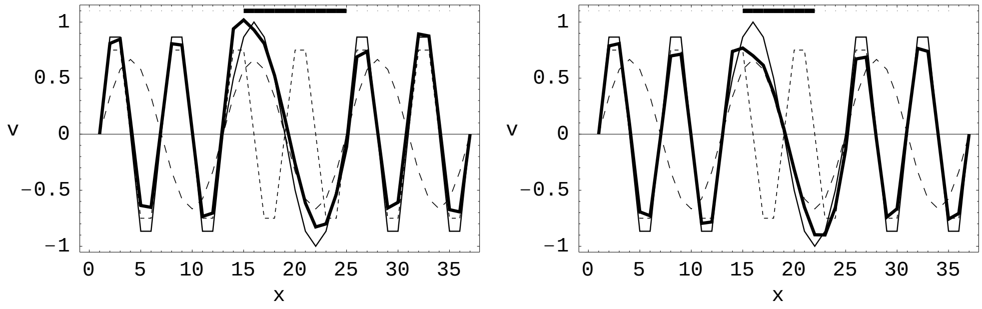

Figure 9: Example of the reconstruction of a potential $v(x)$ from coordinate measurements in a quantum system at finite temperature with likelihood (41) and with a binary auxiliary field (left figure) or a binary hyperfield (right figure), respectively, switching locally between two alternative mean functions (or reference potentials) $t_1(x) = v_1(x)$ and $t_2(x) = v_2(x)$ of a nonparametric Gaussian smoothness prior [24]. In both figures the switching between the two reference potentials v_1, v_2 (induced by the binary hyperfield or the binary auxiliary field, respectively) is indicated by a thick black line above the potentials. Left hand side: Reconstruction with a binary auxiliary field defined in terms of $\phi(x) = v(x)$ as step function $B(x) = \Theta\left(|\omega_1(x)|^2 - |\omega_2(x)|^2\right)$, (which corresponds to Eq. (86) with $\vartheta = 0$ in the limit where the sigmoid approaches a step function), using a prior like in (91) switching between two alternative filtered differences and a penalty on the number of jumps of the auxiliary field. Right hand side: Reconstruction with a binary hyperfield which switches locally between the two nonzero reference potentials v_1, v_2 and a penalty on the number of jumps of the hyperfield as hyperprior.

Therefore the normalization factors have to be included for calculations in Maximum A Posteriori Approximations. This is not the case if we are working with auxiliary fields. Hence, in general MAP solutions for B, $N_d(B)$, and C_B, are different from the MAP solutions for θ, $N_d(\theta)$, and C_θ. If, however, the filtered differences ω_i in Eq. (101) differ only in their reference functions $t_i(x)$, then the normalization term can be skipped. The two MAP equations are then indeed equivalent for $\theta(x) = \Theta\left(|\omega_1(x)|^2 - |\omega_2(x)|^2\right)$, if the threshold vanishes $\vartheta = 0$ and a hyperprior or additional p_B is absent, i.e., if $p(\theta) \propto 1$ and $p(B) \propto 1$. Furthermore it is also easily seen that in that case $\theta(x) = \Theta\left(|\omega_1(x)|^2 - |\omega_2(x)|^2\right)$ is a selfconsistent solution for θ for every given ϕ. Fig. 9 shows two reconstructions of a quantum potential, one using a hyperfield and another one using an auxiliary field.

5 Summary

In nonparametric Bayesian models, where the number of the degrees of freedom of the likelihood is much larger than the number of available data points, the quality of learning does depend essentially on the implemented *a priori* information. Starting from Gaussian process priors we have discussed several methods to implement available *a priori* information explicitly in terms of the likelihood, for example, by choosing a specific prior covariance which corresponds to some approximate symmetry of the likelihood or by adapting the mean function of a prior process using hyperfields. Those techniques for constructing nonparametric priors can be used for many different likelihood models, including density estimation, regression and inverse quantum theory.

References

[1] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer–Verlag, New York, 2nd edition, 1985.

[2] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, 1995.

[3] A. Blake and A. Zisserman. *Visual Reconstruction*. MIT Press, Cambridge, MA, 1987.

[4] N. Bleistein and N. Handelsman. *Asymptotic Expansions of Integrals*. Dover, New York, 1986. Originally published in 1975 by Holt, Rinehart, and Winston, New York.

[5] J. Cardy. *Scaling and Renormalization in Statistical Physics*. Cambridge University Press, Cambridge, 1996.

[6] B. P. Carlin and T. A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*, volume 69 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, 1996.

[7] N. G. de Bruijn. *Asymptotic Methods in Analysis*. Dover, New York, 1981. Originally published by North–Holland Publishing Co., 1958, Amsterdam.

[8] J. L. Doob. *Stochastic Processes*. Wiley, New York, 1953. New edition published 1990.

[9] V. Dose and W. von der Linden. Outlier tolerant parameter estimation. In W. von der Linden, V. Dose, R. Fischer, and R. Preuss, editors, *Maximum Entropy and Bayesian Methods*, Dordrecht, 1999. Kluwer Academic Publishers.

