How to Implement A Priori Information: A Statistical Mechanics Approach *

Jörg C. Lemm [a]

[a] *Institute for Theoretical Physics I*
Münster University
Wilhelm–Klemm–Str. 9
D-48149 Münster, Germany
http://pauli.uni-muenster.de/~lemm
E-mail: lemm@uni-muenster.de

A new general framework is presented for implementing complex a priori knowledge, having in mind especially situations where the number of available training data is small compared to the complexity of the learning task. A priori information is hereby decomposed into simple components represented by quadratic building blocks (quadratic concepts) which are then combined by conjunctions and disjunctions to built more complex, problem specific error functionals. While conjunction of quadratic concepts leads to classical quadratic regularization functionals, disjunctions, representing ambiguous priors, result in non–convex error functionals. These go beyond classical quadratic regularization approaches and correspond, in Bayesian interpretation, to non–gaussian processes. Numerical examples show that the resulting stationarity equations, despite being in general nonlinear, inhomogeneous (integro–)differential equations, are not necessarily difficult to solve. Appendix A relates the formalism of statistical mechanics to statistics and Appendix B describes the framework of Bayesian decision theory.

Keywords: Ambiguous a priori information, Bayesian statistics, regularization approaches, saddle point approximation, mixture and polynomial models for prior density.

Contents

1 Introduction **3**

2 Quadratic concepts **6**
2.1 Definitions 6
2.2 Templates 10
2.3 Covariances and symmetries 11
2.4 From posterior probability to likelihood energy 13

* Report-no: MS-TP1-98-12

3 Combinations of quadratic concepts **16**
3.1 AND: Classical regularization and gaussian processes 16
3.2 OR: Mixture models 21
3.3 An example with ambiguous prior 26
3.4 OR: Landau–Ginzburg regularization and interacting systems . . . 28
3.5 OR: Combination of methods and continuous transformations . . . 31

4 Learning **35**

5 Conclusion **37**

A Statistics and Statistical Mechanics **38**
A.1 Probability 40
A.1.1 Normalization factors or partition sums 40
A.1.2 Log–probabilities, bit numbers, and free energies 41
A.2 Random variables 42
A.2.1 Averages 42
A.2.2 Information 42
A.2.3 Energy 43
A.3 Temperature and external fields 45
A.3.1 Maximum entropy and Boltzmann–Gibbs distributions . . . 46
A.3.2 Annealing methods 47
A.3.3 Generating functions 47
A.4 Conditional probabilities and disordered systems 55

B Bayesian decision theory **58**
B.1 Basic definitions 58
B.1.1 The basic model 58
B.1.2 Learning: predictive and posterior density 59
B.1.3 The risk functional 60
B.2 Interpretation of priors 61
B.2.1 Measured and factorial priors 61
B.2.2 Gaussian likelihoods 62
B.2.3 Generative priors 64
B.2.4 More complex models 65
B.2.5 Structural priors and a posteriori control 66
B.3 Quantum Mechanics 67
B.3.1 Density operators 67
B.3.2 Quantum statistics 69
B.3.3 Quantum mechanical scattering 70
B.3.4 Disordered systems 71
B.3.5 Path integrals 72
B.4 Bayes' rule for complete data 73

B.4.1 Posterior probabilities and likelihoods 73
B.4.2 Posterior and likelihood energies 74
B.5 Saddle point approximation 75
B.5.1 Maximum a posteriori approximation 75
B.5.2 Complete maximum likelihood approximation 77
B.5.3 Loop expansion 79
B.5.4 Stationarity equations 81
B.6 Approximation problems 82

1. Introduction

In the setting of empirical learning training data are used to predict the outcome of future test situations. The principal problem of empirical learning can already be seen in a noise free toy example: Assume we have measured $h(x_1)$ $= 2$ (training data) and have to predict $h(x_2)$ (test data). Clearly, this is a hopeless task unless we know some relations between $h(x_1)$ and $h(x_2)$ [74]. If, however, we know a relation, for example $h(x_2) - h(x_1) = 5$, the task becomes solvable. Such relations do not have the form of standard training data and must be provided by what we will call *a priori information*.

The simple example clarifies two points which are also valid for more complex scenarios: 1. learning is technically merely a reformulation of available (training and a priori) knowledge, and 2. the success of learning is essentially based on empirical control of the implied dependencies between test and training data.

One may now distinguish two principal possibilities of implementing a priori information in learning systems: 1. The dependencies between test and training data can be implemented by restricting the space of possible functions $h(x)$, for example by choosing a parameterized form for $h(x)$. Examples include linear regression models or (finite) neural networks. hereby it is often difficult to interpret such implemented priors in terms of dependencies of function values $h(x)$ [72]. 2. Dependencies between test and training data may also be directly expressed in terms of the functions values $h(x)$ itself, like we have done in the above toy example. This is the approach used for regularization terms in empirical risk minimization and for stochastic prior processes in Bayesian statistics. Here, for example, a smoothness term like

$$\int dx \left(\frac{\partial h(x)}{\partial x} \right)^2 \tag{1}$$

within an error functional can provide the necessary dependencies.

To control and adapt the generalization behavior of learning it seems helpful to treat a priori information as explicit as possible. We choose in this paper therefore the second possibility of *explicit implementation* of a priori information in terms of the function values $h(x)$.

Furthermore, we consider especially low data situations where the number of available training data is small compared to the complexity of h. We do this because those are the cases where a priori information becomes the essential input. Contrasting the well known uninformative priors of Bayesian statistics [5] one may call this a situation with *informative priors*.

As a typical problem, consider an image completion task where only some of the pixels are given and the complete image has to be completed. If we expect for example the image of a face then a priori information to be implemented includes the expectation that a face has two eyes, a mouth and a nose and typical distances between that constituents. Similar problems are times series predictions with a priori informations concerning typically expected relatively well defined structures. Informative priors are also useful in object recognition or pattern classification when the object is relatively complex compared to the number of available training data. For a face detector, for example, a priori information can be implemented in form of a crude a priori model of what a face is which is then refined by the available training data.

In practice, a priori information is usually given in qualitative rather than in quantitative form. The approach used in this paper assumes a priori information stated in form of logical statements like h has (probably) property A AND B (e.g., a face has eyes AND mouth) or property A OR B (eyes may be open OR closed). In a first step properties of h are quantified by so called quadratic concepts introduced in Section 2. Quadratic concepts consist hereby of a prototype or function template t for h, and a corresponding distance measure defining $||t - h||$. In a second step an error functional is constructed implementing the logical statements which represent the given a priori information. This is done in Section 3. An optimal approximation is finally found by minimizing the error functional (Section 4).

Combining quadratic concepts by AND (Section 3.1) yields well known quadratic regularization functionals. Not common is the explicit inclusion of 'continuous data' or template functions t representing approximate reference models for h.

OR–like combinations, however, give rise to non–convex error functionals going beyond classical regularization approaches (Section 3.2). Numerically especially interesting are OR–like combinations of quadratic concepts with equal distance and only differing template functions. Those can be treated without calculating normalization factors which are difficult to obtain. This is exemplified in Section 3.3. Section 3.4 discusses an alternative implementation of OR–like combinations.

In case the number of properties combined by OR is too large to be treated exactly additional approximations are necessary which are discussed in Section 3.5. This is for example the case if an expected structure can be arbitrarily translated or otherwise continuously transformed. An eye, for example, can appear in different shapes/scales/positions and only those variants most consistent with the training data will then be included explicitly.

The proposed approach provides a *interface between symbolic methods and statistics*[1]: A priori information is decomposed by symbolic/logical operations into components, simple enough to be quantified in form of quadratic concepts. Quadratic concepts are then combined, modeling the symbolic operations, resulting in an error functional which is then treated by non–symbolic methods.

Finally, let us add three remarks concerning the numerical requirements of the approach, the language of statistical physics, and the interpretation of error functionals.

Explicit implementation of a priori information is in generally numerically extensive. Due to increasing computational resources numerical methods used in statistical physics or field theory, e.g., Monte Carlo calculations, become more and more applicable to learning problems. Up to now this is mainly the case for one or two dimensional problems, like for example in Bayesian image reconstruction [55, 22,71,76]. In particular, the stationarity equations to be solved to minimize the presented error functionals can in general be nonlinear inhomogeneous integro–differential equations. They are however not necessarily difficult to solve as can be seen in Section 3.3. The reason is that the nonlinearity, in particular the number of minima, is under explicit control and there are limiting cases (at 'high and low temperatures') where the equations become linear. In any case, the presented error functionals can be utilized as a well defined starting point for further approximations.

Due to the background of the author the paper is written mainly in the language of statistical mechanics. Of course, there exists alternative formulations such as function approximation theory or empirical processes, which might be better suited for some purposes. The advantage of using a statistical mechanics formulation, however, is that it provides easy contact to approximations (e.g., saddle point approximation, perturbation theory, high temperature expansions [31,51,77]) and numerical algorithms (e.g., Monte Carlo algorithms [8,9,21,50]) well known from statistical mechanics or statistical field theory and especially useful for large systems. In Appendix A the language of statistics is related to that of statistical mechanics.

Like regularization approaches in empirical risk minimization we use an error functional to find an optimal approximation. We remark, however, that the error functional has a different interpretation from the view point of empirical risk minimization and that of Bayesian statistics. In empirical risk minimization the error functional represents the regularized form of an empirical risk with data terms being an empirical estimate of an expected risk. In the Bayesian interpretation the error function is related to the negative logarithm of the posterior density of h given the training data. Minimizing the error function is then equivalent to a maximum a posteriori approximation under log–loss. As we use the Bayesian interpretation to built up error functionals the necessary background is provided in Appendix B.

2. Quadratic concepts

2.1. Definitions

In this Section quadratic concepts are defined as fundamental building blocks to construct regularization functionals. We concentrate on function approximation or regression problems. Hereby an unknown function or true state of Nature h_N is approximated by a function h chosen from a model space $\mathcal{H}$ using training data $D_T = \{(x_i, y_i)|i \le 1 \le n\} = (x_T, y_T)$. Training data are assumed to consist of n pairs of independent variables x (describing the kind of measurement performed) and dependent variables y (responses or measured values). In classical regularization theory [66,70] the error to be minimized is given by a regularization functional $E(h)$ (see Appendix B for a Bayesian interpretation) which contains the mean–square error terms for training data and an additional prior term, in many cases related to smoothness. We denote prior information by D_0 so $D = D_T \cup D_0$ represents all available data we have. Application to density estimation problems (having an additional normalization constraint and logarithmic terms replacing the mean–square data terms) or classification problems (being a special case of function approximation with integer y) poses no principal new problems and will be reported elsewhere.

Example 1 (Image completion) Consider an image completion or image reconstruction task [55,22,71,76]. Hereby an image h is drawn from a set of images $\mathcal{H}$. Then, given noisy observations $y_i(x_i)$ for some pixels x_i of h a reconstructed and completed image $y = h(x) \in \mathcal{H}$ should be returned by the learning system. For grey level images h is a vector of real numbers containing a two dimensional $n_1 \times n_2$ array of grey level values $g(k,l)$, i.e., with j–components $h_j = g(k,l)$ with $1 \le k \le n_1$, $1 \le l \le n_2$, $1 \le j \le n_1 n_2$ and, for example, $j = k + n_1(l-1)$. Assume now one knows that the image is that of a face. For the class of faces experts can contribute information by verbally describing prototypical forms of constituents (e.g. eyes, nose, mouth), their variants (e.g. open vs. closed mouth, translated, scaled, or deformed eyes) and relations (e.g. typical spatial distances). The related concepts like 'eye' or 'typical distance between eyes', are here linguistic or 'fuzzy' variables [38] which must be quantified to enter a regularization functional. As prototype or template t of an 'eye' an image or drawing of an eye can be chosen. A distance $d(t,h)$ between a reconstructed image h and an 'eye' template t can be defined pixel-wise, e.g. by

$$d^2(t,h) = \sum_x (h(x) - t(x))^2, \tag{2}$$

or, more generally, by a real symmetric, positive definite kernel

$$d_K^2(t,h) = \sum_{x,x'} (h(x) - t(x)) K(x,x') (h(x') - t(x')). \tag{3}$$

The sum over pixels x may be restricted to regions where eyes are expected. Furthermore, eyes may be open OR closed, appear in different sizes and at varying locations. Hence an eye is represented not by a single template but by a set of templates describing the variants in which an eye can appear. For continuous variations we will call such a set a deformable or adaptive template. It is described, for example, by scaling, translation, or deformation parameters. Then the image of a face, incomplete and disturbed by noise, should be reconstructed by approximating the given noisy data points of the incomplete image AND typical constituents of a face and their spatial relations in either one of their variants. Thus, a missing eye should be reconstructed at a typically expected position depending, for example, on the approximated position of mouth and nose, and in a form depending on the form of another, possibly visible eye.

The model treated in Section (3.3) represents a simple example for such a task (for a one–dimensional image or time series, respectively). For comparison, we discuss shortly a possible use of template functions for classification tasks (not treated in this paper):

Example 2 (Face detection) Consider a face detection task. Hereby the independent variable x is drawn from a set of images $\mathcal{X}_R$ by an unknown mechanism which should be approximated by a function $h(x)$. Then, given an image $x \in \mathcal{X}_R$ the output $y \in \{0, 1\}$ of the learning system should indicate whether it is the image of a face, e.g. by $y = h(x) = 1$, or not, e.g. by $y = h(x) = 0$. In contrast to the previous example now x is a vector of real numbers containing a two dimensional array of grey level values. A function template t now represents a prototoype for the classification function h. Classification templates t_j entering the error functional can now, for example, be constructed by reference to image templates t^x representing prototypical face/non–face input vectors x or parts of it. Thus, $E(h) = E(h, \{t_j\})$ where classification templates t_j are defined with the help of image templates $t_j = t_j(x, \{t_i^x\})$. For example face templates t^x for images x can be constructed as described in the previous example. Then, in a second step statements like: 'A face has eyes, nose and mouth in either one of several variations' are expressed in terms of distances d_i^x from image x to image templates t^x. A classification templates t could for example be a simple threshold function responding with $t(\{d_i^x\}) = 1$ if some function of distances $\{d_i^x\}$ to typical faces are below a certain threshold. Finally, distances $d(h, t)$ between classification functions $h \in \mathcal{H}$ have to be chosen to build an error functional $E(h)$ to be minimized. (We refer to Appendix B.6 for the distinction between specifying a reconstruction model, i.e., the probability of face vs. non–face given an image, and specifying a generative model, i.e., the probability of an image provided it represents a face or non–face.)

In general, a *concept* is based upon

1. a *template* $t(x)$ representing a prototype for function h and
2. a *distance* $d(h,t)$ measuring the similarity of a function h to the template t. For practical reasons, a quadratic distance is especially convenient. Concepts with quadratic distances will be called quadratic concepts.
3. A template t can be restricted to a subset of $\mathcal{X}_R$ with the help of a *projection operator*.
4. Complex concepts are constructed as *combination of quadratic concepts*. In this paper variants of AND and OR–like combinations are used.

Now we come to the formal definition of a quadratic concept. Let $\mathcal{H}$ be a Hilbert space of possible hypothesis functions h, and let angular brackets denote scalar products and matrix elements of symmetric operators, e.g. $\langle t|h\rangle = \int d^dx\, t(x)h(x)$ and $\langle t\,|K|\,h\rangle = \int d^dx\, d^dx' t(x)K(x,x')h(x')$ for d–dimensional x. (We will often write in the following simply $\int dx$ also for d–dimensional integrals.) Recall also that a real operator K is positive definite (semi positive definite) if it can be diagonalized, i.e., in terms of eigenfunctions Φ_k of K (with transpose Φ_k^T)

$$K = \sum_k \lambda_k \Phi_k \Phi_k^T, \tag{4}$$

and all eigenvalues $\lambda_k > 0$ ($\lambda_k \geq 0$) are positive (non–negative). Thus, $K = O^T D_K O$ with O orthogonal (i.e., $O^T O = I$ with identity I and transpose O^T) and diagonal operator $D_K > 0$ ($D_K \geq 0$). Positive definite operators define a scalar product by $\langle t|h\rangle_K = \langle t\,|K|\,h\rangle$. Positive (semi) definite operators K can be decomposed $K = W^T W = (OW)^T(OW)$, with real W, invertible if K positive definite, and arbitrary orthogonal O. A quadratic concept is defined as follows:

Definition 1 *(Quadratic concept) A quadratic concept is a pair (t,K) consisting of a template function $t(x) \in \mathcal{H}$ and a real symmetric, positive semi–definite operator, the concept operator K with eigenfunctions (features) Φ_k and eigenvalues (feature weights) λ_k. We call W a concept filter if $K = W^T W$. The operator K defines a concept distance:*

$$d^2(h) = d_K^2(t,h) = \langle h-t\,|K|\,h-t\rangle = ||h-t||_K^2 \tag{5}$$

$$= \langle W(h-t)|W(h-t)\rangle = ||W(h-t)||^2 \tag{6}$$

$$= \sum_k \lambda_k \langle h-t|\Phi_k\rangle \langle \Phi_k|h-t\rangle = \sum_k \lambda_k \;||\, \langle h|\Phi_k\rangle - \langle t|\Phi_k\rangle \,||^2, \tag{7}$$

on subspaces where it is positive definite. The maximal subspace in which the positive semi–definite K is positive definite is the concept space H_K of K. The corresponding hermitian projector P_K in this subspace H_K is the concept projector.

There exists a correspondence between quadratic forms and gaussian processes [70,45,50]. Let x be elements of (the dual of) $\mathcal{H}$, i.e., $h(x)-t(x) = \langle x|h-t\rangle = \left\langle K^{-1}x \left| K \right| h-t\right\rangle = \langle K^{-1}x|h-t\rangle_K$, is a bounded functional according to the Riesz representation theorem. In the context of stochastic processes $\mathcal{H}$ is known as reproducing kernel Hilbert space for x with reproducing kernel K^{-1} [70]. (The term reproducing kernel Hilbert space does not name a certain subclass of Hilbert spaces. Indeed, all separable Hilbert spaces are isomorphic for equal dimension. It characterizes the representation by the coordinates x. For example, the space $\mathcal{L}_2$ of square integrable functions $h(x)$ is not a reproducing kernel Hilbert space with respect to the coordinates x. This is reflected by the fact that functions $h(x) \in \mathcal{L}_2$ are not even defined pointwise.) For reproducing Hilbert spaces the scalar product with kernel K can be related to a covariance of zero mean gaussian variables. Interpreting t as data y_D obtained in situation $x_D = K$ (i.e., measuring features Φ_k weighted by λ_k) we write

$$p(y_D|x_D, h) = p(t|K, h) \propto e^{-d_K^2(t,h)/2} \tag{8}$$

for a distribution of t with mean h and covariance operator K^{-1}. (For a interpretation as posterior $p(h|D)$ see Section 2.4.) The kernel function $K^{-1}(x, x')$ of the covariance operator is also known as Green's function of K, propagator, or two-point correlation function. Hence, an error or energy functional

$$E(h) = \frac{d^2(h)}{2} \tag{9}$$

corresponds up to a constant to a negative log–probability. For approximation problems minimizing an error functional $E(h)$ can, from a Bayesian point of view, be interpreted as a maximimum–a–posteriori approximation (see Appendix B.6).

We remark that we will not discuss in the following problems of infinite spaces. This holds especially for the nonlinear models in the next sections for which the question of a continuum limit is highly non-trivial [32,77]. Hence, if necessary, integrals can in the following be considered as convenient notation for sums. Analogously, derivative operators can be replaced by their discretized lattice versions. This reflects also the fact that finally numerical calculations have to be done in a finite dimensional space.

The next two examples show that the definition of a quadratic concept includes the standard regularization functionals. The first example is a discrete concept with 'trivial' concept distance, the second a continuous concept with 'trivial' template function.

Example 3 (Data template) A standard mean–square error term used for regression is

$$(y_j - h(x_j))^2 = \int_{-\infty}^{\infty} d^d x\, \delta(x - x_j)(h(x) - t_j(x))^2 = \left\langle h - t_j \left| P_j \right| h - t_j \right\rangle. \tag{10}$$

Thus, such a mean–square error term corresponds to a quadratic concept with concept operator with kernel $P_j(x, x') = \delta(x - x_j)\delta(x - x')$ and (data) template t_j is the constant function $t_j(x) \equiv y_j$. The measured features are $h(x) = \langle h|x\rangle$.

Example 4 (Prior template) A typical smoothness functional in regularization theory, corresponding to the Wiener measure for stochastic processes [16,20,34,45] and to the kinetic energy or a free massless scalar Euclidean field in physics [25,32,77], is

$$\int_{-\infty}^{\infty} d^d x \sum_{l=1}^{d} \left(\frac{\partial h(x)}{\partial x_l} \right)^2 = - \left\langle h - t_0 \left| \Delta \right| h - t_0 \right\rangle. \tag{11}$$

Here partial integration has been used under the assumption of vanishing boundary terms. This quadratic concept has the zero function $t_0(x) \equiv 0$ as template. It is a sum of terms with concept filters $\partial/\partial x_l$ which are the generators of infinitesimal translations in d dimensions. Hence, the functional represents a measure of approximate infinitesimal translational symmetry. The concept operator is the negative (semi) definite d–dimensional laplacian with kernel

$$\Delta(x, x') = \delta(x - x') \sum_{l=1}^{d} \frac{\partial^2}{\partial x_l^2}. \tag{12}$$

In the following we will mainly be interested in the non–standard case of a combination of discrete training data with several prior concepts with non–zero template functions. Firstly, we discuss in more detail the two main ingredients of a concept: templates and distances.

2.2. Templates

Template functions can be constructed in various ways. The following list gives some potential applications of template functions in different contexts:

1. (Direct construction by experts) A template can be directly constructed by experts. For financial time series, for example, often an expected trend is included [30].
2. (Combination and extension of arbitrary learning methods) More generally, a template $t(x)$ can be the output of an arbitrary, parametric or non–parametric learning method trained for the same problem. For example, the prototype $t(x)$ can be a rule–based expert system, a regression tree, a neural network or a simple linear regression (for a collection and comparison of methods see for example [52]). Because a prior concept can depend on training data D_T (see Appendix B.4.1) such a $t(x)$ can be obtained by training with the same data D_T which are also used to determine the optimal approximation h^*. If one

wants $t(x)$ to be independent of the training data D_T it can alternatively be obtained by training with an independent set of training data.

3. (Transfer) The template $t(x)$ can also be the result of a learning algorithm for a similar problem and therefore be used to transfer knowledge between tasks. Such a transfer template can be adapted and restricted to certain subspaces by using concept projectors. The new solution h adapts the transfer template t to the new situation according to the new training data and other, additional prior templates.
4. (Learning history) In all cases a template $t(x)$ can be seen to contain in a compressed form the *learning history* prior to the new training data. This allows for example to construct on–line learning procedures. Hereby an intermediate solution h is obtained by using only a part $D_1 \subset D_T$ of the available training data D_T. Then the intermediate solution h is chosen as additional template t and learning is continued with a new data set $D_2 \subset D_T$.
5. (Sampling) A template corresponds to the mean of a gaussian process. For a finite number of x it can therefore be approximated by a sample mean, if samples for $h(x)$ are available. This can be, for example, a set of complete images (or images of constituents) for usage in image reconstruction. Thus, let t_α denote a sample for (a part of) $h(x)$. Then, if $p(h)$ can be approximated by a gaussian, the empirical mean $t = \sum_\alpha t_\alpha$ is a natural candidate for template. Similarly, multimodal distributions $p(h)$ can be approximated by a mixture of gaussians. In that case not only one but a set of centers, i.e., templates t_i has to be chosen (see Section 3). The centers t_i can be obtained by clustering methods.

Clearly, these fields have long publication histories and the advantages and disadvantages of using templates still have to be investigated. In this paper we concentrate mainly on the first possibility.

2.3. Covariances and symmetries

The concept kernel, respectively its inverse the covariance operator, are often related to approximate symmetries of a problem. Sometimes, also a finite rank approximation can be obtained by sampling.

Frequently prior information has the form of symmetries which $h(x)$ has to fulfil, exactly or approximately. We already have seen in Example 4 that approximate symmetry under infinitesimal translations is related to smoothness. The implementation of approximate symmetries requires the definition of a distance to exact symmetry.

We can write for a positive (semi)definite concept operator $K = W^T W = (I - (W - I))^T (I - (W - I))$ with identity I, and concept filter W. Now consider operators S which just change the argument x of $h(x)$ into $\sigma(x)$, i.e.,

$$Sh(x) = h(\sigma(x)). \tag{13}$$

If the transformation $\sigma(x)$ is one–to–one then S permutes the function values and S is a symmetry transformation. Hence, for $K = (I-S)^T(I-S)$, i.e., $S = W - I$ and a zero template $t \equiv 0$ the corresponding squared K–norm

$$||h||_K^2 = \langle h\,|K|\,h\rangle = \langle h - Sh|h - Sh\rangle = ||h - Sh||_I^2 = d_S^2(h) \tag{14}$$

compares h with Sh and measures therefore a degree of symmetry.

A bit more generally, we call a concept operator $K = (I-S)^\dagger K_S(I-S)$ a symmetry concept operator, if S is a symmetry operator. Here K_S can be some positive (semi–)definite operator usually taken equal to the identity I, and the hermitian conjugate $S^\dagger$ is equal to the transpose S^T for real matrices. With zero template we have

$$d_S^2 = \langle h - Sh\,|K_S|\,h - Sh\rangle = \left\langle h \left| (I-S)^\dagger K_S (I-S) \right| h \right\rangle. \tag{15}$$

Continuous symmetries are represented by Lie groups which locally can be written as exponential of m generators $s = (s_1, \cdots, s_m)$ parameterized by a vector θ

$$S(\theta) = e^{\langle\theta|s\rangle} = \sum_{k=0}^{\infty} \frac{1}{k!} \langle\theta|s\rangle^k. \tag{16}$$

using the scalar product notation $\langle\theta|s\rangle = \sum_j \theta_j s_j$ also for vectors of matrices or operators s. The infinitesimal generators s form the corresponding Lie algebra. In particular, the group of d–dimensional translations is generated by the gradient operator ∇. This can be verified by recalling the multidimensional Taylor formula for expansion of h at x

$$S(\theta)h(x) = e^{\langle\theta|\nabla\rangle}h(x) = \sum_{k=0}^{\infty} \frac{\langle\theta|\nabla\rangle^k}{k!} h(x) = h(x+\theta). \tag{17}$$

Up to first order the expansion of the exponential function reads $S \approx 1 + \theta s$. Thus, we can define a distance to an infinitesimal symmetry by

$$d_s^2(h) = \left\langle \frac{h - (1+\theta s)h}{\theta} \left| K_s \right| \frac{h - (1+\theta s)h}{\theta} \right\rangle = \langle sh\,|K_s|\,sh\rangle, \tag{18}$$

with infinitesimal symmetry operator $K = s^\dagger K_s s$. For translations and K_s equal to the identity I this results exactly in (11).

In cases where samples t_α for h are available, like sometimes in image reconstruction or in times series prediction, a finite rank approximation $\hat{K}^{-1}$ of

$$K^{-1}(x, x') = \int dh\, p(h) h(x) h(x') \tag{19}$$

can be obtained by the empirical estimator [57]

$$\hat{K}^{-1}(x, x') = \sum_\alpha t_\alpha(x) t_\alpha(x'). \tag{20}$$

An optimal h^* in the subspace where $\hat{K}^{-1}$ is positive definite and therefore invertible can be found by singular value decomposition. Alternatively, more prior concepts with concepts operators K_i can be added so the sum $\sum_i K_i$ becomes strictly positive definite.

2.4. From posterior probability to likelihood energy

In this Section we discuss the interpretation of quadratic concepts and error functionals in terms of posterior probability and likelihoods (see also Appendix B). We will need this interpretation in the following to combine quadratic concepts to more complex error functionals.

Typically, hypotheses h are defined by specifying the probabilities $p(y_D|x_D,h)$ of finding data y_D in situations x_D under h. These data generating probabilities $p(y_D|x_D,h)$, or (x_D–conditional) likelihoods of h under y_D, are related to posterior probabilities $p(h|D)$ we are interested in according to Bayes' rule

$$p(h|D) = \frac{p(D|h)p(h)}{p(D)} = \frac{p(y_D|x_D,h)p(h)}{p(y_D|x_D)}, \tag{21}$$

with

$$p(y_D|x_D) = \int_{\mathcal{H}} dh\, p(y_D|x_D,h)p(h), \tag{22}$$

assuming $p(h|x_D) = p(h)$ and shorthand notation

$$\int_{\mathcal{H}} dh \cdots = \int_{\mathcal{H}} \prod_x dh(x) \cdots = \int_{\mathcal{H}_1} dh(x_1) \int_{\mathcal{H}_2} dh(x_2) \cdots = \left(\prod_x \int_{\mathcal{H}_x} dh(x) \right) \cdots . \tag{23}$$

The terms of Eq.(21) are in the Bayesian context often referred to as

$$\text{posterior} = \frac{\text{likelihood} * \text{prior}}{\text{evidence}}. \tag{24}$$

For complete data which means for uniform (possibly improper, i.e., non–normalizable) $p(h)$,[1] yields (see also Appendices B.1.1-B.4.1 identifying h with h according to Appendix B.6)

$$p(h|D) = \frac{p(y_D|x_D,h)}{\int dh\, p(y_D|x_D,h)} = \frac{p(y_D|x_D,h)}{Z_L} \propto p(y_D|x_D,h). \tag{25}$$

Being interested in the dependency on h for given data D we can skip the h–independent factor $p(D)$ or $\int dh\, p(D|h)$ and calculate instead of $p(h|D)$ the inverse $p(D|h)$. For the special case (8) of one quadratic concept with template t, so

[1] The probability $p(h)$ is known as prior probability of h. In this paper we treat prior information explicitly and analogous to standard training data. That means we assume $p(h)$ to be h–independent and collect all prior information in D_0. They enter the formalism through $p(h|D_0)$ or $p(D_0|h)$, respectively.

$y_D = t$ and x_D corresponds to K, we find for the denominator in (25) because of the translational invariance of the gaussian measure

$$Z_L = \int dh\, p(t|K,h) = \frac{\int dh\, e^{-d^2/2}}{\int dt\, e^{-d^2/2}} = \frac{\int dh\, e^{-\frac{1}{2}\left\langle h-t\middle|K\middle|h-t\right\rangle}}{\int dt\, e^{-\frac{1}{2}\left\langle h-t\middle|K\middle|h-t\right\rangle}} = 1, \tag{26}$$

and thus $p(h|D) = p(y_D|x_D, h)$.

To obtain error functionals we will now express probabilities in terms of energies and related quantities (see Appendix A.1.2). For example, the data generating probabilities or (x_D–conditional) likelihoods of h can be written in terms of (x_D–conditional) likelihood energies $E(y_D|x_D, h)$ and inverse likelihood temperature β_L

$$p(y_D|x_D, h) = \frac{e^{-\beta_L E(y_D|x_h)}}{Z(\mathcal{Y}_D|x_D, h)} = e^{-\beta_L \left(E(y_D|x_D,h) - F(\mathcal{Y}_D|x_D,h)\right)}, \tag{27}$$

and analogously the posterior probability $p(h|D)$ becomes in terms of posterior energy $E(h|D)$ and inverse posterior temperature β_P

$$p(h|D) = \frac{e^{-\beta_P E(h|D)}}{Z(\mathcal{H}|D)} = e^{-\beta_P \left(E(h|D) - F(\mathcal{H}|D)\right)} \tag{28}$$

or in terms of likelihoods

$$\begin{aligned} p(h|D) &= \frac{p(h)}{p(y_D|x_D)} p(y_D|x_D, h) = \frac{1}{Z_L(h,D)} \frac{e^{-\beta_L E(y_D|x_D,h)}}{Z(\mathcal{Y}_D|x_D, h)} \\ &= e^{-\beta_L \left(E(y_D|x_D,h) - F(\mathcal{Y}_D|x_D,h)\right) + c_L(h,D)}, \end{aligned} \tag{29}$$

with

$$Z_L(h,D) = \frac{p(y_D|x_D)}{p(h)}, \quad c_L(h,D) = \ln p(h) - \ln p(y_D|x_D), \tag{30}$$

free energies,

$$F(\mathcal{H}|D) = -\frac{1}{\beta_P} \ln Z(\mathcal{H}|D), \quad F(\mathcal{Y}_D|x_D, h) = -\frac{1}{\beta_L} \ln Z(\mathcal{Y}_D|x_D, h), \tag{31}$$

and normalization factors or partition sums

$$Z(\mathcal{H}|D) = \int_{\mathcal{H}} dh\, e^{-\beta_P E(h|D)}, \quad Z(\mathcal{Y}_D|x_D, h) = \int_{\mathcal{Y}_D} dy_D\;\, e^{-\beta_L E(y_D|x_D,h)}. \tag{32}$$

Thus, the posterior energy can be expressed by the likelihood energy

$$E(h|D) = F(\mathcal{H}|D) + \frac{\beta_L}{\beta_P}\Big(E(y_D|x_D, h) - F(\mathcal{Y}_D|x_D, h)\Big) + \frac{1}{\beta_P} c_L(h,D). \tag{33}$$

For complete data, i.e., h–independent $p(h)$

$$Z_L(h,D) = Z_L(D) = \int dh\, p(y_D|x_D,h) = \int dh\, \frac{e^{-\beta_L E(y_D|x_D,h)}}{Z(\mathcal{Y}_D|x_D,h)}. \tag{34}$$

In that case we have already seen that maximizing the likelihood $p(y_D|x_D,h)$ with respect to h is equivalent to maximizing the posterior $p(h|D)$. However, minimizing the posterior energy is not necessarily equivalent to minimizing the (conditional) likelihood energy. This is due to the possibility of a h–dependent normalization of the likelihood energy. If, however, in addition to complete data also the likelihood normalization is h–independent, i.e., $Z(\mathcal{Y}_D|x_D,h) = Z(\mathcal{Y}_D|x_D)$, then

$$Z_L = \frac{\int dh\, e^{-\beta_L (y_D|x_D,h)}}{Z(\mathcal{Y}_D|x_D)} = \frac{Z(y_D|x_D,\mathcal{H})}{Z(\mathcal{Y}_D|x_D)} \tag{35}$$

and the posterior becomes according to Eq.(29)

$$p(h|D) = \frac{e^{-\beta_L E(y_D|x_D,h)}}{Z(y_D|x_D,\mathcal{H})} \propto e^{-\beta_L E(y_D|x_D,h)}, \tag{36}$$

where $Z(y_D|x_D,\mathcal{H}) = \int dh\, e^{-\beta_L E(y_D|x_D,h)}$ is an integral over the conditional variable h. Thus, *for complete data and h–independent likelihood normalization maximizing the posterior is equivalent to minimizing the (x_D–conditional) likelihood energy.*

Remark 1 (Mixed representation by likelihood and posterior energies)
Up to here we expressed posterior energies completely by likelihood energies. It is often also useful, however, to express the posterior energy by a sum of likelihood energy and posterior energy terms. A prior term for example, may describe a h (and not data) generating process directly given by $p(h|D_0)$. Let the data $D = (x_D, y_D)$ be therefore divided in measured data $D_L = (x_L, y_L)$ which enter as likelihoods and a generating prior $D_P = (x_P, y_P)$ describing the posterior of the h–generating process. Thus, $D = D_L \cup D_P$ and $x_D = x_L \cup x_P$, $y_D = y_L \cup y_P$. Then, according to Bayes' theorem and $p(h|x_L, D_P) = p(h|D_P)$

$$p(h|D) = \frac{p(y_L|x_L,h)p(h|D_P)}{p(y_L|x_L,D_P)} = \frac{e^{-\beta_L E(y_L|x_L,h) - \beta_P E(h|D_P) + \ln p(y_L|x_L,D_P)}}{Z(Y_L|x_L,h)\, Z(\mathcal{H}|D_P)}. \tag{37}$$

In this case, minimizing the mixed energy $\beta_L E(y_L|x_L,h) + \beta_P E(h|D_P)$ is equivalent to maximizing the posterior $p(h|D)$ if the normalization of the likelihood terms $Z(Y_L|x_L,h)$ is h–independent.

3. Combinations of quadratic concepts

3.1. AND: Classical regularization and gaussian processes

In the classical situation of regularization theory the aim is to approximate all available data, training data D_T as well as prior information D_0. In logical terms, the aim is to approximate D_0 AND D_{T_1} AND D_{T_2} AND $\cdots D_{T_n}$. A logical AND of events corresponds to a product for probabilities $p(A,B) = P(A)p(B|A)$ or $p(A,B) = P(A)p(B)$ for independent events. A product for probabilities corresponds to a sum for log–probabilities. This holds also for concepts, if we interpret concepts as linear functions of log–probabilities, i.e., as energies (see Appendix A) or errors (see Appendix B.6). In the following we will identify the events A, B, etc. with data in the form of template functions t_j.

Consider a set of quadratic concepts d_j^2 with templates $T_N = \{t_j | 1 \le j \le N\}$ and concept operators $K_N = \{K_j | 1 \le j \le N\}$. Assume we have data $y_D = t_1$ AND t_2 AND $\cdots$, i.e.,

$$p(h|D) = p(h|T_N, K_N) \propto p(y_D|x_D, h) = p(T_N|K_N, h) \tag{38}$$

$$= \prod_j^N p(t_j|K_j, h) = \frac{e^{-\beta_L \sum_j^N E(t_j|K_j,h)}}{Z(\mathcal{T}^N|K_N, h)} = e^{-\beta_L E(T_N|K_N,h)}, \tag{39}$$

with $E(T_N|K,h) = E(y_D|x_D,h) = \sum_j E(t_j|K_j,h) = \sum_j^N d_j^2(h)/2$. In writing $p(t_j|K_j,h)$ we suppressed the dependency on the temperature $1/\beta_L$. The normalization factor $Z(\mathcal{T}^N|K_N,h)$ factorizes

$$Z(\mathcal{T}^N|K_N,h) = \prod_j^N Z(\mathcal{T}|K_j,h) \tag{40}$$

with

$$Z(\mathcal{T}|K_j,h) = \int \left(\prod_x dt_j(x)\right) e^{-\beta_L E(t_j|K_j,h)}. \tag{41}$$

For quadratic concepts the integrals are gaussian and therefore independent of the mean h, i.e.,

$$Z(\mathcal{T}^N|K_N,h) = Z(\mathcal{T}^N|K_N). \tag{42}$$

Thus, according to the results of the last Section we may minimize the likelihood energy $E(y_D|x_D,h) = E(T_N|K_N,h)$ instead of the posterior energy $E(h|D)$. Writing now for simplicity $E(T_N|K_N,h) = E(h)$, the following proposition is obtained by straightforward calculation:

Proposition 1 (Probabilistic AND) A sum of squared distances $E_j = d_j^2(h)/2 = d(t_j,h)/2$ can be written

$$E(h) = \sum_j^N E_j(h) = \frac{1}{2}\sum_j^N d_j^2(t_j,h) = \frac{1}{2}\sum_j^N \left\langle h-t_j \left| K_j \right| h-t_j \right\rangle$$

$$= \frac{1}{2}\left(\langle h\,|K|\,h\rangle + \sum_j^N \left\langle t_j \left| K_j \right| t_j \right\rangle \right) - \left\langle h \right| \sum_j K_j t_j \Big\rangle$$

$$= \frac{1}{2}\left(\langle h-t\,|K|\,h-t\rangle + \sum_j^N \left\langle t_j \left| K_j \right| t_j \right\rangle - \langle t\,|K|\,t\rangle \right)$$

$$= \frac{d^2(0,h)}{2} - \langle \tilde{t}|h\rangle + \frac{N}{2}M_2$$

$$= \frac{N}{2}\left(\bar{d}^2(t,h) + V\right) = \frac{d^2(t,h)}{2} + E_{min} \tag{43}$$

as sum of one squared distance $d^2(t,h)$ from template average t and a h–independent minimal energy E_{min}. Squared distances are defined as

$$d^2(0,h) = \langle h\,|K|\,h\rangle\,, \tag{44}$$

$$d^2(t,h) = \langle h-t\,|K|\,h-t\rangle\,, \tag{45}$$

$$\bar{d}^2(t,h) = \frac{d^2(t,h)}{N} = \langle h-t\,|\bar{K}|\,h-t\rangle\,, \tag{46}$$

with template average

$$t = K^{-1}\tilde{t}, \quad \tilde{t} = \sum_j^N K_j t_j, \tag{47}$$

concept operators

$$K = \sum_j K_j, \quad \bar{K} = \frac{K}{N} = \frac{1}{N}\sum_j K_j, \tag{48}$$

and h–independent minimal energy and template variance V

$$E_{min}(T_N) = \frac{N}{2}V, \quad V = M_2 - M_1^2, \tag{49}$$

with

$$M_2 = \frac{1}{N}\sum_j^N \left\langle t_j \left| K_j \right| t_j \right\rangle, \quad M_1^2 = \langle t\,|\bar{K}|\,t\rangle\,. \tag{50}$$

The linear stationarity equation for a functional $E = d^2(t,h)/2 + E_{min}$ reads

$$0 = K(t-h) = \tilde{t} - Kh. \tag{51}$$

For positive definite, i.e., invertible K, this has solution $h = t = K^{-1}\tilde{t}$.

We see that for quadratic concepts addition does not lead to something really new: The sum of quadratic functions is a quadratic function, or in a probabilistic interpretation, a product of gaussians is a gaussian. *It is also interesting to note that for such additive combinations all template functions t_i corresponding to infinite data can be eliminated from the formalism by combining them in one term with template average t and solving for a shifted $h' = h - t$. This elimination of 'infinite data' templates will not be possible in the nonlinear combinations presented in the next sections.*

Remark 2 (Normalization) The normalization factor $Z(\mathcal{T}^N|K_N,h)$ to obtain $p(T_N|K_N,h)$ being a product of q–dimensional gaussian integrals can be calculated explicitly (see Eq.(314))

$$Z(\mathcal{T}^N|K_N,h) = \prod_j^N \int \left(\prod_x^q dt_j(x)\right) e^{-\frac{\beta_L}{2} d_j^2(t_j,h)} \tag{52}$$

$$= \prod_j^N \left((2\pi)^{\frac{q}{2}} \beta_L{}^{-\frac{q}{2}} \left(\det K_j\right)^{-\frac{1}{2}}\right) = (2\pi)^{\frac{qN}{2}} \beta_L{}^{-\frac{qN}{2}} \left(\det \prod_j^N K_j\right)^{-\frac{1}{2}}. \tag{53}$$

For gaussian integrals the exact solution (53) coincides with a saddle point approximation (see B.5). Eq.(53) differs from a T–dependent normalization over $h \in \mathcal{H}$ which would be necessary to obtain $p(h|D)$ if we would interpret a sum of quadratic concepts as posterior energy $E(h|D) = \sum_j E(h|t_j,K_j) = \sum_j d_j^2/2$,

$$Z(\mathcal{H}|D) = \int dh\, e^{-\beta_P E(h)} = \int \left(\prod_x^q dh(x)\right) e^{-\frac{\beta_P}{2}\sum_j^N d_j^2(t_j,h)}$$

$$= e^{-\beta_P E_{min}} \int dh\, e^{-\beta_P \frac{d^2(0,h)}{2}} = e^{-\beta_P E_{min}} (2\pi)^{\frac{q}{2}} \beta_P{}^{-\frac{q}{2}} \left(\det \sum_j^N K_j\right)^{-\frac{1}{2}}. \tag{54}$$

Denoting by $\langle \cdots \rangle_K$ a gaussian average over h with covariance $\propto K^{-1} = (\sum_j K_j)^{-1}$ we recognize the moment generating function for $h(x)$

$$M(\beta_P \tilde{t}) = e^{\beta_P E_{min}} Z(\mathcal{H}|D) = \left\langle e^{\beta_P \langle\, \tilde{t}\,|\,h\,\rangle} \right\rangle_K$$

$$= \sum_{n=0}^{\infty} \frac{\beta_P^n}{n!} \int dx_1 \cdots dx_n \, \tilde{t}(x_1) \cdots \tilde{t}(x_n) \left\langle h(x_1) \cdots h(x_n) \right\rangle_K . \tag{55}$$

Hence (functional) derivatives of $M(\beta_P \tilde{t})$ with respect to $\beta_P \tilde{t}(x)$ generate moments of $h(x)$ (see Section A.3.3, in particular Wick's theorem 182).