[10] R. Fischer, K. M. Hanson, V. Dose, and W. von der Linden. Background estimation in experimental spectra. *Phys. Rev. E*, 61:1152, 2000.

[11] C. W. Gardiner. *Handbook of Stochastic Methods*. Springer–Verlag, Berlin, 2nd edition, 1990.

[12] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, New York, 1995.

[13] D. Geman and G. Reynolds. Constraint restoration and the recover of discontinuities. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 14:367–383, 1992.

[14] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984. Reprinted in Shafer & Pearl, editors, 1990. *Readings in Uncertainty Reasoning*. Morgan Kaufmann, San Mateo, CA, 1990.

[15] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural network architectures. *Neural Computation*, 7(2):219–269, 1995.

[16] J. M. Hammersley and D. C. Handscomb. *Monte Carlo Methods*, volume 3 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1964.

[17] J. Honerkamp. *Stochastic Dynamical Systems. Concepts, Numerical Methods, Data Analysis*. VCH Verlag, New York, 1994.

[18] J. Honerkamp. *Statistical Physics*. Springer–Verlag, New York, 1998.

[19] D. P. Landau and K. Binder. *A Guide to Monte Carlo Simulations in Statistical Physics*. Cambridge University Press, Cambridge, 2000.

[20] J. C. Lemm. Bayesian field theory. Technical Report MS-TP1-99-1, Universität Münster, 1999. arXiv:physics/9912005.

[21] J. C. Lemm. Mixtures of Gaussian process priors. In *ICANN 99. Proceedings of the 9th International Conference on Artificial Neural Networks, Edinburgh, UK, September 7–10, 1999*, London, 1999. IEEE Conference Publication.

[22] J. C. Lemm. Inverse time–dependent quantum mechanics. *Phys. Lett. A*, 276(1–4):19–24, 2000. arXiv:quant-ph/0002010.

[23] J. C. Lemm and J. Uhlig. Bayesian inverse quantum theory. *Few–Body Systems*, 29:25–52, 2000. arXiv:quant-ph/0006027.

[24] J. C. Lemm, J. Uhlig, and A. Weiguny. Bayesian reconstruction of approximately periodic potentials for quantum systems at finite temperatures. *The European Physical Journal B*, 20(3):349–366, 2001. MS-TP1-00-4, arXiv:quant-ph/0005122.

[25] D. J. C. MacKay. Bayesian methods for backpropagation networks. In E. Domany, J. L. van Hemmen, and K. Schulten, editors, *Models of Neural Networks III*, pages 211–254. Springer–Verlag, New York, 1996.

[26] D. J. C. MacKay. Bayesian non–linear modeling for the prediction competition. In G. Heidbreder, editor, *Maximum Entropy and Bayesian Methods, Santa Barbara 1993*, pages 221–234, Dordrecht, 1996. Kluwer Academic Publishers.

[27] D. J. C. MacKay. Comparison of approximate methods for handling hyperparameters. *Neural Computation*, 11:1035–1068, 1999. A preprint version is available at `http://wol.ra.phy.cam.ac.uk/mackay/alpha.ps.gz`.

[28] A. Messiah. *Quantum Mechanics*. North–Holland, Amsterdam, 1961.

[29] I. Montvay and G. Münster. *Quantum Fields on a Lattice*. Cambridge University Press, Cambridge, 1994.

[30] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. Pure Applied Math.*, 42:577–684, 1989.

[31] R. M. Neal. *Bayesian Learning for Neural Networks*. Springer–Verlag, New York, 1996.

[32] R. M. Neal. Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Technical Report 9702, Department of Statistics, University of Toronto, Canada, 1997.

[33] J. W. Negele and H. Orland. *Quantum Many–Particle Systems*, volume 68 of *Frontiers in Physics Series*. Addison–Wesley, Redwood City, CA, 1988.

[34] C. P. Robert. *The Bayesian Choice*. Springer–Verlag, New York, 1994.

[35] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2001.

[36] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*. North–Holland, Amsterdam, 1992.

[37] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[38] W. von der Linden, V. Dose, J. Padayachee, and V. Prozesky. Signal and background separation. *Phys. Rev. E*, 59:6527–6534, 1999.

[39] J. von Neumann. *Mathematical Foundations of Quantum Mechanics*. Princeton University Press, Princeton, NJ, 1955.

[40] G. Wahba. *Spline Models for Observational Data*. SIAM, Philadelphia, 1990.

[41] G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In B. Schölkopf, C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods: Support Vector Machines*, pages 69–88. MIT Press, Cambridge, MA, 1998.

[42] C. K. I. Williams and C. E. Rasmussen. Gaussian processes for regression. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 514–520, Cambridge, MA, 1996. MIT Press.

[43] G. Winkler. *Image Analysis, Random Fields, and Dynamic Monte Carlo Methods*. Springer–Verlag, Berlin, 1995.

[44] S. C. Zhu and D. Mumford. Prior learning and Gibbs reaction–diffusion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(11):1236–1250, 1997.

[45] S. C. Zhu, Y. N. Wu, and D. Mumford. Minimax entropy principle and its application to texture modeling. *Neural Computation*, 9(8):1627–1660, 1997.