Remark 3 (Kernel methods) Practically important is the case where the stationarity equation can be solved in a space with dimension $\tilde{n} \le n$ being the number of different x values in the training data. This can be much less then the space necessary for a reasonable discretization of the whole function h. To see this we consider the classical situation of mean–square data terms and one additional prior concept

$$E(h) = \langle h - t_T \, | K_T | \, h - t_T \rangle + \langle h - t_0 \, | K_0 | \, h - t_0 \rangle \tag{56}$$

with

$$\langle h - t_T \, | K_T | \, h - t_T \rangle + V = \sum_j^n \left\langle h - t_j \, \middle| K_j \middle| \, h - t_j \right\rangle = \sum_j^n (y_j - h(x_j))^2 . \tag{57}$$

Thus, the operator $K_T = \sum_j K_j$ is diagonal, commutes with the projector P_T into the space $\mathcal{X}_T$ of training data and has matrix elements

$$K_T(x, x') = \delta(x - x') \sum_j^n \delta(x - x_j) = \delta(x - x') \, \delta(0) n_x, \quad n_x = \sum_j^n \frac{\delta(x - x_j)}{\delta(0)}, \tag{58}$$

containing, for discrete x where $\delta(0) = \delta_{0,0} = 1$ represents a Kronecker–δ, the number n_x of data for every x. For continuous x the factor $\delta(0)$ becomes infinite but will cancel in the following calculations. The data template

$$t_T = K_T^{-1} \sum_j K_j t_j, \quad t_T(x) = \frac{\sum_j^n \delta(x - x_j) y_j(x_j)}{\sum_j^n \delta(x - x_j)} = \frac{1}{n_x} \sum_j^{n_x} y_j(x_j) \tag{59}$$

is the average of y values for given x. Here K_T^{-1} is defined on $\mathcal{X}_T$ and $t_T(x) = 0$ for x not in the training data, i.e, with $n_x = 0$. The stationarity equation

$$0 = K_T(h - t_T) + K_0(h - t_0)$$

yields

$$h = K_0^{-1} a + t_0, \tag{60}$$

with $a = K_T(t_T - h)$. For known K_0^{-1} it is sufficient to solve

$$a = [I + K_T K_0^{-1}]^{-1} [K_T(t_T - t_0)] \tag{61}$$

for a function defined on the space $\mathcal{X}_T$ with dimension $\mathrm{Tr} P_T = \tilde{n} \le n$. The formulas can be adapted to the case where K_0 has zero eigenvalues so it is not invertible over the whole space [70]. Classical examples of such *kernel methods*

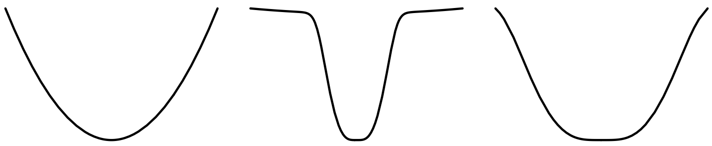

Figure 1. Gaussian vs. two 'robust' gaussian mixtures: Left: Logarithm of one gaussian (parabola), Middle: Logarithm of mixture with two gaussians with different variance and equal mean (insensitive for large deviations), Right: Logarithm of mixture with three gaussians with different variances and different means.

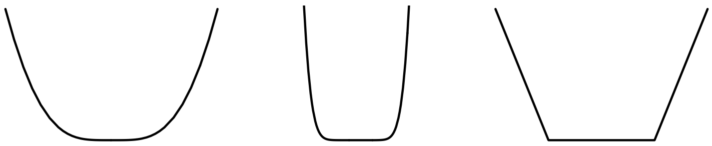

Figure 2. Three robust error functions which are insensitive to small errors. Left: Logarithm of mixture with two gaussians with equal variance and different means, Middle: Logarithm of mixture with 11 gaussians with equal variance and different means, Right: ϵ–insensitive error.

are gaussian Radial Basis Functions which use as concept operator the pseudo–differential operator

$$K_0 = \sum_{m=0}^{\infty} (-1)^m \frac{\sigma^{2m}}{m!\, 2^m} \Delta^m, \tag{62}$$

where Δ^m is the m–iterated laplacian and which has a gaussian inverse. They also include piecewise linear interpolation ($K_0 = -\Delta$) and various spline methods (e.g., $K_0 = \Delta^2$) [70,56,23].

Remark 4 (Robust error functions and support vector machine) As a generalization of quadratic concepts one can allow nonquadratic functions U of filtered differences $W(h-t)$ [76]. In the simplest case U is only applied to the data concept (57) of a functional of the classical form (56). Then the nonlinear stationarity equation can also be restricted to the $\tilde{n}$–dimensional space $\mathcal{X}_T$. The data term can for example be replaced by a gaussian mixture model (see Section 3.2). Robust error functions have flat regions and are therefore insensitive to, i.e., robust against, changes in that flat region [28,29]. Also numerically flat regions in the error surface can be useful because there the gradient vanishes and those regions do not contribute to the stationarity equations. Robust error functions can for example down–weight large errors. Typical cases are filters used in image processing where edges represent large discontinuities (regions with low smoothness) but are relatively likely. Fig.1 shows examples obtained by gaussian mixtures. Another variant of robust error functions is insensitive to small errors. An example is an ϵ–insensitive error (zero between $\pm\epsilon$, linear outside, see Fig.2) for the data term. For example, expanding h in a basis of eigenfunctions Φ_k of the prior concept operator

$$K_0 = \sum_k \lambda_k \Phi_k \Phi_k^T, \quad h(x) = \sum_k n_k \Phi_k(x) \tag{63}$$

yields for functional (56)

$$E(h) = \sum_i \left(\sum_k n_k \Phi_k(x_i) - y_i \right)^2 + \sum_k \lambda_k |n_k|^2. \tag{64}$$

Replacing the mean–square error data term by an ϵ–insensitive error results in a standard quadratic programming problem and is equivalent to Vapnik's support vector machine [69,24,64,65].

3.2. OR: Mixture models

Assume we believe that a function can be similar to either one of two templates. Often a list of possible alternatives can be given. For example, one may simply state that a face is that of a women or that of a man, eyes may usually be open or closed, or we may be able to give a list of possible pattern prototypes for a time series. For example, electrocardiograms, or similarly earthquakes, have typical patterns which can appear in distinguishable variants. Thus, we discuss here the case where we want to approximate t_1 OR t_2 OR $\cdots t_n$. For probabilities we have $p(A \,\mathrm{OR}\, B) = P(A) + p(B) - P(A,B)$ where the last term vanishes for exclusive events. For log–probabilities L of exclusive events this means

$$L(A \,\mathrm{OR}\, B) = \ln\left(e^{L(A)} + e^{L(B)}\right). \tag{65}$$

Relating now concepts to log–probabilities by interpreting them as energies or errors, respectively, shows that alternative quadratic concepts lead to gaussian mixture models.

To be specific, let us discuss the cases of discrete input, output and generation noise (See Section B.2.4). *Input noise* means that the measurement device producing outcome y is not known exactly, i.e.,

$$p(y|x,h) = \sum_i p(y|x_i,h)\, p(x_i|x). \tag{66}$$

For gaussian $p(y|x_i,h)$ hypothesis h defines the mean and the independent variable x_i defines the covariance K_i^{-1}/β_L of the measurement producing y. Thus, model (66) leads to a mixture of gaussians with different covariances. Similarly, under *output noise* a measured value cannot be read off exactly,

$$p(y|x,h) = \sum_i p(y|y_i)\, p(y_i|x,h). \tag{67}$$

Thus having found y the true measurement result was one of the y_i. For gaussian $p(y_i|x,h)$ where y_i is represented by a template function t_i this leads to a likelihood for h being a mixture of gaussians with different means. In particular, consider a situation of ambiguous data where a set of y_i cannot be distinguished by a measurement procedure so all y_i lead to outcome y. In that case

$$p(y|y_i) = \frac{1}{N_i}\sum_i^{N_i} \delta(y - y_i), \tag{68}$$

so the likelihood becomes a simple sum over alternatives

$$p(y|x,h) = \frac{1}{N_i}\sum_i^{N_i} p(y_i|x,h). \tag{69}$$

Generation noise on the other hand means that the h–producing probability is modeled as a mixture

$$p(h) = \sum_i p(i)p(h|i), \tag{70}$$

where i could determine mean and covariance of a gaussian $p(h|i)$.

For a maximum posterior approximation we have to maximize the posterior density

$$p(h|D) = p(h|D_L, D_P) = \frac{p(y_L|x_L,h)p(h|D_P)}{p(y_L|x_L,D_P)}. \tag{71}$$

Including input, output, and generation noise the posterior becomes

$$p(h|D) \propto \sum_{ijk} p(k)p(h|k)p(y|i)p(j|x)p(i|j,h), \tag{72}$$

skipping the h–independent factor $p(y_L|x_L,D_P)$. We will especially consider the case where $p(i|j,h) = \prod_l^{N_L} p(i_l|j_l,h)$ factorizes into N_L independent components

$$p(h|D) \propto \sum_{ijk} p(k)p(h|k)p(y|i)p(j|x)\prod_l^{N_L} p(i_l|j_l,h), \tag{73}$$

with $i = \{i_l|0 \le l \le N_L\}$ and $j = \{j_l|0 \le l \le N_L\}$. In particular, if also input and output noise factorize, i.e., $p(y|i)p(j|x) = \prod_l p(y_l|i_l)p(j_l|x_l)$, this would read

$$\begin{aligned} p(h|D) &\propto \sum_{l'}^{N_L}\sum_{i_{l'}j_{l'}k} p(k)p(h|k)\prod_l^{N_L} p(y_l|i_l)p(i_l|j_l,h)p(j_l|x_l) \\ &= \sum_k p(k)p(h|k)\prod_l^{N_L}\sum_{i_lj_l} p(y_l|i_l)p(i_l|j_l,h)p(j_l|x_l). \end{aligned} \tag{74}$$

Choosing now quadratic concepts for h–dependent (likelihood and generation) energies, Eq.(73) becomes

$$p(h|D) \propto \sum_{ijkm} p(k,m)p(h|K_k,t_m)p(T|i)p(j|K)\prod_l p(t_{i_l}|K_{j_l},h), \tag{75}$$

where $T = \{t_l|0 \le l \le N_L\}$, $K = \{K_l|0 \le l \le N_L\}$, and separate summation variables for mean and covariance of the h–generating process have been introduced.

To find an error functional to be minimized we express the posterior in terms of energies and free energies, the latter determined by normalization factors. Two kinds of free energies have to be included in an error functional.

1. Free energies cannot be skipped if depending on variables with respect to those we have to maximize the posterior. Thus, in general free energies of all likelihood terms have to be included. For the special case of gaussians, however, normalization of likelihood terms is h–independent.
2. Furthermore, it is often easier and more meaningful to specify instead of a joint probability, e.g., $p(h,k)$, conditional and marginal probability, e.g., $p(k)$ and $p(h|k)$. In that case also free energies depending on summation variables i, j, k, m have to be included. Otherwise, such free energies would contribute to marginal energies like $p(k)$, and terms like $E(k)$ could not be interpreted as energy for marginal $p(k)$. Systems specified by conditional energies are also known as *disordered systems* (see Section A.4).

Thus, in terms of energies,

$$p(h|D) \propto \sum_{ijk} e^{-\beta_h E(k)} e^{-\beta_P (E(h|k)-F(\mathcal{H}|k))} e^{-\beta_y (E(y|i)-F(\mathcal{Y}|i))}$$
$$\times e^{-\beta_L \sum_l (E(i_l|j_l,h)-F(\mathcal{I}_l|j_l,h))} e^{-\beta_x (E(j|x)-F(\mathcal{J}|x))}, \tag{76}$$

for $i_l \in \mathcal{I}_l$, $j \in \mathcal{J}$, $y \in \mathcal{Y}$, $h \in \mathcal{H}$. In particular for quadratic concepts this becomes, choosing $\beta_P = \beta_L = \beta$,

$$p(h|D) \propto \sum_{ijkm} p(k,m)p_{ij} e^{-\beta\left((E(h|K_k,t_m)-F(\mathcal{H}|K_k,t_m))+\sum_l\left(E(t_{i_l}|K_{j_l},h)-F(\mathcal{T}_{i_l}|K_{j_l},h)\right)\right)}, \tag{77}$$

with $p_{ij} \propto p(T|i)p(j|K)$. Notice that despite $p(T|i)$ is not normalized over i we could nevertheless choose p_{ij} to be normalized over i and j by taking $p_{ij} = p(T|i)p(j|K)/\sum_i p(T|i)$. For k– and j–independent covariances this becomes

$$p(h|D) \propto \sum_{im} p(t_m)\, p_i\, e^{-\beta\left(E(h|K_P,t_m)+\sum_l E(t_{i_l}|K_l,h)\right)}, \tag{78}$$

with $p_i \propto \prod_l p(T|i)$. The posterior can be written completely in likelihood form by using that for quadratic concepts the free energy is h–independent $F(\mathcal{H}|K_k,t_m) = F(\mathcal{T}_m|K_k,h)$ and $p(h|K_P,t) = p(t|K_P,h)$ under uniform prior. The h–generating energy $E(h|K_P, t_m)$ can therefore be included in the sum over l. Hence, combining i, j, k, m into one multi–index i and writing $t_{i_l} = t_{il}$, $K_{i_l} = K_{il}$ Eq.(77) reads

$$p(h|D) \propto \sum_i p_i \, e^{-\beta \sum_l^{N_L+1} \left(E(t_{il}|K_{jl},h) - F(\mathcal{T}_{il}|K_{jl},h) \right)}. \tag{79}$$

Hence, we obtain:

Theorem 1 (Mixture model) Alternative quadratic concepts can be implemented by the mixture model

$$E_M(h) = -\ln p(h|D) = -\ln \left(\sum_i^N p_i \, e^{-\beta E_i(h) + c_i} \right), \tag{80}$$

where the component energies

$$E_i(h) = \sum_j^{N_i} E(t_{ij}|K_{ij}, h) = \sum_j^{N_i} E_{ij}(h) \tag{81}$$

are additive combinations of quadratic concepts

$$E_{ij}(h) = \frac{d_{ij}^2(h)}{2}, \quad d_{ij}^2(h) = \left\langle h - t_{ij} \,\middle|\, K_{ij} \,\middle|\, h - t_{ij} \right\rangle. \tag{82}$$

If i–dependent the normalization integrals

$$c_i = -\sum_j^{N_i} \ln \left(\int dt_{ij} e^{-\beta \, E_{ij}} \right) + c, \tag{83}$$

have to be calculated up to an arbitrary constant c so they do not interfere with mixture probabilities p_i. The model has the stationarity condition[2]

$$0 = t_M(h) - K_M(h) h, \tag{84}$$

with

$$K_M = \sum_i a_i(h) K_i, \quad K_i = \sum_j K_{ij}, \quad a_i(h) = p_i \, e^{-\beta E_i(h) + c_i} \tag{85}$$

and

$$t_M = \sum_i a_i(h) K_i t_i = \sum_i a_i(h) \tilde{t}_i, \quad t_i = K_i^{-1} \tilde{t}_i, \quad \tilde{t}_i = \sum_j K_{ij} t_{ij}. \tag{86}$$

[2] For h–dependent $c_i = c_i(h)$ additional terms arising from $\delta c_i / \delta h$ contribute to the stationary equations.

Proof: The form of E_M follows directly from Eq.(79). The stationary equation is obtained by skipping the strictly monotone logarithm and using

$$\frac{\delta}{\delta h}\, d_{ij}^2 = 2K_{ij}(h - t_{ij}), \tag{87}$$

where the left hand side denotes the function with components $(\delta/\delta h(x))d_{ij}^2$. *q.e.d.*

It is in general nonlinear and can have multiple solutions. Indeed, the model (80) is in contrast to classical regularization functionals in general non–convex. The mixture model energy E_M has with respect to the summation variable i the form of a free energy for a system at finite temperature. Notice, also, the difference to mixture models like they are used frequently in density estimation. In such approaches h is assumed as gaussian mixture. In contrast, model (80) represents a mixture model for the posterior density and does not restrict h to be a gaussian mixture.

Remark 5 (Separate saddle point approximation for each component) Alternatively, a maximum posterior approximation (which would be exact for gaussian integrals) can also be applied to the i components separately. Then, however, in general a second minimization step has to be performed also for approximation problems (see Appendix B.6). Notice that also for separate saddle point approximations weighting factors $(\det K_i)^{-1/2}$ have to be calculated in case of unequal component covariances (see [14,12,51] and Appendix B.5.1).

Remark 6 (Low and high temperature limits) In the interpretation of error (or energy) minimization as saddle point approximation of a Bayesian risk (see Appendix B) the result becomes exact for zero temperatures, provided that only one dominating stationary point survives at zero temperature (see Appendix B.5). Thus we expect error minimization to be good at low enough temperature, i.e., large β. At low temperatures, however, the stationarity equations generally have many solutions, making it difficult to find the dominating one. (For a more detailed discussion of stable low temperature solutions for the mixture model see [44]). In practice one often uses annealing methods which start solving the stationarity equations at high temperature and iteratively adapt the found solution to lower and lower temperatures (see Appendix A.3.2). Interestingly, a saddle point approximation is for gaussian functions also exact at arbitrary temperatures $1/\beta$. Figures 17 and 18 in Appendix B show that for high temperatures a sum of two gaussians with different centers becomes approximately a gaussian again. Thus, in that case a saddle point approximation is also a good approximation at high temperatures. More generally, one may perform a (moment or high temperature) expansion of the exponential in powers of β,

$$\sum_i p_i\, e^{-\beta E_i} = \left\langle e^{-\beta E_i} \right\rangle_{\mathcal{I}} = \left\langle 1 - \beta E_i + \frac{\beta^2}{2} E_i^2 + \cdots \right\rangle_{\mathcal{I}} . \tag{88}$$

Thus, at high enough temperature minimizing $E_M = -c_i - \ln \left\langle e^{-\beta E_i} \right\rangle_{\mathcal{I}}$ with i–independent c_i becomes minimizing $\langle E_i \rangle_{\mathcal{I}} = \sum_i p_i\, E_i$. This is just the AND–case of Proposition 1. At medium temperatures larger differences to a full Bayesian approach have to be expected. (Especially at 'phase transitions' where solutions of the stationarity equations vary strongly with temperature.)

3.3. An example with ambiguous prior

Consider the following situation with ambiguous prior: Assume we want to implement that a function can be similar to prototype t_a OR another prototype t_b. Thus, we take the two prototypes as template functions $t_a(x)$ and $t_b(x)$. Assume further, we choose the same concept distances $d_a(h,t_a)$, $d_b(h,t_b)$ by taking for both templates the same concept operator $K_{a,0} = K_{b,0} = K_0$. In the following we consider a smoothness related sum of iterated laplacians

$$K_0 = \sum_l \lambda_l (-\Delta)^l , \tag{89}$$

where we understand $(-\Delta)^0 = I$. For equal mixture probabilities $p(1) = p(2)$ we obtain for (80)

$$p^M = p_1^M + p_2^M \propto e^{-E_M} = e^{-\beta E_1} + e^{-\beta E_2} = e^{-\frac{\beta}{2}(d_D^2 + d_a^2)} + e^{-\frac{\beta}{2}(d_D^2 + d_b^2)}, \tag{90}$$

with mean–square data concept $d_D^2 = \langle h - t_T \left| K_T \right| h - t_T \rangle$ as in Eq.(57), $d_a^2 = \langle h - t_a \left| K_0 \right| h - t_a \rangle$ and analogously for b. The stationarity condition for Eq.(90) is

$$h = p_1^M t_1 + p_2^M t_2 , \tag{91}$$

with component template averages

$$t_1 = (K_T + K_0)^{-1} \left(K_T t_T + K_0 t_a \right), \quad t_2 = (K_T + K_0)^{-1} \left(K_T t_T + K_0 t_b \right). \tag{92}$$

Because p_1^M and p_2^M are not functions but only two temperature dependent convex mixing coefficients the solutions for arbitrary temperatures have to be on the one dimensional line spanned by the two single components solutions t_1 and t_2. The stationarity condition (91) can be rewritten

$$h = t + \tanh \left(\frac{\beta}{2} (E_2 - E_1) \right) \frac{t_1 - t_2}{2}, \tag{93}$$

with total template average

$$t = (K_T + 2K_0)^{-1} \left(K_T t_T + K_0 (t_a + t_b) \right). \tag{94}$$

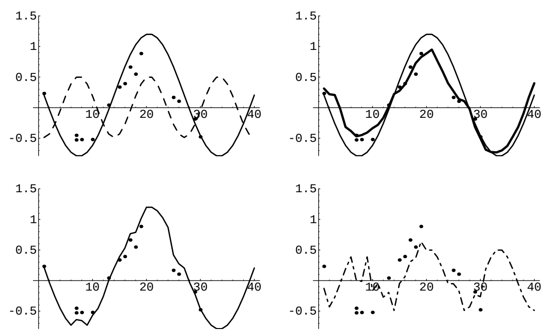

Figure 3. Example with ambiguous priors. The upper left diagram shows the two templates t_a and t_b and data D_T drawn from the interval $[0, 30]$. The upper right diagram shows the true state of Nature h_N (thick line) in comparison with t_a. The second row shows two solutions t_1 and t_2 for vanishing smoothness coefficients $\lambda_1 = 1$, $\lambda_i = 0$, $i > 1$.

In the one–dimensional space of solutions spanned by the two component averages, the equation is equivalent to that of the celebrated ferromagnet. Thus, the two template mixture model shows the typical ferromagnetic bifurcation (related to a 'phase transition' of the underlying physical system) switching with decreasing temperature from one to two stable solutions.

Fig.4 shows a numerical example of the two–template situation with

$$t_a(x) = -\sin\left(\frac{3\pi(x-1)}{m-1}\right) + a_1, \quad t_b(x) = \sin^2\left(\frac{3\pi(x-1)}{m-1}\right) + a_2, \tag{95}$$

with $m = 40$ and the constants a_i adjusted so the mean over the interval $[0, 40]$ becomes zero (see Fig.3). The shown results have been obtained by the EM (expectation–maximization) algorithm (see Section 4).

Fig.5 summarizes the temperature dependence of the model. While the two *low temperature limits* $\beta \to \infty$ are given by the single component solutions t_1 and t_2, the *high temperature limit* $\beta \to 0$ is the total template average t. All three solutions t_1, t_2 and t correspond to a quadratic minimization problem (or a gaussian process) with therefore linear stationarity equation.

The fact that the solutions for arbitrary temperatures are in the convex hull spanned by the low temperature solutions t_i is easily generalized to more then two alternative templates with equal covariances. This is a numerical important observation, because the low (and high) temperature limits are for quadratic concepts determined by linear equations. As explained in Section 3.1 those linear equations can be solved in a $\tilde{n} \leq n$ dimensional space given by the training data D_T. If the number of components is small the optimal mixture of low temperature solutions t_i may then for example be obtained by cross–validation

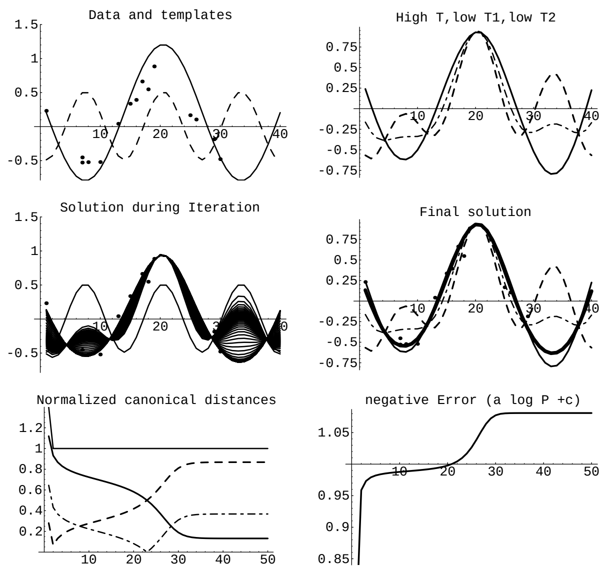

Figure 4. Mixture model (90) near a bifurcation. The chosen templates are $t_a(x) = -\sin\left(\frac{3\pi(x-1)}{m-1}\right) + a_1$ and $t_b(x) = \sin^2\left(\frac{3\pi(x-1)}{m-1}\right) + a_2$ with $m = 40$ and the constants a_i adjusted to obtain zero mean over the interval $[0, 40]$. 15 data points have been sampled on the interval $[0, 30]$ from a function similar to t_a. The value $\beta = 0.105$ is near the critical value shortly before the second solution occurs. The parameters for K_0 of Eq.(89) are $\lambda_0 = 0.1$, $\lambda_1 = \alpha^2\lambda_0$, $\lambda_2 = \alpha^4\lambda_0$, $\lambda_i = 0$, $i > 2$, with $\alpha = (m-1)/(3\pi)$. Row 1. Left: Shown are the templates t_a, t_b, (dashed) and the training data (dots). Right: The solution in the high temperature limit t (template average of data with t_a and t_b) $\beta \to 0$ (dashes with dots) and the two low temperature solutions t_1, t_2 which average the data with one of the templates t_a or t_b, respectively. Row 2. Left: The evolving solution during iteration with $A = K_M$ and relaxation factor $\eta = 1$ (see Section 4) starting from 'wrong' initial guess $h(x) = t_b(x)$. Right: The final solution compared with the high temperature and the two low temperature solutions. Row 3. Left: Shown are the squared distances of the solution h to the low and high temperature solutions t_1, t_2, t during iteration: $n^2_{t_i}(h) = \langle h - t_i\,|K_T + K_0|\,h - t_i\rangle / d^2(t_1, t_2)$, for $t_i = t_1$(dashes), $t_i = t_2$(thick) and $t_i = t$(dashes with dots)) . After one iteration step the sum $n^2_{t_1}(h) + n^2_{t_2}(h) = 1$ (thin) meaning that the solution moves along the one dimensional line connecting the two low temperature limits t_1 and t_2. Right: The negative error $-E_M(h)$ during iteration.

or similar techniques. Then it is not necessary to discretize the whole function h which is costly or even practically impossible for higher dimensional x. This makes such calculations feasible also for higher dimensional $\mathcal{X}_R$ spaces.

3.4. OR: Landau–Ginzburg regularization and interacting systems

The interpretation of quadratic concepts as energies is not the only possibility. Their quantitative relation to probabilities may be different. For example, combination of concepts may also be modeled by fuzzy logical operations. Those exist in many variations but coincide on their boundaries with Boolean opera-

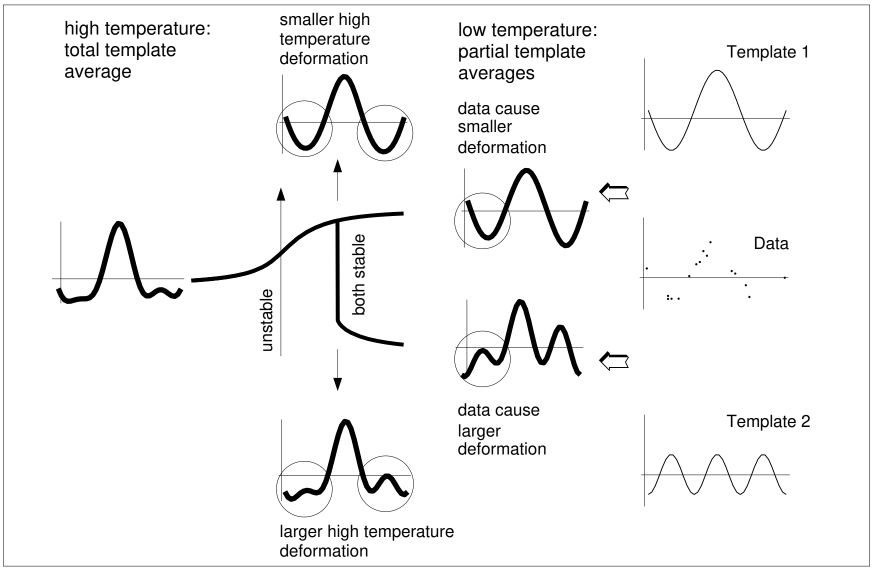

Figure 5. Scheme of temperature (β^{-1}) dependence of solutions for the error functional E^M of model (90). The two prior templates t_i (template 1 = t_a, template 2 = t_b) and the training data D_T (dots) are shown on the right hand side. At low temperature two solutions exist, being the template average of the data D_T with either one of the templates t_1 or t_2. Going to higher temperatures the less well fitting solution disappears at a critical temperature $1/\beta^*$. The better fitting solution survives and transforms for higher temperatures slowly into the high temperature solution which is the template average of all three templates: data D_T, t_1 and t_2.

tions [37,38]. We choose for concept distances $d(h,t)$ the logical interpretation of $d^2(h,t) = 0$, i.e., $h = t$, as 'true' and of $d^2(h,t) = \infty$ as 'false'. For such variables a typical implementation of a logical OR is a product

$$d^2_{(A\,\mathrm{OR}\,B)} = d^2_A d^2_B. \tag{96}$$

A product implementation for alternative concepts is especially convenient because then arbitrary combinations of quadratic concepts by (an additive) AND and (a multiplicative) OR are polynomial expressions in the concept distances. The stationarity equations of such polynomial models are easily obtained by setting the functional derivatives with respect to h to zero and given in the following theorem:

Theorem 2 (Landau–Ginzburg regularization) The polynomial model

$$E_{LG} = \frac{1}{2\beta_{LG}} \sum_{i=1}^{N} \prod_{j=1}^{N_i} \left(\beta_{LG}\, d^2_{ij}\right) = \frac{1}{2} \sum_{i=1}^{N} \beta_{LG}^{N_i-1} \prod_{j=1}^{N_i} d^2_{ij}, \tag{97}$$

has stationarity equation

$$K_{LG}(h)h = t_{LG}(h). \tag{98}$$

Here

$$K_{LG}(h) = \sum_i^N \sum_j^{N_i} M_{ij}(h) K_{ij} \tag{99}$$

and

$$t_{LG}(h) = \sum_i^N \sum_j^{N_i} M_{ij}(h) K_{ij} t_{ij}, \tag{100}$$

with

$$M_{ij}(h) = \beta_{LG}^{N_i-1} \prod_{k\neq j}^{N_i} d_{ik}^2(h), \tag{101}$$

and

$$d_{ij}^2(h) = \left\langle h - t_{ij} \,\middle|K_{ij}\middle|\, h - t_{ij} \right\rangle. \tag{102}$$

Proof: The stationary equation follows from (87) using the product rule for derivatives.

$$\frac{\delta}{\delta h} \prod_j d_{ij}^2 = \sum_j \left(\frac{\delta}{\delta h}\, d_{ij}^2 \right) \prod_{k\neq j} d_{ik}^2. \tag{103}$$

q.e.d.

In the high temperature limit $\beta_{LG} \to 0$ only quadratic terms survive and the stationarity equation becomes linear. The polynomial model E_{LG} resembles the phenomenological Landau–Ginzburg treatment of phase transitions in physics [39,10].

Remark 7 (Mixed likelihood and posterior interpretation of error/energy) A product $\prod_j d_j^2$ of quadratic concepts or log–probabilities does not implement a probabilistic OR for exclusive events i with gaussian probability according to d_i^2. But like for E_M of the mixture model, one may also try to interpret additive parts E_i of $E_{LG} = \sum_i E_i/(2\beta_{LG})$ as probabilistic AND for independent events. Hereby only terms E_i depending on only one template $\{t_{ij}\} = \{t_i\}$ can be interpreted as likelihood energies $E_i(y_i|x_i,h)$. Indeed, for an error E_{LG} of form(97), being composed out of squared distances $d_{ij}^2(h,t_{ij})$, the normalization $Z(\mathcal{Y}_i|x_i,h) = \int dy_i\, e^{-\beta_L E_i(y_i|x_i,h)}$ for additive parts E_i depending on only one t_i is h– and t_i–independent. This can be seen using $d^2(h,t) = d^2(h-t) = d^2(t-h) = d^2(t,h)$, so that also arbitrary functions $g(d^2)$ of squared distances fulfil $g(h,t) = g(d_{ij}^2(h,t)) = g(h-t) = g(t-h)$. For such functions $\int dh\, g(h-t) =$

$\int dt\, g(t-h) = \int dt\, g(h-t)$ for unrestricted or periodic domain of integration. Thus, an integral $Z(\mathcal{Y}_D|x_D, h)$ depending on only one t is both, h– and t–independent. For terms E_i, however, depending on different t_{ij}, like for example a product $d_1^1 d_2^2$, this is not true in general. Within a probabilistic interpretation such terms must either be interpreted as posterior energy, or the h (and possibly i)–dependent logarithm of the normalization constant $\ln Z(\mathcal{Y}_D|x_D, h)$ has to be subtracted.

Example 5 (The two concept model again) The two–template example of Section 3.3 may alternatively be parameterized as follows:

$$E_{LG} = d_D^2 + \beta_{LG} d_a^2 d_b^2. \tag{104}$$

The stationarity equation is cubic with either one or two stable solutions depending on β_{LG}. Variations are

$$E_{LG} = (d_D^2 + d_a^2)(d_D^2 + d_b^2), \tag{105}$$

or

$$E_{LG} = d_D^2 + d_a^2 + d_b^2 + \beta_{LG}\left(d_D^2 + d_a^2 d_b^2\right). \tag{106}$$

In the latter formulation β_{LG} interpolates between OR at low temperatures and AND at high temperatures, similar to the situation for mixture models. For numerical results of the Landau–Ginzburg model for the example of Section 3.3 see [44].

3.5. OR: Combination of methods and continuous transformations

The number of components i of a mixture model $\sum_i p_i e^{-\beta E_i + c_i}$ can be too large to be treated exactly. Consider a set of template functions $t(\theta)$, with function values denoted $t(x,\theta)$, parameterized by a continuous parameter (vector) $\theta \in \Theta$. This results in a continuous mixture model $\int d\theta\, p(\theta) e^{-\beta E(\theta) + c(\theta)}$ where the sum $\sum_i$ is replaced by an integral $\int d\theta$. Thus, analogous to the Bayesian integral over h (or h, see Appendix B), the θ–integral has to be solved by an approximation. This approximation can be of low temperature type (restricting to the most important contributions, e.g. saddle point approximation) or high temperature type (starting from the mean, e.g., cumulant or moment expansion relying on an expansion of the exponential) or a Monte Carlo integration (random sampling). From the fact that the norm of a gaussian does not depend on its mean it follows that changing templates does not change the normalization constant as long as the covariance K^{-1} is unaltered. Hence, in cases with θ–independent K also the partition sum $Z(\theta)$ and thus $c(\theta)$ is θ–independent. For θ–dependent K the replica method or supersymmetric approaches (see Appendix A.4) can be useful in some cases.

Two important situations have to be mentioned where continuously parameterized or *adaptive templates* appear naturally.

1. (Approximate structural models or combining and extending arbitrary learning methods) We have already discussed in 2.2 that a template can be given by an arbitrary parameterized function, for example regression models like decision trees or neural networks [27,11,59,26]. Then θ denotes the parameter vector of such a model and the integral $\int d\theta$ is a Bayesian integral over the possible parameter values with prior probability $p(\theta)$. The usual way to proceed is to use a learning algorithm to find an optimal approximation θ^*. This corresponds to a saddle point approximation of the θ–integral. Including such templates $t(\theta)$ in the regularization functional means that several different approximation methods can be combined and restrictions given by their parameterization can be overcome. Using a combined saddle point approximation for h and θ leads to stationarity equations which couple the optimal h^* and the optimal θ^*. Such a simultaneous saddle point approximation for h and θ uses the same training data D_i to determine both, h^* and θ^*. (This can be compared with boosting methods which have received much attention recently [63].) Using parameterized function spaces $t(x,\theta)$ as adaptive templates corresponds to the prior assumption that the structure of the data generating process is approximately captured by their parameterization. For example, $t(x,\theta)$ can model a probable hierarchical organization of the generating process.
2. (Approximate continuous symmetries) Typically, templates $t(x)$ can appear in several variations $t(x,\theta)$. These variations may be described by a continuous parameter vector θ. For example, one may wish to include translated, rotated, scaled or otherwise deformed 'eye'–templates in the regularization functional. Again, a combined saddle point approximation can be used to find the best fitting variant of the template $t^* = t(\theta^*)$ and an optimal approximation h^* simultaneously.

Consider now a general case of input, output, and generation noise. Furthermore let $i = (i_x, i_y, i_h)$ be a discrete mixture variable for which we want to treat the summation exactly, and $\theta = (\theta_x, \theta_y, \theta_h)$ a continuous mixture variable to be treated in maximum a posteriori approximation. It follows from Eq.(71) that

$$p(h|D) \propto \sum_i \int d\theta \, p(i_h) p(h|i_h) p(y|i_y,\theta_y) p(i_x,\theta_x|x) \prod_l^{N_L} p(i_{y_l},\theta_{y_l}|i_{x_l},\theta_{x_l},h). \quad (107)$$

Especially for quadratic concepts,

$$p(h|D) \propto \sum_i \int d\theta \, p(i_h,\theta_h) p(h|K_{i_h,\theta_h}, t_{i_h,\theta_h}) p(T|i_x,\theta_x) p(i_y,\theta_y|K) \times \prod_l p(t_{i_l,\theta_{x_l}}|K_{j_l,\theta_{y_l}}, h). \quad (108)$$

To obtain an error functional for minimization we write this in terms of energies and free energies

$$p(h|D) \propto \sum_i \int d\theta \, e^{-\beta_h E(i_h,\theta_h)} e^{-\beta_P (E(h|i_h,\theta_h) - F(\mathcal{H}|i_h,\theta_h))}$$
$$\times e^{-\beta_y (E(y|i_y,\theta_y) - F(\mathcal{Y}|i_y,\theta_y))} e^{-\beta_x (E(i_x,\theta_x|x) - F(\mathcal{I}_x,\Theta_x|x))}$$
$$\times e^{-\beta_L \sum_l \left(E(i_{y_l},\theta_{y_l}|i_{x_l},\theta_{x_l},h) - F(\mathcal{I}_l,\Theta_{y_l}|i_{x_l},\theta_{x_l},h)\right)}, \tag{109}$$

where $\theta_x \in \Theta_x$, $\theta_y \in \Theta_y$. In particular, for quadratic concepts, choosing $\beta_P = \beta_L = \beta$ and writing the generation probability $p(h|K_k, t_m)$ in likelihood form under uniform prior

$$p(h|D) \propto \sum_i \int d\theta \, p_{i,\theta} \, e^{-\beta \sum_l^{N_L+1} \left(E(t_{il}(\theta_{y_l})|K_{jl}(\theta_{x_L}),h) - F(\mathcal{T}_{il,\theta_{y_l}}|K_{il}(\theta_{x_l}),h)\right)}, \tag{110}$$

with $t_{i_{y_l},\theta_{y_l}} = t_{il}(\theta_{x_l})$, $K_{i_{x_l},\theta_{x_l}} = K_{il}(\theta_{y_l})$, and

$$p_{i,\theta} = \frac{p(i_h,\theta_h) p(T|i_y,\theta_y) p(i_x,\theta_x|x)}{\sum_{i_y} \int d\theta_y \, p(T|i_y,\theta_y)}. \tag{111}$$

Hence, instead of maximizing the posterior probability $p(h|D)$ we can minimize an error functional

$$E = -\ln \sum_i \int d\theta \, p_{i,\theta} \, e^{-\beta E_i(\theta) + c_i(\theta,h)}. \tag{112}$$

For h–, θ–, i–independent normalization factor $Z(\mathcal{T}_{il,\theta_{y_l}}|K_{il}(\theta_{x_l}),h)$ also

$$c_i(\theta,h) = -\ln Z(\mathcal{T}_{il,\theta_{y_l}}|K_{il}(\theta_{x_l}),h) + c, \tag{113}$$

with arbitrary constant c, is θ–, i– or h–independent and can thus be skipped. This is the case for gaussians with θ–, i–independent covariances K^{-1}/β.

Consider for example component energies E_i

$$E_i(\theta) = E_{i,1} + E_{i,2}(\theta), \quad E_{i,1} = \sum_j^{N_i} \frac{d_{ij}^2}{2}, \quad E_{i,2}(\theta) = \frac{d_i^2(\theta)}{2}, \tag{114}$$

where we separated an θ–independent part, with squared distances

$$d_{ij}^2 = \left\langle h - t_{ij} \left| K_{ij} \right| h - t_{ij} \right\rangle, \quad d_i^2(\theta) = \left\langle h - t_i(\theta) \left| K_i \right| h - t_i(\theta) \right\rangle. \tag{115}$$

Performing for differentiable $E_i(\theta)$ the integral in saddle point approximation and assuming $p_{i,\theta}$ to be, compared to $E_i(\theta)$, slowly varying at the stationary point, e.g. being uniform, (otherwise the derivative of $p_{i,\theta}$ has to be included) results in the stationarity equation

$$0 = \frac{dE}{d\theta} \Rightarrow 0 = \sum_i^{n_i} p_{i,\theta} \, e^{-\beta E_i(\theta) + c_i} \left(\frac{\partial E_{i,2}(\theta)}{\partial \theta} \right), \tag{116}$$

with

$$\frac{\partial E_{i,2}(\theta)}{\partial \theta} = \left\langle \frac{\partial t_i(\theta)}{\partial \theta} \middle| K_i \middle| t_i(\theta) - h \right\rangle. \tag{117}$$

This equation has to be solved simultaneously with the stationarity equation for h. Thus, to find the optimal h^* two (sets of) coupled stationarity equations have to be solved. As those are usually nonlinear, a self–consistent solution has to be found by iteration, starting from some initial guesses h^0 and θ^0. The iteration can be performed, for example, according to the following steps:

1. Obtaining h^0: Restrict the problem first to the training space $\mathcal{X}_T$, i.e., the space of x_k which are present in the training data D_T. For equally weighted, standard mean–square data terms a natural initial guess for $h(x_k)$ in that subspace is the observed average of y_k for the given x_k, i.e., $h^0(x_k) = \frac{1}{n_x}\sum_i^{n_x} y_i(x_i)$ $= t_T(x_k)$ with n_x the number of training data for x.
2. Obtaining θ^1: In the second step h^0 is inserted into Eq.(117), projected to the training subspace $\mathcal{X}_T$, and iterated with initial guess θ^0 to obtain a new θ^1 optimal in $\mathcal{X}_T$ for that h^0.
3. Obtain full h: The intermediate solution θ^1 is then used to solve for $h(x)$ for all $x \in \mathcal{X}_R$.
4. Continue iteration: The stationarity equations are solved by iteration until self–consistency is reached.

Consider the special case of choosing K proportional to the projector P_T into the space $\mathcal{X}_T$ of training data and using as initial guess for h the data template (see Eq.(59)) $h^0(x_k) = t_T(x_k)$. This is equivalent to a standard mean–square error minimization

$$0 = \sum_k \frac{\partial t_i(x_k, \theta)}{\partial \theta}(t_i(x_k, \theta) - y(x_k)). \tag{118}$$

For example, the template $t_i(x, \theta)$ can represent a neural network with weights and biases included in θ. In that case Eq.(118) could be solved by backpropagation. It is important to note that in our context the resulting network $t_i(\theta^*)$, optimal on the training data, is only the initial guess to be used for further iteration to obtain an optimal h^*. In later stages of the iteration $t_i(\theta)$ can for example be retrained by virtual examples drawn from $h(x)$.

In the case of continuous symmetry transformations $S_i(\theta)$ the transformed template is given by

$$t_i(\theta) = S_i(\theta)t_i = e^{\theta s_i} t_i, \tag{119}$$

where θ is the parameter vector of a continuous Lie group with vector of generators s (see 2.3). This yields the derivative $\partial t_i(\theta)/\partial\theta = s_i t_i(\theta) = s_i S_i(\theta) t_i$ hence

$$\frac{\partial E_{i,2}(\theta)}{\partial\theta} = \left\langle s_i t_i(\theta) \left| K_i \right| t_i(\theta) - h \right\rangle = \left\langle S_i(\theta) s_i t_i \left| K_i \right| h - S_i(\theta) t_i \right\rangle \tag{120}$$

for $s_i S_i = S_i s_i$. Using $S^{-1}(\theta) = S(-\theta)$, this can also be written as

$$\left\langle s_i t_i | S_i^T(\theta) K_i S_i(\theta) \left(t_i - S_i(-\theta) h\right) \right\rangle. \tag{121}$$

For translations, for example, one finds with $S(\theta)t(x) = t(x+\theta)$ and $S^T(\theta) = S^{-1}(\theta) = S(-\theta)$ for vanishing commutator $[K_i, s] = K_i s - s K_i = 0$

$$\left\langle st(\theta) \left| K_i \right| t(\theta) - h \right\rangle = \int dx\, dx' \, \left(t(x') - h(x'-\theta)\right) K(x,x') \frac{dt}{dx'}. \tag{122}$$

Example 6 (Adaptive templates) Figure 6 shows a simple example of an adaptive template depending on two parameters. In Figure 7 numerical results are presented for a corresponding one–dimensional two template model

$$E(\theta,\theta') = -\ln\left(e^{-\beta E_1(\theta)} + e^{-\beta E_2(\theta')}\right), \tag{123}$$

with mean–square data terms and the negative laplacian $-\Delta$ as concept operator K, i.e.,

$$E_1(\theta) = \frac{1}{2}\sum_k (y_k - h(x_k))^2 + \frac{1}{2}\int dx \, \left(\frac{\partial(h(x) - t_1(x,\theta_1,\theta_2))}{\partial x}\right)^2, \tag{124}$$

$$E_2(\theta') = \frac{1}{2}\sum_k (y_k - h(x_k))^2 + \frac{1}{2}\int dx \, \left(\frac{\partial(h(x) - t_2(x,\theta_1',\theta_2'))}{\partial x}\right)^2. \tag{125}$$

The initial guesses h^0, θ^0, θ'^0 have been obtained by projection into the space $\mathcal{X}_T$ according to the method discussed above. Then h and θ have been updated alternately using complete search for θ and θ' and an expectation–maximization (EM) algorithm for h.

4. Learning

The stationarity equations of the presented models are in general nonlinear, inhomogeneous integro–differential equations. One may remark, that similar equations appear for example in quantum mechanical scattering theory, where, similarly to templates or data, the inhomogeneities represent measurable asymp-

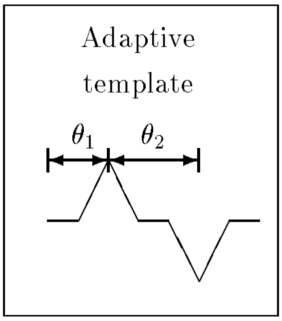

Figure 6. The figure shows one of the templates depending on two parameters used for model (123). Parameter θ_1 describes the location of the first maximum, θ_2 the distance between maximum and minimum.

totic states ("channels") of the system [43]. Nonlinear equations have to be solved by iteration [19,3,7]. Consider the equation to be solved written in a form

$$K(h)h = t(h). \tag{126}$$

Then an iteration procedure or *learning algorithm* is obtained by selecting an operator A, usually positive definite, and a relaxation factor η, to be used with the updating rule

$$h^{k+1} = h^k + \eta A^{-1}(t - Kh^k). \tag{127}$$

For η small enough and a positive definite A the function to be minimized decreases till reaching a local minimum. A may depend on the iteration step k and h^k. The gradient algorithm, for example, is obtained when taking $A = I$ equal to the identity. It does require matrix multiplication but no inversion. A gaussian A^{-1} on the other hand can approximate local correlations induced by differential operators. Choosing $A = K_M$ corresponds for error functional E_M to the expectation–maximization (EM) algorithm and Newton's method takes the negative Hessian. Figure 8 compares for the mixture model (90) gradient algorithm, an EM algorithm (labelled relaxation), and iteration with gaussian A^{-1} with the two different standard deviations $\sigma = 2$, $\sigma = 1$. It shows that the gradient has extreme difficulties with long range correlations. The gaussian A^{-1} performs well, at least at the beginning of the iteration. That means it captures well the covariance structure of that particular problem. This is useful, because in this case A^{-1} is given and so no inversion is needed. It can be seen that the gaussian algorithm especially at larger variance takes longer to adapt the fine structure of the function. This suggests to change the variance during iteration or to combine it with the gradient algorithm. The EM algorithm, which works here quite well, requires at every step inversion of the h–dependent K_M. The performance of specific algorithms is clearly problem dependent. Recently, multiscale or multigrid methods have been become particularly popular.

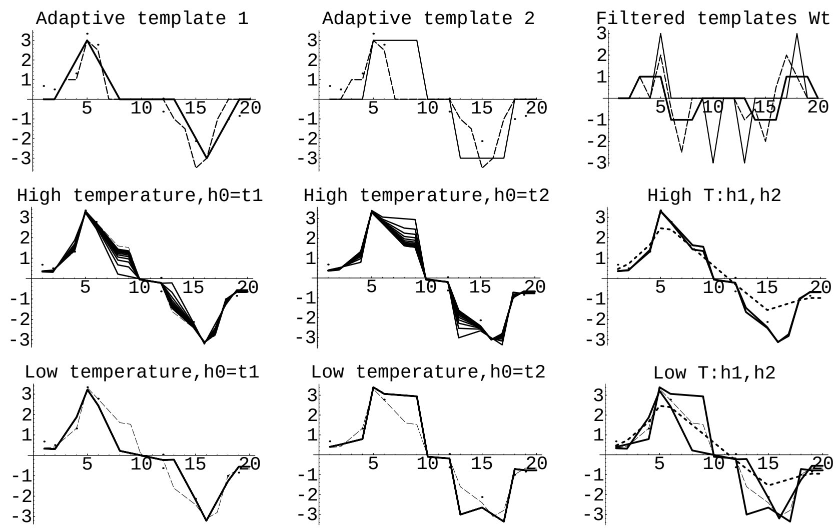

Figure 7. The figure presents numerical results for the model (123) with two adaptive templates $t_1(\theta_1, \theta_2)$ (shown in Fig.6 and another $t_2(\theta_1', \theta_2')$ on a mesh with 20 points. The figures in the first row show 1. ten data points (dots), the true function h (dashes) from which the data have been sampled with gaussian distribution (with $\sigma = 0.5$) and t_1 (thick) with θ_1, θ_2 optimized for the given data, 2. the same with t_2, and 3. the linearly transformed templates Wt_i and Wh. Hereby W is the derivative operator which is a concept filter for the negative laplacian. The second row shows for a high temperature case (here $\beta = 0.1$) 1. the solution h during iteration for initial guess $h^0 = t_1$ (thick), and the high temperature limit (thin dashes). The high temperature limit is the template average of data and both adaptive templates, and is the limiting case where the OR–like mixture becomes a gaussian AND. 2. The same for initial guess $h^0 = t_2$. 3. The final solutions for both cases and the classical solution for only one laplacian prior concept with zero template $t_0 \equiv 0$ (thick dashes). Notice, that in the high temperature case the two solutions coincide. The third row shows the same for a low temperature case (here $\beta = 0.5$) Notice that here the two solutions evolving from the two initial guesses $h^0 = t_1$ and $h^0 = t_2$ do not coincide.

5. Conclusion

A new and relatively general method has been proposed to construct problem specific error functionals (or posterior densities) utilizing complex, 'informative' a priori knowledge. For that purpose a priori information is decomposed into simple components representing constraints, like measured training data or approximate symmetries, which the function to be approximated is expected to fulfil. The constraints are represented by quadratic error terms which are combined using logical operations (conjunctions and disjunctions). Conjunctions of quadratic concepts lead to classical quadratic regularization functionals or, in Bayesian interpretation, to gaussian processes. Disjunctions, representing situations with ambiguous data or ambiguous priors, result in nonconvex models

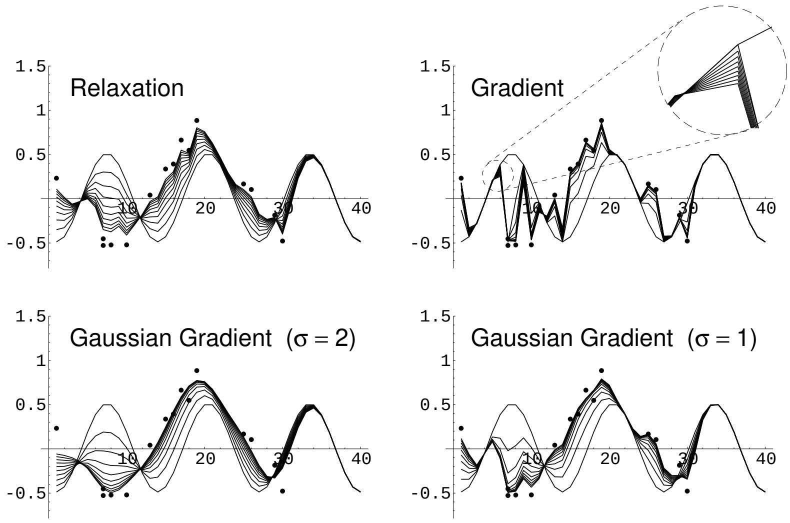

Figure 8. Comparison of learning algorithms for mixture model (90). Shown are the first iterations (starting with template t_b, see Fig.5) for the four iteration matrices $A = K_M$ (EM for mixture model (90), labelled relaxation) $A = I$ (gradient) and two gaussian A^{-1} with standard deviations $\sigma = 2$, $\sigma = 1$. The gradient has obviously difficulties capturing the long distance correlations and requires an extremely small relaxation factor (step width) η. The gaussian algorithm, which like the gradient requires no inversion, performs relatively well on a global scale without reaching the fine structure as fast.

going beyond classical regularization and gaussian process approaches. Two variants to treat ambiguous a priori informations have been discussed in more detail: mixture models for posterior densities and polynomial models related to the Landau–Ginzburg treatment of phase transitions. For simple numerical examples the feasibility of the approach has been demonstrated.

The presented approach might especially be useful for complex learning tasks with relatively few available training data. It seems also worth to study its relations to knowledge transfer and combination of learning systems in more detail.

Appendix

A. Statistics and Statistical Mechanics

Both disciplines, statistics and statistical mechanics, deal with probabilistic models. Their differences in language and methods can be traced back to differences in their typical applications.

Statistical mechanics has been developed for extremely large systems, like they appear in condensed matter physics. Typical systems of statistical mechanics are of high regularity, defined on a two or three dimensional grid with local variables having the same range of possible values (e.g. ± 1 for spins or real numbers for scalar fields), and only local interactions. There are relatively few prototypical systems, which have been studied extensively, for example, the celebrated Ising model. A complex machinery has been developed to obtain results with very high accuracy but requiring long and costly calculations.

Statistics, on the other hand, is mainly interested in the solution of a larger variety and more application oriented problems. Compared to prototypical systems studied in statistical mechanics, the corresponding models are therefore often smaller but less regular. The main practical problem consists in constructing adequate models. The models are usually not expected to allow very precise predictions. Hence, their is no need to achieve, by costly calculations, a numerical accuracy which is beyond the validity of the model. Needed are fast, flexible, and robust algorithms.

However, due to the increasing computing power becoming widely available, the gap between the two disciplines gets smaller. Their is now a growing and seminal interaction between the two disciplines, like in the development of Monte Carlo methods or graphical models. Typical areas where methods of statistical mechanics can be applied easiest have a large amount of quantitative data of the same kind, organized on a one, two, or three dimensional grid with dependencies dominated by local interactions. This, for example, is the case in Bayesian image reconstruction or when predicting financial time series. As it becomes therefore important to understand the languages of both approaches we will discuss in the following the relations between concepts of statistics and statistical mechanics.

We begin with some remarks concerning the construction of probability distributions:

1. (Normalization constants or partition sums) The specification of a probability $p(x)$ starts with unnormalized numbers $Z(x)$. To ensure the normalization condition $\int dx\, p(x) = 1$ a normalization constant $Z = \int dx\, Z(x)$ has to be calculated. For simple systems like a dice this is easy. For large systems, like in typical systems studied in statistical mechanics, this can be a highly non–trivial task. Thus, unnormalized functions $Z(x)$ instead of probabilities $p(x)$ are the natural starting point for large systems. Despite the fact that the Z are not probabilities because not normalized we will call them in the following *'unnormalized probabilities'* or *partition sums*.
2. (Log–probabilities, information and energy) A large system is usually constructed out of simpler subsystems, with the probabilities of the subsystems combined by multiplication, according to $p(x_1, x_2) = p(x_1)p(x_2|x_1)$. Sums, however, are easier to handle than products and represent a somewhat more intuitive concept. So it is easier to deal with infinite sums, or in the continuous case, with integrals, than with infinite or continuous products. This means, that large systems are easier constructed in terms of *logarithms of probability*

$\ln p(x)$. We will show in the following how *information* is related to expectations of the logarithmus of probabilities and *energy* to expectations of the logarithm of unnormalized probabilities Z.

3. (Conditional probabilities and disordered systems) Instead of directly constructing a complicated probability distribution $p(x)$ it is often easier to break $p(x)$ down in simpler parts. This is done by selecting conditions y under which $p(x|y)$ and $p(y)$ are relatively easy to specify. The total probability is then obtained by combining the alternatives y according to $p(x) = \int dy\, p(y)p(x|y)$. For example, it can be more easily to specify probabilities of symptoms for given specific diseases and probabilities of diseases than to write down directly a probability for the symptoms averaged over all diseases in one step. This, however, also means that due to the normalization requirements $\int dx\, P(x|y) = 1$, a normalization constant $Z(y)$ over x must be calculated for every y. In statistical mechanics such models are called *disordered systems*. Such averages over energy functions occur for example for spin glasses [46,18,53].

Thus, while statistics can often be formulated directly in terms of probabilities, statistical mechanics uses a formulation in terms of logarithms of unnormalized probabilities. In the following their relations to the concepts of energy and free energy will be discussed.

A.1. Probability

Especially for large and complex systems, it is convenient to work with the logarithm of 'unnormalized probabilities' instead directly with probabilities.

A.1.1. Normalization factors or partition sums

Let X be a random variable with possible values $x \in \mathcal{X}$ and assume a probability measure $p(A)$ defined on a σ–algebra of possible events $\mathcal{A}$ being subsets of a set $\mathcal{X}$. The event x may be represented by a vector of real numbers.

Denoting unnormalized probabilities by $Z(x) \propto p(x)$ we write

$$p(x) = \frac{Z(x)}{Z(\mathcal{X})} \tag{128}$$

and for general events A with $A \subseteq \mathcal{X}$, i.e., including $A = \mathcal{X}$

$$Z(A) = \int_{x\in A} dx\, Z(x) \propto p(A) = \int_{x\in A} dx\, p(x), \quad p(A) = \frac{Z(A)}{Z(\mathcal{X})}. \tag{129}$$

Introducing a second random variable Y and using the compatible normalization

$$Z(y) = Z(\mathcal{X}, y) = \int dx\, Z(x, y) \propto p(y) = \int dx\, p(x, y)) \tag{130}$$

gives

$$Z(\mathcal{X},\mathcal{Y}) = \int dx \int dy\, z(x,y) = \int dy\, z(y) = Z(\mathcal{Y}) \tag{131}$$

and therefore

$$p(y) = \frac{Z(y)}{Z(\mathcal{Y})} = \frac{Z(y)}{Z(\mathcal{X},\mathcal{Y})} \quad \text{and } p(x|y) = \frac{p(x,y)}{p(y)} = \frac{Z(x,y)}{Z(y)}. \tag{132}$$

Choosing $Z(y)$ to be not equal but proportional to $Z(\mathcal{X},y)$ one obtaines

$$p(x|y) = \frac{p(x,y)}{p(y)} = \frac{Z(x,y)\, Z(\mathcal{Y})}{Z(y)\, Z(\mathcal{X},\mathcal{Y})}. \tag{133}$$

In slight generalization of the language of statistical physics we call unnormalized probabilities Z also *partition sums*.

A.1.2. Log–probabilities, bit numbers, and free energies

Log–probabilities L are defined by

$$L(x) = \ln p(x), \quad p(x) = e^{L(x)} \tag{134}$$

and for $A \subseteq \mathcal{X}$

$$L(A) = \ln p(A), \quad p(A) = e^{L(A)}. \tag{135}$$

In terms of log–probabilities L a product like

$$p(x,y) = p(x)p(y|x) \tag{136}$$

becomes a sum

$$e^{L(x,y)} = e^{L(x)+L(x|y)}. \tag{137}$$

Log–probabilities are widely used in practice due to the fact that it is often more convenient to deal with sums than with products. Also the 'quenching' effect of the logarithm can be important in numerical calculations when $p(x)$ varies over several orders of magnitudes.

Common are especially negative log–probabilities b also called bit numbers

$$b(x) = -\ln p(x), \quad p(x) = e^{-b(x)} \tag{138}$$

and for $A \subseteq \mathcal{X}$

$$b(A) = -\ln p(A), \quad p(A) = e^{-b(A)}. \tag{139}$$

Analogously the *free energy* is defined for unnormalized probabilities Z by

$$F(x) = -\frac{1}{\beta}\ln Z(x), \quad Z(x) = Z(x|\beta) = e^{-\beta F(x)}, \quad p(x) = e^{-\beta(F(x)-F(\mathcal{X}))}. \tag{140}$$

and for $A \subseteq \mathcal{X}$

$$F(A) = -\frac{1}{\beta} \ln Z(A), \quad Z(A) = Z(A|\beta) = e^{-\beta F(A)}, \quad p(A) = e^{-\beta(F(A)-F(\mathcal{X}))}. \tag{141}$$

The dependency on β will in the following not always be denoted explicitly. For $x \in \mathcal{X}$ the factor $e^{-\beta F(x)}$ is also known as Boltzmann weight of x. The reasons for the introduction of the parameter β will be discussed later in detail.

A.2. Random variables

Information and energy are averages of special random variables. We discuss now their connection to bit numbers and free energy.

A.2.1. Averages

Recall the definition of an expectation or average of a random function $g(x)$ over $\mathcal{X}$

$$g_X = \langle g(x) \rangle_{\mathcal{X}} = \int dx\, p(x) g(x). \tag{142}$$

Analogously, we define an expectation over a subset $A \subseteq \mathcal{X}$

$$g_X(A) = \langle g(x,A) \rangle_A = \int dx\, p(x|A) g(x,A) = \int_{x \in A} dx\, \frac{p(x)}{p(A)} g(x,A) \tag{143}$$

using $p(x,A) = p(x)$ and $p(x|A) = p(x)/p(A)$ for $x \in A$. Using a second random variable Y this can be generalized to a conditional expectation of a random function $g(x,y)$ over event $Y = y$

$$g_X(y) = \langle g(x,y) \rangle_{\mathcal{X}|y} = \int dx\, p(x|y) g(x,y). \tag{144}$$

An average of form (143) is obtained by choosing in (144) a random variable Y taking the value y everywhere on A but not on its complement.

A.2.2. Information

A random variable C on $\mathcal{X}$ corresponds to a function $C(x)$ defined for every $x \in \mathcal{X}$. A special example is a transformation T of $p(x)$ which defines a corresponding random variable by $C(x) = T(p(x))$ for $x \in \mathcal{X}$. We will call in the following C the canonical random variable of the transformation T on $\mathcal{X}$. Specifically, for every distribution $p(x)$ the corresponding bit number $b(x) = -\ln p(x)$ can be considered as random function with the property of being defined not only x– but $p(x)$–dependent. The canonical random variable on $\mathcal{X}$ for bit numbers, i.e., for the transformation $-\ln p(x)$, will be called *information* $I(x) = b(x) = -L(x)$.

(For an axiomatic approach, properties of information and the definition of related quantities see for example [2,4,15].) Accordingly, Eqs.(138) can be written

$$I(x) = -\ln p(x), \quad p(x) = e^{-I(x)}. \tag{145}$$

Like any random variable the bit number or information $b(x)$ can be averaged over the whole set $\mathcal{X}$, over a subset $A \subseteq \mathcal{X}$, or conditioned on $Y = y$. One finds for the average or first moment of the bit number or information b the *average information* I_X

$$I_X(\mathcal{X}) = \langle b(x) \rangle_{\mathcal{X}} = -\langle \ln p(x) \rangle_{\mathcal{X}} \tag{146}$$

$$I_X(A) = \langle b(x) \rangle_A = -\langle \ln p(x) \rangle_A \tag{147}$$

$$I_X(y) = \langle b(x,y) \rangle_{\mathcal{X}|y} = -\langle \ln p(x,y) \rangle_{\mathcal{X}|y} \tag{148}$$

In Eq.(147) we used $b(x,A) = b(x)$ for $x \in A$ and $p(x|A) \ln p(x,A) = 0 \ln 0 = 0$ for $x \notin A$ and in Eq.(148) we allowed y–dependent $p(x,y)$. It follows in accordance with the definition from Eq.(147) for $A = \{x\}$ that $I_X(A = \{x\}) = I_X(x) = I(x) = b(x)$ or analogously $I_X(Y = X = x) = b(x,x) = b(x)$ from Eq.(148). While I and b coincide on events $x \in \mathcal{X}$, in general the difference between bit number and average information, i.e., the difference between transformed probability and expectations of the corresponding canonical random variable, is given by

$$I_X(y) - b(y) = \langle b(x,y) \rangle_{\mathcal{X}|y} - b(y) = -\langle \ln p(x,y) \rangle_{\mathcal{X}|y} + \ln p(y)$$

$$= -\left\langle \ln \frac{p(x,y)}{p(y)} \right\rangle_{\mathcal{X}|y} = -\langle \ln p(x|y) \rangle_{\mathcal{X}|y} = \langle b(x|y) \rangle_{\mathcal{X}|y}. \tag{149}$$

Defining the *entropy* (conditional information)

$$H_X(y) = \langle b(x|y) \rangle_{\mathcal{X}|y} \tag{150}$$

this can be written

$$I_X(y) - b(y) = H_X(y). \tag{151}$$

including

$$I_X(A) - b(A) = H_X(A). \tag{152}$$

The relations $b(\mathcal{X}) = 0$ and $I_X(x) = b(x)$ yield the special cases

$$H_X(\mathcal{X}) = I_X(\mathcal{X}), \quad H_X(x) = 0. \tag{153}$$

A.2.3. Energy

In the same way as $b(x)$ also $F(x)$ may be interpreted as random variable on $\mathcal{X}$ called *energy* $E(x) = F(x)$. (E is also called (euclidian) *action* in field theory). Hence, energy is the canonical random variable of the transformation $-\frac{1}{\beta}\ln(Z(\mathcal{X})p(x))$, and one can write for Eqs.(140)

$$E(x) = -\frac{1}{\beta}\ln Z(x), \quad Z(x) = Z(x|\beta) = e^{-\beta E(x)}, \quad p(x) = e^{-\beta(E(x)-F(\mathcal{X}))}. \tag{154}$$

In general βF can be decomposed into $\beta F = \sum_i \alpha_i E_i$. (This is the case, for example, in the grand canonical ensemble of statistical physics where, compared to the canonical ensemble, a component corresponding to the particle number is added.)

The analogue of average informations I are then averages of the energy $E(x) = F(x)$ called *average energies* E_X

$$E_X(\mathcal{X}) = \langle F(x)\rangle_{\mathcal{X}} = -\frac{1}{\beta}\langle \ln Z(x)\rangle_{\mathcal{X}} \tag{155}$$

$$E_X(A) = \langle F(x)\rangle_A = -\frac{1}{\beta}\langle \ln Z(x)\rangle_A, \tag{156}$$

$$E_X(y) = \langle F(x,y)\rangle_{\mathcal{X}|y} = -\frac{1}{\beta}\langle \ln Z(x,y)\rangle_{\mathcal{X}|y}. \tag{157}$$

It follows in accordance with the definition from Eq.(156) for $A = \{x\}$ that $E_X(A = \{x\}) = E_X(x) = E(x) = F(x)$ or analogously $E_X(Y = X = x) = F(x,x) = F(x)$ from Eq.(157). We will skip the subscripts X in the following. While free energy and energy coincide on events $x \in \mathcal{X}$ in general their difference is given by

$$\beta E(y) - \beta F(y) = -\langle \ln Z(x,y)\rangle_{\mathcal{X}|y} + \ln Z(y)$$

$$= -\left\langle \ln\frac{Z(x,y)}{Z(y)}\right\rangle_{\mathcal{X}|y} = -\left\langle \ln\left(p(x|y)\frac{Z(\mathcal{Y})}{Z(\mathcal{X},\mathcal{Y})}\right)\right\rangle_{\mathcal{X}|y}. \tag{158}$$

Choosing

$$F(y) = -\frac{1}{\beta}\ln Z(y) = -\frac{1}{\beta}\ln\int dx\, e^{-\beta F(x,y)} \tag{159}$$

i.e.,

$$Z(y) = Z(\mathcal{X},y) = \int dx\, Z(x,y) \tag{160}$$

it follows

$$Z(\mathcal{Y}) = \int dy\, Z(y) = \int dx \int dy\, Z(x,y) = Z(\mathcal{X},\mathcal{Y}). \tag{161}$$

Therefore

$$-\langle \ln Z(x,y)\rangle_{\mathcal{X}|y} + \ln Z(y) = -\langle \ln p(x|y)\rangle_{\mathcal{X}|y} \tag{162}$$

and

$$\langle F(x,y)\rangle_{\mathcal{X}|y} - F(y) = \frac{1}{\beta}\langle b(x|y)\rangle_{\mathcal{X}|y}. \tag{163}$$

Assuming $F(x|y)$ here to be defined as

$$F(x|y) = -\frac{1}{\beta}\ln\frac{Z(x,y)}{Z(y)} = -\frac{1}{\beta}\ln\frac{e^{-\beta F(x,y)}}{e^{-\beta F(y)}} = -\frac{1}{\beta}\ln\frac{e^{-\beta F(x,y)}}{\int dx\, e^{-\beta F(x,y)}}, \tag{164}$$

i.e.,

$$Z(x|y) = \frac{Z(x,y)}{Z(y)} = p(x|y), \tag{165}$$

one can also write for Eq.(163)

$$\langle F(x,y)\rangle_{\mathcal{X}|y} - F(y) = \langle F(x|y)\rangle_{\mathcal{X}|y}, \tag{166}$$

parallelizing the corresponding equation(149) for bit numbers or informations b. Using the definitions of entropy $H(y) = \langle b(x|y)\rangle_{\mathcal{X}|y}$ and energy $E(y) = \langle F(x,y)\rangle_{\mathcal{X}|y}$ Eq.(163) states that the β–scaled difference between energy and free energy is the entropy

$$\beta E(y) - \beta F(y) = H(y). \tag{167}$$

This is a generalization of the well known relation of statistical physics

$$\beta E(\mathcal{X}) - \beta F(\mathcal{X}) = H(\mathcal{X}), \tag{168}$$

where the argument $\mathcal{X}$ is usually skipped.

The following table summarizes some of the relations (recall that the variable y in the table can be replaced by $x \in \mathcal{X}$ or $A \subseteq \mathcal{X}$ and note that for $Z = \beta = 1$ free energy and energy become identical to bit number and information):

Transformed probability	Averages of canonical random variable	Difference
bit number $b(y) = -\ln p(y)$	information $I(y) = -\langle \ln p(x)\rangle_{\mathcal{X}\|y}$	entropy $H(y) = -\langle \ln p(x\|y)\rangle_{\mathcal{X}\|y}$ $= I(y) - b(y)$
free energy $F(y) = -\frac{1}{\beta}\ln Z(y)$	energy $E(y) = -\frac{1}{\beta}\langle \ln Z(x)\rangle_{\mathcal{X}\|y}$	entropy $H(y) = -\langle \ln p(x\|y)\rangle_{\mathcal{X}\|y}$ $= \beta(E(y) - F(y))$

A.3. Temperature and external fields

Now we look at the role of the parameter β. Its inverse $T = 1/\beta$ is called *temperature* in statistical mechanics. The parameter β can also be interpreted as an *external source or field* coupling to the conjugated random variable energy.

The energy can thereby be subdivided into several components E_i with corresponding conjugated β_i. One β_i, for example, can be proportional to a magnetic field (or chemical potential, pressure, $\cdots$) coupling to a magnetic moment $E_i = M$ (or particle number, volume, $\cdots$). Also the calculation of moments or cumulants of other random variables is often facilitated by introducing an external source coupling to that variable.

We will discuss the following roles of β:

1. β is Lagrange parameter determining the expectation of the energy $\langle E\rangle = \int dx\, p(x)E(x)$. Its variation defines an exponential family.
2. β is a homotopy parameter used by annealing methods, interpolating between easy and difficult to solve problems.
3. β represents an external source or field coupling to the energy. The cumulants of E can be obtained as responses to a changing external field, i.e., as derivatives of $\ln Z$ with respect to β. For example, $\langle E\rangle_{\mathcal{X}} = -(\partial/\partial\beta)\ln Z(\mathcal{X})$.

A.3.1. Maximum entropy and Boltzmann–Gibbs distributions

It is well known that minimizing the entropy and fixing normalization condition $\int dx\, p(x) = 1$ and expectations $\int dx\, p(x)E_i(x) = E_i(\mathcal{X})$ by the Lagrange multiplier method yield Boltzmann–Gibbs (or generalized canonical) distributions. Indeed, adding the constraints with Lagrange multipliers α_i and setting to zero the functional derivative of

$$H(y) - \sum_{i=1} \alpha_i E_i(y) - \alpha_0 \langle 1\rangle_{\mathcal{X}|y} \tag{169}$$

$$= -\int dx \left(p(x|y)\ln p(x|y) - \sum_{i=1} \alpha_i E_i(x,y)p(x|y) - \alpha_0 p(x|y)\right), \tag{170}$$

with respect to $p(x|y)$

$$\begin{aligned} 0 \;&=\; \frac{\delta}{\delta p(x|y)}\left(H(y) - \sum_{i=1}\alpha_i E_i(y) - \alpha_0 \langle 1\rangle_{\mathcal{X}|y}\right) &(171)\\ &=\; -\ln p(x|y) - 1 - \sum_{i=1}\alpha_i E_i(x,y) - \alpha_0, &(172)\end{aligned}$$

one finds

$$p(x|y) = e^{-\sum_{i=1}\alpha_i E_i(x,y)-\alpha_0-1} = \frac{e^{-\beta F(x|y)}}{Z(\mathcal{X}|y)}, \tag{173}$$

with

$$\sum_{i=1}\alpha_i E_i(x,y) = \beta E(x,y) = \beta F(x|y) \quad\text{and}\quad Z(\mathcal{X}|y) = e^{\alpha_0+1}. \tag{174}$$

For $p(x|y) = p(x)$ this gives Eq.(154). Thus any probability distribution $p(x)$ can be seen as result of a maximum entropy procedure with normalization constraint and fixed expectation of the energy $E(x)$.

A.3.2. Annealing methods

The Lagrange multiplier β, respectively the temperature $T = 1/\beta$, determines the average energy. Introduction of several Lagrange multipliers allows the fixation of several components E_i of $E(x)$. Varying β defines an exponential family with canonical parameter β and and canonical statistic E. In the high temperature limit $T \to \infty$, i.e., $\beta \to 0$, all $p(x)$ become equal. In the low temperature limit $T \to 0$, i.e., $\beta \to \infty$, only events x^* with maximal $p(x^*) \ge p(x)$, $\forall x \in \mathcal{X}$ survive, while all other events x with $p(x) < p(x^*)$, $\exists x^* \in \mathcal{X}$ are damped exponentially with decreasing temperature. The temperature dependency is used by annealing methods which are specific realizations of general homotopy or parameter continuation methods and very important in practice [36,46,62,75]. They solve a difficult (e.g. multimodal) problem at finite or zero temperature by beginning with an easier (e.g. convex) high temperature problem and then slowly decrease the temperature.

A.3.3. Generating functions

Moments or cumulants of random variables can often be conveniently calculated by the use of generating functions. The nth *moment* M_n of a random function $g(x)$, with $E(x)$ being a special case, is the expectation of its nth power

$$M_n(g) = \langle g^n(x)\rangle_{\mathcal{X}}\,, \quad \text{e.g.} \quad M_n(E) = \langle E^n(x)\rangle_{\mathcal{X}}\,. \tag{175}$$

For a vector valued function g with components g_i (e.g. E_i), $i \in \mathcal{I}$ the moments become the functions (*unconnected correlation functions*)

$$M_{i_1,i_2,\cdots,i_n} = \langle g_{i_1}(x) g_{i_2}(x)\cdots g_{i_n}(x)\rangle_{\mathcal{X}}\,. \tag{176}$$

The *cumulants* (or *connected correlation functions*) are given by [20,48]

$$C_{i_1,i_2,\cdots,i_n} = \sum_{\mathcal{P}} (-1)^{m-1}(m-1)!\; M_{j_1,j_2,\cdots,j_{p_1}} M_{k_1,k_2,\cdots,k_{p_2}} \cdots M_{l_1,l_2,\cdots,l_{p_m}} \tag{177}$$

with inverse

$$M_{i_1,i_2,\cdots,i_n} = \sum_{\mathcal{P}} C_{j_1,j_2,\cdots,j_{p_1}} C_{k_1,k_2,\cdots,k_{p_2}} \cdots C_{l_1,l_2,\cdots,l_{p_m}} \tag{178}$$

where $\mathcal{P}$ denotes a partition of the n indizes into non–empty subsets and m is the number of factors in the summand and one takes $C_0 = 0$. For a small number m of components i moments and cumulants may be more conveniently indexed by "occupation numbers" n_i

$$M_{(n_1,n_2,\cdots,n_m)} = \langle g_1^{n_1}(x) g_2^{n_2}(x)\cdots g_n^{n_m}(x)\rangle_{\mathcal{X}}\,. \tag{179}$$

For scalar function g, i.e., in the one component case $\mathcal{I} = \{1\}$, we will write the nth moment $M_{1,1,\cdots,1} = M_{(n)} = M_n$ and nth cumulant $C_{1,1,\cdots,1} = C_{(n)} = C_n$, skipping the bracket for the sake of simplicity. Hence, for a scalar g

$$M_0 = 1, \quad M_1 = C_1, \quad M_2 = C_2 + (C_1)^2, \quad M_3 = C_3 + 3C_2C_1 + (C_1)^3, \quad (180)$$

$$C_0 = 0, \quad C_1 = M_1, \quad C_2 = M_2 - (M_1)^2, \quad C_3 = M_3 - 3M_2M_1 + 2(M_1)^3, \quad (181)$$

where the second cumulant is the well known variance. Unlike moments, the cumulants are additive for independent subsystems, i.e., $p(x_1, x_2) = p(x_1)p(x_2) \Rightarrow C_n(p(x_1, x_2)) = C_n(p(x_1)) + C_n(p(x_2))$. Another significant property of cumulants is the possibility to set consistently $C_n = 0$ for all $n > 2$ (for gaussian distributions), which is not possible for moments. If, however, one $C_n \neq 0$ for $n > 2$ then automatically an infinite number of other C_m do also not vanish (See for example [20]). For a multidimensional gaussian distribution with vanishing means $M_i = 0$ Eq.(178) reduces to a sum over two–point connected correlation functions (or propagators) C_{ij}

$$M_{i_1,i_2,\cdots,i_{2n}} = \sum_{\text{Pairings}} C_{j_1,k_1} C_{j_2,k_2} \cdots C_{j_n,k_n}. \quad (182)$$

This relation is known as Wick's theorem.

For a scalar function $g(x)$ the *moment generating function* is given by

$$M(\gamma) = \frac{Z(\gamma)}{Z} = \left\langle e^{\gamma g(x)} \right\rangle_{\mathcal{X}} = \int dx\, p(x) e^{\gamma g(x)}$$
$$= \frac{1}{Z} \int dx\, e^{-\beta E(x) + \gamma g(x)} = \sum_{n=0}^{\infty} \frac{\gamma^n}{n!} M_n, \quad (183)$$

with

$$Z(\gamma) = \int dx\, e^{-\beta E(x) + \gamma g(x)}. \quad (184)$$

For vector g (with discrete or continuous index set $\mathcal{I}$ one has $\gamma g = \sum_i \gamma_i g_i$ (or $\int di$ for continuous i) and $\gamma^n M_n$ has to be understood as $\langle (\sum_i \gamma_i g_i(x))^n \rangle_{\mathcal{X}}$, i.e.,

$$M(\gamma) = \left\langle e^{\sum_i^m \gamma_i g_i(x)} \right\rangle_{\mathcal{X}} = \sum_{n=0}^{\infty} \frac{1}{n!} \left\langle \left(\sum_i^m \gamma_i g_i(x) \right)^n \right\rangle_{\mathcal{X}}$$

$$= \sum_{n=0}^{\infty} \sum_{n_1,\ldots,n_m}^{n} \delta(\sum_i^m n_i - n) \frac{\gamma_1^{n_1} \cdots \gamma_n^{n_m}}{n_1! n_2! \cdots n_m!} M_{(n_1,n_2,\cdots,n_m)},$$

$$= \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{i_1,\ldots,i_n}^{m} \gamma_{i_1} \cdots \gamma_{i_n} M_{i_1,i_2,\cdots,i_n}. \quad (185)$$

It is easy to verify that the moments can be obtained by

$$\frac{\partial^n}{\partial\gamma^n}M(\gamma)\Big|_{\gamma=0} = \frac{\partial^n}{\partial\gamma^n}\left\langle e^{\gamma g(x)}\right\rangle_{\mathcal{X}}\Bigg|_{\gamma=0}$$

$$= \left\langle \frac{\partial^n}{\partial\gamma^n} e^{\gamma g(x)}\right\rangle_{\mathcal{X}}\Bigg|_{\gamma=0} = \left\langle g^n(x) e^{\gamma g(x)}\right\rangle_{\mathcal{X}}\Bigg|_{\gamma=0} = \langle g^n(x)\rangle_{\mathcal{X}} = M_n(g). \tag{186}$$

or in the multidimensional case

$$\frac{\partial^n}{\partial\gamma_{i_1}\cdots\partial\gamma_{i_n}}M(\gamma)\Big|_{\gamma=0} = \frac{\partial^n}{\partial\gamma_{i_1}\cdots\partial\gamma_{i_n}}\left\langle e^{\sum_i \gamma_i g_i(x)}\right\rangle_{\mathcal{X}}\Bigg|_{\gamma=0} = M_{i_1,i_2,\cdots,i_n}(g). \tag{187}$$

The cumulants are generated by differentiating the logarithm $\ln M(\gamma) = C(\gamma)$

$$\frac{\partial^n}{\partial\gamma^n}C(\gamma)\Big|_{\gamma=0} = \frac{\partial^n}{\partial\gamma^n}\ln\left\langle e^{\gamma g(x)}\right\rangle_{\mathcal{X}}\Bigg|_{\gamma=0} = C_n(g) \tag{188}$$

$$\frac{\partial^n}{\partial\gamma_{i_1}\cdots\partial\gamma_{i_n}}C(\gamma)\Big|_{\gamma=0} = \frac{\partial^n}{\partial\gamma_{i_1}\cdots\partial\gamma_{i_n}}\ln\left\langle e^{\sum_i \gamma_i g_i(x)}\right\rangle_{\mathcal{X}}\Bigg|_{\gamma=0} = C_{i_1,i_2,\cdots,i_n}(g). \tag{189}$$

Hence, $C(\gamma)$ is the *cumulant generating function* with Taylor expansion around $\gamma = 0$

$$C(\gamma) = \ln M(\gamma) = \ln\left\langle e^{\gamma g(x)}\right\rangle_{\mathcal{X}} = \sum_{n=0}^{\infty}\frac{\gamma^n}{n!}C_n(g), \tag{190}$$

or in the multidimensional case

$$C(\gamma) = \ln\left\langle e^{\sum_i \gamma_i g_i(x)}\right\rangle_{\mathcal{X}} = \sum_{n=0}^{\infty}\frac{1}{n!}\sum_{i_1,\ldots,i_n}\gamma_{i_1}\cdots\gamma_{i_n}C_{i_1,i_2,\cdots,i_n}. \tag{191}$$

Analogous to β the parameter γ can be thought as an external source or field (e.g., a magnetic field) coupling to g. Even if a field γ is not present in 'reality' the cumulants can still be calculated as derivatives at $\gamma = 0$. If a nonzero field γ_0 is present we can incorporate $\gamma_0 g$ in a new energy $-\beta E = -\beta' E' + \gamma_0 g$ replacing a given $\beta' E'$ and proceed as before. Because sometimes useful, especially to obtain the Legendre transform of C, we will give in the following some of the formulae explicitly for nonzero field γ_0. Assuming that derivatives and integration can be interchanged one has

$$\frac{\partial^n}{\partial\gamma^n}M(\gamma)\Big|_{\gamma=\gamma_0} = \frac{\partial^n}{\partial\gamma^n}\left\langle e^{\gamma g(x)}\right\rangle_{\mathcal{X}}\Bigg|_{\gamma=\gamma_0} = \left\langle g^n(x) e^{\gamma_0 g(x)}\right\rangle_{\mathcal{X}} = \frac{Z(\gamma_0)}{Z}M_{n,\gamma_0}(g). \tag{192}$$

Including the γ_0–term in the expectation $\left<\cdots\right>_{\mathcal{X}|\gamma_0}$ we find for

$$M_{\gamma_0}(\gamma) = \frac{Z(\gamma)}{Z(\gamma_0)} = \frac{M(\gamma)}{M(\gamma_0)} = \left< e^{(\gamma-\gamma_0)g} \right>_{\mathcal{X}|\gamma_0} \tag{193}$$

as derivatives

$$\frac{\partial^n}{\partial\gamma^n} M_{\gamma_0}(\gamma)\Big|_{\gamma=\gamma_0} = \frac{\partial^n}{\partial\gamma^n} \frac{Z(\gamma)}{Z(\gamma_0)}\Bigg|_{\gamma=\gamma_0} = \frac{\partial^n}{\partial\gamma^n} \left(\frac{\int dx\, e^{-\beta E(x)+\gamma g(x)}}{\int dx\, e^{-\beta E(x)+\gamma_0 g(x)}} \right) \Bigg|_{\gamma=\gamma_0}$$

$$= \left< g^n e^{(\gamma-\gamma_0)g} \right>_{\mathcal{X}|\gamma_0} \Big|_{\gamma=\gamma_0} = \left< g^n(x) \right>_{\mathcal{X}|\gamma_0} = M_{n,\gamma_0}(g). \tag{194}$$

Cumulants for nonzero fields are generated by $C_{\gamma_0} = \ln M_{\gamma_0}$

$$\frac{\partial^n}{\partial\gamma^n} C_{\gamma_0}(\gamma)\Big|_{\gamma=\gamma_1} = \frac{\partial^n}{\partial\gamma^n} \ln \left< e^{(\gamma-\gamma_0)g(x)} \right>_{\mathcal{X}|\gamma_0} \Bigg|_{\gamma=\gamma_1} = C_{n,\gamma_1}(g), \tag{195}$$

where $\gamma_1 = \gamma_0$ is possible. Because an additive constant does not change the derivatives of a cumulant generating function also $C(\gamma) = \ln M(\gamma) = \ln M_{\gamma_0}(\gamma) + \ln Z(\gamma_0) - \ln Z(\gamma) = \ln M_{\gamma_0}(\gamma) + \ln M(\gamma_0)$ can be used as cumulant generating function for $\gamma_0 \neq 0$. The expansion of the generating functions M_{γ_0} and C_{γ_0} around γ_0 in powers of $(\gamma - \gamma_0)$ becomes

$$M_{\gamma_0}(\gamma) = \left< e^{(\gamma-\gamma_0)g(x)} \right>_{\mathcal{X}|\gamma_0} = \sum_{n=0}^{\infty} \frac{(\gamma-\gamma_0)^n}{n!} M_{n,\gamma_0}(g), \tag{196}$$

$$C_{\gamma_0}(\gamma) = \ln M(\gamma) = \ln \left< e^{(\gamma-\gamma_0)g(x)} \right>_{\mathcal{X}|\gamma_0}$$

$$= \sum_{n=0}^{\infty} \frac{(\gamma-\gamma_0)^n}{n!} C_{n,\gamma_0}(g) = \sum_{n=0}^{\infty} \frac{(\gamma-\gamma_1)^n}{n!} C_{n,\gamma_1}(g), \tag{197}$$

and analogously for the multidimensional case. Moments and cumulants for different fields γ_0 and γ_1 are related according to

$$M_{n,\gamma_0} = \frac{Z(\gamma_1)}{Z(\gamma_0)} \frac{\partial^n}{\partial\gamma^n} M_{\gamma_1}(\gamma)\Big|_{\gamma=\gamma_0} = \frac{\left< g^n e^{(\gamma_0-\gamma_1)g} \right>_{\mathcal{X}|\gamma_1}}{\left< e^{(\gamma_0-\gamma_1)g} \right>_{\mathcal{X}|\gamma_1}} = \frac{\sum_{m=0}^{\infty} \frac{(\gamma_0-\gamma_1)^m}{m!} M_{m+n,\gamma_1}}{\sum_{k=0}^{\infty} \frac{(\gamma_0-\gamma_1)^k}{k!} M_{k,\gamma_1}}, \tag{198}$$

and

$$C_{n,\gamma_0} = \frac{\partial^n}{\partial\gamma^n} C_{\gamma_1}(\gamma)\Big|_{\gamma=\gamma_0} = \sum_{m=0}^{\infty} \frac{(\gamma_0-\gamma_1)^m}{m!} C_{m+n,\gamma_1}, \tag{199}$$

i.e., for the difference

$$\Delta C_n(\gamma_0,\gamma_1) = C_{n,\gamma_0} - C_{n,\gamma_1} = \sum_{m=1}^{\infty} \frac{(\gamma_0-\gamma_1)^m}{m!} C_{m+n,\gamma_1}. \tag{200}$$

Because inverse temperature β can be regarded as a special nonzero field, the moments and cumulants of the energy can be obtained by

$$\frac{\partial^n}{\partial(-\beta)^n} Z = Z M_n(E) \tag{201}$$

$$\frac{\partial^n}{\partial(-\beta)^n} \ln Z = C_n(E) \tag{202}$$

like

$$\frac{\partial^2}{\partial(-\beta)^2} \ln Z = C_2(E) = \left\langle E^2(x) \right\rangle_{\mathcal{X}} - \langle E(x) \rangle_{\mathcal{X}}^2 . \tag{203}$$

The generalization to the multidimensional case is analogous to Eqs.(185), (187). Equations for the second cumulants connecting the derivative with respect to an external field (response, dissipation) with a variance (fluctuation) are also known under the name dissipation–fluctuation theorems.

Moments of a function g of x can be expressed by moments of $\Delta x = x - x_0$ by expanding $g(x)$ around x_0

$$g(x) = \sum_{n=0}^{\infty} \frac{(x-x_0)^n}{n!} \frac{\partial^n}{\partial x^n} g(x)\Big|_{x=x_0} = e^{\langle \Delta x | \nabla' \rangle} g(x')\Big|_{x'=x_0} = e^{\langle \Delta x | \nabla \rangle} g(x_0), \tag{204}$$

with gradient $\nabla' g(x') = (\partial/\partial x') g(x') = g^{(1)}(x')$ and analogously $(\nabla')^n g(x') = (\partial^n/\partial(x')^n) g(x') = g^{(n)}(x')$. We understand here and in the following the expression $e^{\langle \Delta x | \nabla \rangle}$ or $\langle \Delta x | \nabla \rangle^n$ to be "normal ordered", meaning that the derivatives act only to the right and not on Δx. Analogously in the multidimensional case for example $\nabla^2 = \Delta$ creates the Hessian matrix. This yields,

$$M_n = \langle g(x) \rangle_{\mathcal{X}} = \left\langle g(x_0) + \Delta x\, g^{(1)}(x_0) + \frac{(\Delta x)^2}{2} g^{(2)}(x_0) + \cdots \right\rangle_{\mathcal{X}} \tag{205}$$

$$= \left\langle e^{\langle \Delta x | \nabla \rangle} g(x_0) \right\rangle_{\mathcal{X}} = g(x_0) + \langle \Delta x \rangle_{\mathcal{X}} g^{(1)}(x_0) + \frac{\left\langle (\Delta x)^2 \right\rangle_{\mathcal{X}}}{2} g^{(2)}(x_0) + \cdots . \tag{206}$$

We have seen that an expansion in moments or cumulants depends on the the choice of γ_0 or, equivalently, on the splitting of the x–dependent terms in one term $-\beta E(x)$ which defines an expectation $\langle \cdots \rangle_{\mathcal{X}}$ and a field term $\gamma g(x)$. Thus, there is a practically important freedom in choosing different moment or cumulant expansions as approximations for the same exponential. Assume for example that moments for $p'(x) \propto e^{-\beta' E'(x)}$ can be easily calculated. For $-\beta E = -\beta' E' + \gamma g(x)$

we can then write Eq.(198) for an expectation $\langle\cdots\rangle_{\mathcal{X}}$ of a function $q(x)$ under $p(x) \propto e^{-\beta E(x)}$

$$\langle q(x)\rangle_{\mathcal{X}} = \langle q e^{\gamma g}\rangle'_{\mathcal{X}} \frac{Z'}{Z} = \frac{\langle q e^{\gamma g}\rangle'_{\mathcal{X}}}{\langle e^{\gamma g}\rangle'_{\mathcal{X}}} = \frac{\langle q\,(1+\gamma g+\cdots)\rangle'_{\mathcal{X}}}{\langle 1+\gamma g+\cdots\rangle'_{\mathcal{X}}}, \tag{207}$$

with $Z' = \int dx\, e^{-\beta' E'(x)}$ and $\langle\cdots\rangle'_{\mathcal{X}} = \int dx\, p'(x)\cdots$. By expansion around x_0 this can also be expressed in terms of moments of Δx

$$\langle q(x)\rangle_{\mathcal{X}} = \frac{e^{\gamma g(x_0)}\left(q(x_0) + \langle\Delta x\rangle'_{\mathcal{X},(\gamma)}\left(q^{(1)}(x_0) + q(x_0)\gamma g^{(1)}(x_0)\right) + \cdots\right)}{e^{\gamma g(x_0)}\left(1 + \langle\Delta x\rangle'_{\mathcal{X},(\gamma)}\,\gamma g^{(1)}(x_0) + \cdots\right)}, \tag{208}$$

where the prefactor cancels. This is the basis of saddle point approximation (or loop expansion) which will be discussed in Section B.5 and also the basis of importance sampling in Monte Carlo calculations [8,9,21,50].

To obtain equations which go beyond a cumulant expansion it is useful to consider the Legendre transform of $C(\gamma)$. The *Legendre transform* (or *effective action*) $\Gamma(\phi)$ of $C(\gamma)$ is defined by requiring [61,25,77]

$$\Gamma(\phi) + C(\gamma) - \sum_i \gamma_i\phi_i, \tag{209}$$

to be stationary with respect to variations of fields γ_i (coupling to g_i) at ϕ fixed. (We may remark here that $C(\gamma)$ and thus its Legendre transform Γ depends on the choice of g_i. A typical case is $g_i = x_i$.) This means

$$\frac{\partial C(\gamma)}{\partial\gamma_i} = \phi_i = \langle g\rangle_{\mathcal{X}|\gamma} = C_{i,\gamma}, \tag{210}$$

using also that C is the cumulant generating function. Using the chain rule one finds by differentiating functional (209) with respect to ϕ_i

$$\frac{\partial\Gamma(\phi)}{\partial\phi_i} = -\sum_i \frac{\partial C(\gamma)}{\partial\gamma_i}\frac{\partial\gamma_i}{\partial\phi_i} + \gamma_i + \sum_i \phi_i\frac{\partial\gamma_i}{\partial\phi_i}, \tag{211}$$

and thus with Eq.(210)

$$\frac{\partial\Gamma(\phi)}{\partial\phi_i} = \gamma_i. \tag{212}$$

Now set

$$\Gamma(\phi) = \sum_{n=1}^{\infty}\frac{1}{n!}\sum_{i_1,\cdots,i_n}\Gamma_{i_1,\cdots,i_n}\Delta C_{i_1}\cdots\Delta C_{i_n}, \tag{213}$$

with $\Delta C_i = C_{i,\gamma} - C_i = \phi_i - C_i$. This defines the *(proper) vertex functions* $\Gamma_{i_1,\cdots,i_n}$ for which

$$\Gamma_{i_1,\cdots,i_n} = \left.\frac{\partial^n \Gamma}{\partial \phi_{i_1}\cdots\partial\phi_{i_n}}\right|_{\gamma=0} = \left.\frac{\partial^{n-1}\gamma_{i_n}(\phi)}{\partial\phi_{i_1}\cdots\partial\phi_{i_{n-1}}}\right|_{\gamma=0}, \tag{214}$$

using Eq.(212) and noting that $\Delta C = 0$ if $\gamma = 0$. Inverting the multidimensional version of Eq(200) (setting $\gamma_1 = 0$) to obtain $\gamma_i(\phi)$ in terms of ΔC_i, and defining *amputated correlation functions*

$$A_{i_1,\cdots,i_n} = \sum_{j_1,\cdots j_n} C^{-1}_{i_1,j_1}\cdots C^{-1}_{n_1,j_n} C_{j_1,\cdots,j_n}, \tag{215}$$

one finds

$$\Gamma_{i_1} = 0, \tag{216}$$
$$\Gamma_{i_1,i_2} = C^{-1}_{i_1,i_2}, \tag{217}$$
$$\Gamma_{i_1,i_2,i_3} = -A_{i_1,i_2,i_3}, \tag{218}$$
$$\Gamma_{i_1,i_2,i_3,i_4} = -A_{i_1,i_2,i_3,i_4} + \sum_{i,j}\left(A_{i_1,i_2,i}C_{i,j}A_{j,i_3,i_4} + A_{i_1,i_3,i}C_{i,j}A_{j,i_2,i_4} + A_{i_1,i_4,i}C_{i,j}A_{j,i_2,i_3}\right), \tag{219}$$
$$\cdots$$

These equations can easily be inverted

$$C_{i_1,i_2} = \Gamma^{-1}_{i_1,i_2}, \tag{220}$$
$$A_{i_1,i_2,i_3} = -\Gamma_{i_1,i_2,i_3}, \tag{221}$$
$$A_{i_1,i_2,i_3,i_4} = -\Gamma_{i_1,i_2,i_3,i_4} + \sum_{i,j}\left(\Gamma_{i_1,i_2,i}C_{i,j}\Gamma_{j,i_3,i_4} + \Gamma_{i_1,i_3,i}C_{i,j}\Gamma_{j,i_2,i_4} + \Gamma_{i_1,i_4,i}C_{i,j}\Gamma_{j,i_2,i_3}\right), \tag{222}$$
$$\cdots$$

Often it is only possible to find (full or perturbed) vertex functions $\Gamma_{i_1,\cdots,i_n}$ by expanding around known vertex functions $\Gamma^0_{i_1,\cdots,i_n}$ for a solvable (reference or unperturbed) system. For example, Eq.(217) can be reformulated in terms of the "self energy" being the difference between perturbed and unperturbed two point vertex functions

$$(\Sigma)_{i_1,i_2} = \Gamma_{i_1,i_2} - \Gamma^0_{i_1,i_2}. \tag{223}$$

This results in

$$C_{i_1,i_2} = C^0_{i_1,i_2} - \sum_{j,k} C^0_{i_1,j}\,(\Sigma)_{j,k}\,C_{k,i_2}, \tag{224}$$

which expresses the full connected two point correlation function C_{i_1,i_2} in terms of an unperturbed $C^0_{i_1,i_2}$. An approximation $(\hat{\Sigma})$ for the self energy can be obtained for example by perturbation theory with respect to the unperturbed reference system. Then a corresponding self-consistent solution $\hat{C}_{i_1,i_2}$ can be found by iteration according to

$$\hat{C}_{i_1,i_2} = C^0_{i_1,i_2} - \sum_{j,k} C^0_{i_1,j} (\hat{\Sigma})_{j,k} C^0_{k,i_2} + \sum_{j,k,l,m} C^0_{i_1,j} (\hat{\Sigma})_{j,k} C^0_{k,l} (\hat{\Sigma})_{l,m} C^0_{m,i_2}). \quad (225)$$

The following two Tables show generating functions for zero field

generating function	derivatives
$Z(\gamma) = \int dx\, e^{-\beta E(x)+\gamma g(x)}$	$\frac{\partial^n}{\partial \gamma^n} Z(\gamma)\Big\vert_{\gamma=0} = Z M_n(g)$
$\ln Z(\gamma) = \ln \int dx\, e^{-\beta E(x)+\gamma g(x)}$	$\frac{\partial^n}{\partial \gamma^n} \ln Z(\gamma)\Big\vert_{\gamma=0} = C_n(g)$
$M(\gamma) = \frac{Z(\gamma)}{Z} = \left\langle e^{\gamma g(x)} \right\rangle_{\mathcal{X}}$	$\frac{\partial^n}{\partial \gamma^n} M(\gamma)\Big\vert_{\gamma=0} = M_n(g)$
$C(\gamma) = \ln M(\gamma) = \ln \left\langle e^{\gamma g(x)} \right\rangle_{\mathcal{X}}$	$\frac{\partial^n}{\partial \gamma^n} C(\gamma)\Big\vert_{\gamma=0} = C_n(g)$

and for nonzero field β or γ_0, respectively

generating function	derivatives
$Z = \int dx\, e^{-\beta E(x)}$	$\frac{\partial^n}{\partial(-\beta')^n} Z(\beta')\Big\vert_{\beta'=\beta} = Z M_n(E)$
$\ln Z = \ln \int dx\, e^{-\beta E(x)}$	$\frac{\partial^n}{\partial(-\beta')^n} \ln Z(\beta')\Big\vert_{\beta'=\beta} = C_n(E)$
$Z(\gamma) = \int dx\, e^{-\beta E(x)+\gamma g(x)}$	$\frac{\partial^n}{\partial\gamma^n} Z(\gamma)\Big\vert_{\gamma=\gamma_0} = Z(\gamma_0) M_{n,\gamma_0}(g)$
$\ln Z(\gamma) = \ln \int dx\, e^{-\beta E(x)+\gamma g(x)}$	$\frac{\partial^n}{\partial\gamma^n} \ln Z(\gamma)\Big\vert_{\gamma=\gamma_0} = C_{n,\gamma_0}(g)$
$M(\gamma) = \frac{Z(\gamma)}{Z} = \left\langle e^{\gamma g(x)} \right\rangle_{\mathcal{X}}$	$\frac{\partial^n}{\partial\gamma^n} M(\gamma)\Big\vert_{\gamma=\gamma_0} = M(\gamma_0) M_{n,\gamma_0}(g)$
$C(\gamma) = \ln M(\gamma) = \ln \left\langle e^{\gamma g(x)} \right\rangle_{\mathcal{X}}$	$\frac{\partial^n}{\partial\gamma^n} C(\gamma)\Big\vert_{\gamma=\gamma_0} = C_{n,\gamma_0}(g)$
$M_{\gamma_0}(\gamma) = \frac{Z(\gamma)}{Z(\gamma_0)} = \left\langle e^{(\gamma-\gamma_0)g(x)} \right\rangle_{\mathcal{X}\vert\gamma_0}$	$\frac{\partial^n}{\partial\gamma^n} M_{\gamma_0}(\gamma)\Big\vert_{\gamma=\gamma_0} = M_{n,\gamma_0}(g)$
$C_{\gamma_0}(\gamma) = \ln M_{\gamma_0}(\gamma) = \ln \left\langle e^{(\gamma-\gamma_0)g(x)} \right\rangle_{\mathcal{X}\vert\gamma_0}$	$\frac{\partial^n}{\partial\gamma^n} C_{\gamma_0}(\gamma)\Big\vert_{\gamma=\gamma_1} = C_{n,\gamma_1}(g)$

A.4. Conditional probabilities and disordered systems

We already discussed that it is often useful to look for conditions under which the energy is easy to specify and to combine the different possible conditions in a second step to obtain the complete probability. Thus, a joint probability $p(x,y)$ can be specified either by a 'joint' (or annealed) energy function $E(x,y)$ with conjugated joint (or annealed) temperature $1/\beta$ or a conditional (or quenched) energy $E(x|y)$ with conjugated conditional (or quenched) temperatures $1/B(y)$ and a mixture energy $E(y)$ with mixture temperature $1/b$. Hence,

$$p(x,y) = \frac{e^{-\beta E(x,y)}}{Z(\mathcal{X},\mathcal{Y})} = p(y)p(x|y) = \frac{e^{-bE(y)}e^{-B(y)E(x|y)}}{Z(\mathcal{Y})Z(\mathcal{X}|y)}, \tag{226}$$

with

$$p(y) = e^{-bE(y)}/Z(\mathcal{Y}), \tag{227}$$

$$p(x|y) = e^{-B(y)E(x|y)}/Z(\mathcal{X}|y), \tag{228}$$

and

$$Z(\mathcal{X},\mathcal{Y}) = \int dx \int dy\, e^{-\beta E(x,y)}, \tag{229}$$

$$Z(\mathcal{Y}) = \int dy\, e^{-bE(y)}, \tag{230}$$

$$Z(\mathcal{X}|y) = \int dx\, e^{-B(y)E(x|y)}. \tag{231}$$

One may remark, that choosing $-\beta E(x,y) = -bE(y) - B(y)E(x|y)$ produces a joint probability

$$p'(x,y) = \frac{e^{-\beta E(x,y)}}{Z(\mathcal{X},\mathcal{Y})} = \frac{e^{-bE(y)-B(y)E(x|y)}}{\int dy\, e^{-bE(y)} Z(\mathcal{X}|y)}, \tag{232}$$

which is different from $p(x,y)$ of Eq.(226).

If interested only in variable x one integrates (marginalizes) over y. Working with joint energies this gives

$$p(x) = \int dy\, p(x,y) = \frac{\int dy\, e^{-\beta E(x,y)}}{Z(\mathcal{X},\mathcal{Y})} = \frac{e^{-\beta E(x)}}{Z(\mathcal{X})}, \tag{233}$$

whereas working with conditional energies yields

$$p(x) = \int dy\, p(x,y) = \int dy\, \frac{e^{-\ bE(y)} e^{-B(y)E(x|y)}}{Z(\mathcal{Y})Z(\mathcal{X}|y)} = \int dy\, p(y) \frac{e^{-B(y)E(x|y)}}{Z(\mathcal{X}|y)}. \tag{234}$$

In the formulation with joint probabilities $E(x,y)$ is the canonical variable. Its averages can be obtained by differentiation with respect to β

$$E_{ann} = E(\mathcal{X}) = \langle E(x)\rangle_{\mathcal{X}} = \langle E(x,y)\rangle_{\mathcal{X},\mathcal{Y}} = -\frac{\partial}{\partial\beta} \ln Z(\mathcal{X}). \tag{235}$$

In the formulation with conditional probabilities the expectation of $E(x|y)$ can be obtained by differentiation with respect to $B(y)$

$$E_{quen} = E(\mathcal{X}|\mathcal{Y}) = \langle E(\mathcal{X}|y)\rangle_{\mathcal{Y}} = \langle E(x|y)\rangle_{\mathcal{X},\mathcal{Y}} = -\left\langle \frac{\partial}{\partial B(y)} \ln Z(\mathcal{X}|y) \right\rangle_{\mathcal{Y}}. \tag{236}$$

Even for y–independent $B(y) = B$ the averaging of $\ln Z(\mathcal{X}|y)$ remains

$$\langle E(x|y)\rangle_{\mathcal{X},\mathcal{Y}} = -\frac{\partial}{\partial B} \langle \ln Z(\mathcal{X}|y)\rangle_{\mathcal{Y}}. \tag{237}$$

For y–independent normalization $Z(\mathcal{X}|y) = Z(\mathcal{X})$ both approaches are equivalent and $\langle \ln Z(\mathcal{X}|y)\rangle_{\mathcal{Y}} = \ln Z(\mathcal{X})$. In general, however, the expectations E_{ann} and E_{quen} are different. Also, despite of the equality (226), the exponential families defined by varying the parameters β or B (or $B(y)$) are not the same. The conditional temperatures or fields $B(y)$ do not influence the distribution $p(y)$, while the joint temperature $1/\beta$ does.

In practice, for example, it may take some time after changing temperature or an external field until a stationary distribution $p(x, y)$ is reached. Assume the dynamic of y being much slower than that of x (e.g. lower energy barriers for x and higher energy barriers for y). In a magnetic substance x may stand for fast adapting local spins and y for very slowly moving impurities. Then changing the physical temperature will on short time scales (or low temperatures) only change the distribution of the fast adapting spins x, while the slow impurities y remain quenched. Then the relevant physical temperature is the conditioned or quenched field $B(x)$. On a much longer time scale (or at high enough temperatures, i.e., in an 'annealed system') also the impurities will approach an equilibrium distribution. Then the physical temperature is a joint or annealed field β. In addition, ensemble averages for variables which reach a stationary distribution on short enough time scales are often measured as time averages under a (quasi–)ergodic dynamic. In contrast, averages over slow variables can in practice not be obtained as time averages and must be realized as ensemble averages.

Eq.(237) requires the calculation of a partition sum $Z(\mathcal{Y}, x)$ for every x. This is possible for a small number of different x values or if the x–dependence can be calculated analytically and the average over $\mathcal{X}$ be performed. In general, however, Eq.(237) is rather difficult to solve. One possibility to proceed is using the identity

$$\ln Z = \lim_{n\to 0} \frac{Z^n - 1}{n}, \tag{238}$$

which is verified by expanding

$$Z^n = e^{n \ln Z} = 1 + n \ln Z + \sum_{k=2} \frac{n^k}{k!} (\ln Z)^k. \tag{239}$$

Typically, the average over Z^n for integer n is easier to perform than over $\ln Z$. Because Z^n describes a system with n independent 'replicas' of the same system this approach is known under the name replica method [46,18,27,53]. Performing the average over x, however, results usually in a coupling between the different replicas. Also one has to be careful, because calculating Z^n for integer n does not uniquely determine the analytic continuation to $n \to 0$. For non–interacting disordered systems a supersymmetric approach (expressing the two–point correlation function as a product of two gaussian integrals, one over commuting variables and another over anticommuting (Grassmann) variables) can avoid such difficulties of the replica approach ([17,49]).

B. Bayesian decision theory

B.1. Basic definitions

B.1.1. The basic model

Consider the following scenario[73]. We assume that we can prepare a specific situation $x \in \mathcal{X}$ and measure outcome $y \in \mathcal{Y}$. Furthermore, we assume that the probability $p(y|x,h)$ of outcome y is determined by x and additional variables h which we cannot observe directly. These additional variables h will be called collectively 'state of Nature'. Furthermore, we assume all knowledge f we have accumulated about Nature in the past has been combined in form of a probability density $p(h|f)$ over the possible states of Nature $h \in \mathcal{H}$. The aim of learning is to update our knowledge about Nature $p(h|f) \rightarrow p(h|f'(D,f))$ as more data D arrive under the assumption that the underlying 'true state of Nature' h_N producing the data does not change.

Hence, to define the basic model formally we split the random variables of interest into the two groups of

1. *hidden* (not directly measurable) variables $h \in \mathcal{H}$ (*model states*, *possible state of Nature*), assuming the *true state of Nature* h_N is in $\mathcal{H}$, and of
2. *visible* (directly measurable) variables consisting of (potential) *data* $\{x, y\}$ and *state of knowledge* f, where
   a. the vector $x \in \mathcal{X}$ collects all *independent variables* (independent of h, may also be called questions, measurement devices, conditions/situations of measurement, measured quantities, observables),
   b. the vector $y \in \mathcal{Y}$ encompasses all *dependent variables* (depending on h, may also be called answers, measurement results, responses, observed values), and
   c. the state of knowledge $f \in \mathcal{F}$ includes all *determining variables* (determining h, i.e., parameterizing the probability $p(h)$ of the h–producing process and describing thus the situation under study).

Thus, a state of Nature or model state h is described by specifying its data generation densities (x–conditional y–likelihoods of h) $p(y|x,h)$. All together, the joint probability of the basic model factorizes according to

$$p(x,y,h|f) = p(h|f)\,p(x|h)\,p(y|x,h) = p(h|f)\,p(x)\,p(y|x,h). \tag{240}$$

The variables x, y may be vectors of i.i.d. sampled (vector) variables. Repeated independent sampling under constant h, i.e.,

$$p(x) = \prod_i^n p(x_i), \quad p(y|x,h) = \prod_i p(y_i|x_i,h), \tag{241}$$

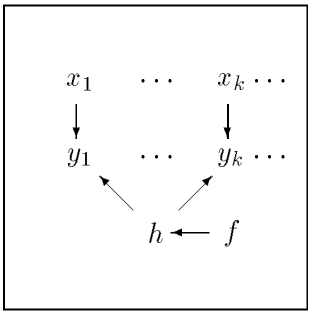

Figure 9. Graphical representation of a probabilistic model factorizing according to $p(x, y, h|f) = p(h|f) \prod_k p(x_k)p(y_k|x_k, h)$.

where x_i can contain (components of) x_0, gives for the example of a discrete set of $x_i \in \mathcal{X}$

$$p(x, y, h) = p(h|f) \prod_i p(x_i)\, p(y_i|x_i, h). \tag{242}$$

Fig.9 shows a graphical representation of that model as a directed acyclic graph [54,41,33,59].

B.1.2. Learning: predictive and posterior density

By *learning* we mean the change in state of knowledge $f \to f'(D, f)$ as new data D arrive. We will distinguish potentially interesting and actually known data:

1. *relevant* or *test* data $D_R = (x_R, y_R)$, $x_R \in \mathcal{X}_R$, $y_R \in \mathcal{Y}_R$ correspond to potential (future) application situations of interest, being thus the data we are actually interested in and which we want to predict, and
2. *available* data D contribute to our *state of knowledge* about Nature. For practical purposes those may be further divided in
   a. *training data* $D_T = \{D_{T,i}|1 \le i \le n\} = \{(x_i, y_i)|1 \le i \le n\} = (x_T, y_T)$, being an empirical sample of D_R, i.e., a finite number of pairs (x_i, y_i) drawn i.i.d. according to $p(x_i)p(y_i|x_i, h_N)$ for relevant $x_i \in X_R$ under the true state of Nature h_N, and
   b. *prior data* $D_0 = \{(x_0, y_0), f, S\}$ collecting all other available knowledge (a priori information) not contained in the training data. Prior data can appear as
      i. *measured prior*, corresponding to measured data (x_0, y_0) not considered as training data, as
      ii. *generative prior* f, (*preparation control*) corresponding to knowledge about the (probabilistic) preparation process which generates the true state of Nature h_N, or as

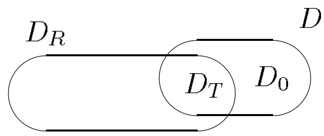

Figure 10. Typical relation between relevant data D_R, training data $D_T \subset D_R$, prior data D_0 and available data $D = D_T \cup D_0$.

iii. *structural prior* S (*model control*) refering to all knowledge concerning the specified dependency structure of the model variables.

Fig.10 summarizes the relations between the different data types.

Being interested in the relevant data the aim of learning is to find the *predictive density* $p(y_R|x_R, f'(D, f))$, or more shortly, $p(y_R|x_R, D)$, after receiving training and prior data. Inserting the hidden variables h the predictive density becomes

$$p(y_R|x_R, f')p(y_R|x_R, D) = \int dh\, p(h|f(D))p(y_R|x_R, h). \tag{243}$$

Thus, the space $\mathcal{F}$ of possible states of knowledge is the convex hull of the the space $\mathcal{H}$ of possible states of Nature. The essential ingredient to be calculated in Eq.(243) is the *posterior density* $p(h|f'(D, f))$, or more shortly $p(h|D)$, which can be obtained by obtained by inverting the model (242) using Bayes' rule

$$p(h|D) = \frac{p(y_D|x_D, h)p(h)}{p(y_D|x_D)}. \tag{244}$$

B.1.3. The risk functional

Next we consider a set of possible actions $a(x) \in \mathcal{A}$ from which we can choose in situation x before having seen y. The action $a(x)$ can for example be the value we expect for y (regression) or a complete density $p(y|x, a)$ (density estimation) we expect for y under x. Furthermore, assume we suffer loss $l(x, y, a(x))$ if y appears after we have chosen $a(x)$. Common loss functions are for example log-loss $l(x, y, a) = -\ln p(y|x, a)$ for density estimation or mean square loss $(y - a(x))^2$, absolute loss $|y - a(x)|$, or $\delta(y - a(x))$ for regression [5].

The decision model we consider has a graphical representation shown in Fig.11, where actions and loss are deterministic variables

$$p(l|x, y, a) = \delta(l - l(x, y, a)), \tag{245}$$

and

$$p(a|x) = \delta(a - a(x)). \tag{246}$$

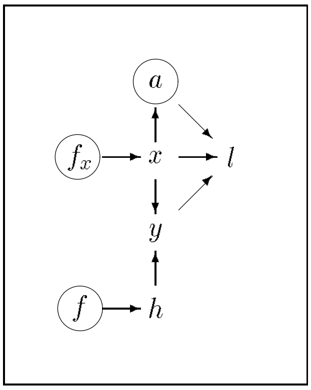

Figure 11. The setting of Bayesian decision theory in a graphical model. Circles indicate known variables. In this figure the variable f_x determining the probability $p(x)$ is shown explicitly. f_x is implicit in the other figures.

Decision theory aims in minimizing a functional of the loss posterior $p(l|a,f)$. The most common functional chosen to be minimized is the *expected risk* (expected loss)

$$\begin{aligned} r(a,f) &= \int dl\; l(x,y,a)\, p(l|a,f) \\ &= \int dx\; \int dy\, p(x)p(y|x,f)\, l(x,y,a) \\ &= \int dh \int dx \int dy\, p(x)\, p(y|x,h)\, p(h|f)\, l(x,y,a). \end{aligned} \tag{247}$$

B.2. Interpretation of priors

B.2.1. Measured and factorial priors

The state of knowledge before evidence has been received for data D appears as visible variable. This visible variable, however, must also be based on some information which can be considered being data. One may wish, therefore, to express a state of knowledge by giving explicitly the data it is based on. To do this, one has to include prior data in form of a measured prior (x_0, y_0). Such a measured prior represents a situation where value y_0 have been found for h as result of a (probabilistic) measurement of property x_0 (e.g., smoothness). Thus, model (242) becomes

$$\begin{aligned} p(x,y,h) &= p(h|f)p(x_0)\, p(y_0|x_0,h) \prod_{i=1} p(x_i)\, p(y_i|x_i,h) \\ &= p(h|f) \prod_{i=0} p(x_i)\, p(y_i|x_i,h). \end{aligned} \tag{248}$$

Its graphical representation is shown in Fig.12.

Even, however, if measured priors are included the variable f (characterizing now a lower level generative prior) remains in the model. The question therefore arises what kind of f should be chosen as a natural starting point of learning. Consider therefore (as lower level generative prior) a *factorial prior*, for which $p(h|f)$ factorizes with respect to relevant data,

$$p(h|f) = \prod_{k=1} p(h_k|f_k). \tag{249}$$

The model, shown in Fig.13, becomes

$$p(x,y,h|f(\text{factorial})) = p(x_0)p(y_0|x_0,h)\prod_{k=1} p(h_k|f_k)p(x_k)p(y_k|x_k,h_k). \tag{250}$$

Without prior term $p(y_0|x_0,h)$ this corresponds to a diagram where only variables with the same index i are connected, which means that receiving data for x_i would not allow any generalization to relevant data $D_{j\neq i}$

$$p(y_j|x_j, f(D_{i\neq j})) = p(y_j|x_j, f(\text{factorial})). \tag{251}$$

A factorial prior is therefore a natural choice for a formal starting point of learning as it contains no information which allows generalization from training to non–training data. With respect to a factorial prior as lower level prior it is therefore essential for generalization that additional higher level a priori information $D_0 = (y_0, x_0)$ depends on all h_i.

Interestingly, the asymptotic end points of learning, i.e., the pure model states h, also represent factorial states,

$$p(y|x,h) = \prod_i p(y_i|x_i,h), \tag{252}$$

because the variables (x_i, y_i) for different i are independent conditioned on h. Hence, in this picture learning starts and ends asymptotically in a factorial state of knowledge. The distance from a factorial state could be measured by the (x–averaged) *mutual information* [40,15]

$$M(f) = \int dx \int dy\, p(y|x,f) \frac{p(y|x,f)}{\prod_i p(y_i|x_i,f)}. \tag{253}$$

Thus, starting at zero the mutual information would be expected to increase at the beginning of learning and to approach finally zero again asymptotically. The final factorial state could again be starting point of learning under a new, finer model $\mathcal{H}'$.

B.2.2. Gaussian likelihoods

A model state h is a shorthand notation for a parameter vector ξ specifying the data generating densities $p(y|x,h) = p(y|x,\xi)$. In this paper we mainly consider gaussian regression problems for which relevant and training data are

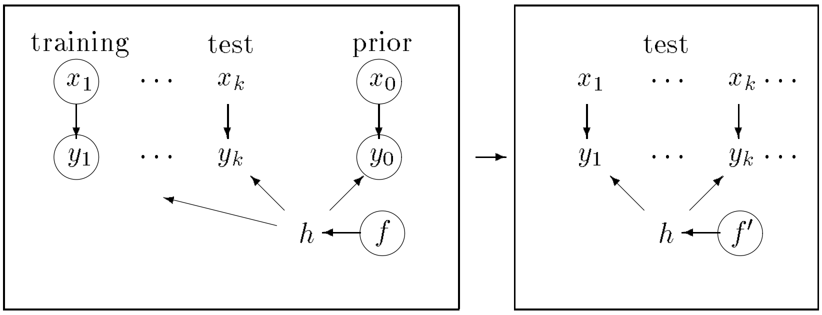

Figure 12. Graphical representation of a model with measured and with generative prior $p(x, y, h|f) = p(h|f)\ p(x_0)p(y_0|x_0, h) \prod_{k=1} p(x_k)p(y_k|x_k, h)$. Hereby x_0, y_0 may also factorize into independent components. Circles indicate measured variables, i.e., the available data D. The predictive density for relevant or test data D_R reads $p(y_R|x_R, f'(D, f)) = \int dh\, p(y_R|x_R, h)$ $p(h|D)$ and requires calculation of the posterior density $p(h|D) \propto p(y_D|x_D, h)p(h)$. The figure on the right shows the situation after learning where data D has been used to obtain the new state of knowledge $f'(f, D)$.

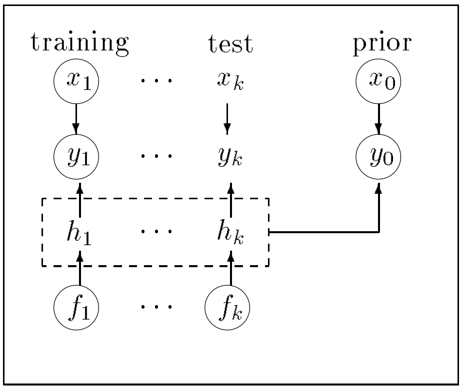

Figure 13. Graphical representation of a factorial prior (with respect to relevant or test data) $p(x, y, h|f(\text{factorial})) = p(x_0)p(y_0|x_0, h) \prod_{k=1} p(h_k|f_k)p(x_k)p(y_k|x_k, h_k)$. Without additional higher level prior D_0 with response y_0 depending on h_k and h_l no generalization is possible from training data $D_{l\neq k}$ to test data D_k for that k and l.

produced by gaussians with mean specified by h and h– and x–independent variance. For example, for one–dimensional y

$$p(y|x, h) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-h(x))^2}{2\sigma^2}}. \tag{254}$$

Hence, in this case h is parameterized by the regression function $\xi(x) = h(x)$. In general density estimation problems one use a parameter function $\xi(x, y) = p(y|x, h)$ under the additional normalization constraint $\int dy\, \xi(y, x) = \int dy\, p(y|x, h) = 1$, $\forall x, h$. Thus, in this case the y–likelihood of h is not gaussian in ξ.

The integration $\int dh$ necessary to obtain the predictive density stands for the integration $\int \prod_m d\xi_m$ over all components ξ_m of parameter vector ξ. In case h is parameterized by a regression function $\xi_m \to h(x)$ as in Eq.(254) the integral reads

$$\int \prod_m d\xi_m \to \int \prod_x dh(x). \tag{255}$$

As far as well defined, this becomes for continuous x variable a functional integral. Similarly, for general density estimation problems

$$\int \prod_m d\xi_m \to \int \prod_x \delta(\int dy'\, p(y'|x,h) - 1) \prod_y dp(y|x,h). \tag{256}$$

Prior data have to depend on all $h(x)$ to allow for generalization. Analogously to the gaussian y–likelihood of (257), we will in this paper mainly consider gaussian prior information, e.g., for one–dimensional regression problems

$$p(y_0|x_0,h) = (2\pi)^{-\frac{d}{2}} (\det K)^{\frac{1}{2}} e^{-\frac{1}{2}\left\langle h-y_0 \middle| K(x_0) \middle| h-y_0 \right\rangle}, \tag{257}$$

for symmetric, positive (semi–)definite K and

$$\left\langle h - y_0 \middle| K(x_0) \middle| h - y_0 \right\rangle = \sum_{x=1}^{d} \sum_{x'=1}^{d} (h(x) - y_0(x)) K(x,x') (h(x') - y_0(x')). \tag{258}$$

An infinite normalization factor in Eq.(257) appearing for $d \to \infty$ will cancel when calculating expectations. We will use therefore the convenient integral notation $\sum_x \Delta_x \to \int dx$ in the following. The variable x_0 determines the kind of measurement for which y_0 is regarded as output. In the gaussian case (257) x_0 defines the operator K and thus the covariance K^{-1}. To measure smoothness, for example, one may choose the laplacian $K = \Delta$. The choice $K(x) = |x><x|$ yields the standard gaussian training or test data of Eq.(regression).

B.2.3. Generative priors

If the process generating h_N is under explicit control it is natural to model prior data as generative prior. For example, consider a situation where h is produced by a gaussian density

$$p(f_0|x_0,y^0) = (2\pi)^{-\frac{d}{2}} (\det K)^{\frac{1}{2}} e^{-\frac{1}{2}\left\langle h-y_0 \middle| K(x_0) \middle| h-y_0 \right\rangle}, \tag{259}$$

with mean y_0 and covariance K^{-1} under control. Then the corresponding generative prior model is shown in Fig.14.

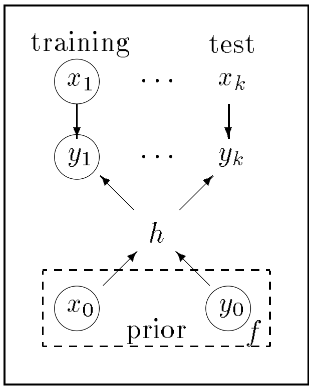

Figure 14. Graphical representation of a generative prior model $p(x,y,h|D_0) = p(h|x_0,y_0)$ $\prod_{k=1} p(x_k)p(y_k|x_k,h)$. Here, a priori information enters as posterior density of a h–generating process. The variables $\theta_h = (x_0, y_0)$ are parameters of the state generating process. For gaussians, for example, they may determine mean and covariance structure. The posterior density of h under data D becomes $p(h|D) \propto p(y_T|x_T,h)p(h|D_0)$.

B.2.4. More complex models

The basic model of Section (B.1.1) is completely general. It can however be useful to implement additional structure. *Input noise*, for example, corresponds to a decomposition

$$p(y|x,h) = \int d\theta_x p(y|\theta_x,h)p(\theta_x|x). \tag{260}$$

For a finite number of discrete θ_x and gaussian components $p(y|\theta_x,h)$ this constitutes for $p(y|x,h)$ a mixture of gaussians with varying mean. Similarly, *output noise*

$$p(y|x,h) = \int d\theta_y p(y|\theta_y)p(\theta_y|x,h), \tag{261}$$

may be used to construct gaussian mixtures with varying covariances. Analogously, *generation noise*

$$p(h|f) = \int d\theta_h p(h|\theta_h)p(\theta_h|f), \tag{262}$$

can be used to obtain for example a mixture of gaussians with varying mean and covariance.

The variables $\theta = (\theta_x, \theta_y, \theta_h)$ represent additional hidden variables of the model. Restricting for convenience the variables denoted by h to that hidden variables which determine the relevant data (relevant state of Nature), an (extended/complete) state of nature $\tilde{h}$ is given by specifying $\tilde{h} = (h, \theta)$. A graphical representation of a model with additional noise variables $\theta = (\theta_x, \theta_y, \theta_h)$ is shown in Fig.15.

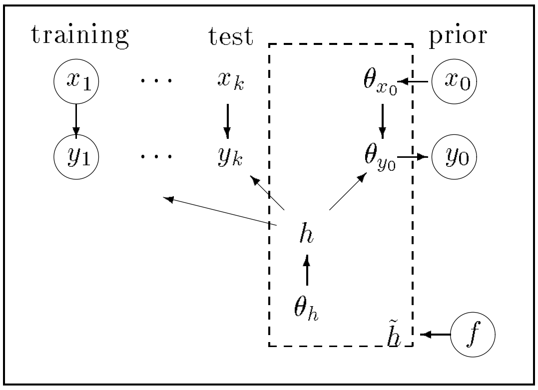

Figure 15. Graphical representation of a model including besides h additional hidden variables $p(x, y, \tilde{h}) = p(x, y, h, \theta_{x_0}, \theta_{y_0}, \theta_h) = p(\theta_h)p(h|\theta_h)\; p(x_0)p(y_0|\theta_{x_0})\; p(\theta_{y_0}|\theta_{x_0}, h)p(\theta_{x_0}|x_0)$ $\prod_{k=1} p(x_k)p(y_k|x_k, h)$. The additional hidden variables $\theta = (\theta_{x_0}, \theta_{y_0}, \theta_h)$ parameterize input, output, and generation noise. The posterior density of h requires marginalization (integration) over θ and becomes $p(h|D) \propto \int d\theta_{x_0}\, d\theta_{y_0}\, d\theta_h\; p(y_T|x_T, h)\; p(y_0|\theta_{y_0})\; p(\theta_{y_0}|\theta_{x_0}, h)\; p(\theta_{x_0}|x_0)$ $p(h|\theta_{f_0})\; p(\theta_{f_0})$.

B.2.5. Structural priors and a posteriori control

We have up to now discussed how prior data can enter a model in the form of empirically measured data or as generative prior controlling the preparation of f. A third possibility consists in the direct specification of the necessary dependency relations between relevant and training data within the model.

The empirical validity a a priori information implemented by a specific model is based the possibility to control the required dependency structure between relevant and training data (model control). As those dependencies have to be controled at the time of testing (which is usually after having received training data) the validity of a priori information can be ensured by *a posteriori control*. As a trivial example consider a bound $y_R < c$ which can be enforced by ignoring values larger than c at the time of testing.

Thus, it is useful to note that a device producing y with probability density $p(y|x, h)$ has a 'passive' interpretation as *measurement device* or an 'active' interpretation as *control device*. In the passive interpretation the device is said to measure the property x of Nature h, in the active interpretation y is said to be produced or controled by preparing x, but modified by unknown parameters h. The active interpretation is evident if the probability density of y is determined by a function g of x and y only

$$p(y|x, h) = g(x, y), \tag{263}$$

which in the extreme case may even be deterministic

$$p(y|x, h) = \delta(y - g(x)). \tag{264}$$

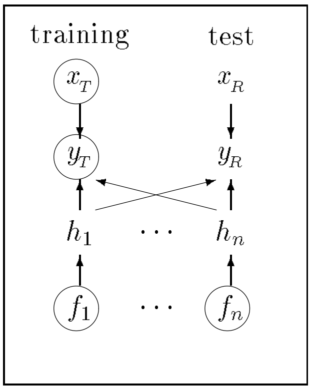

Figure 16. Graphical representation of a structural prior model $p(x, y, h) = p(x_T)\ p(x_R)\ p(y_T|x_T, h)\ p(y_R|x_R, h) \prod_n p(h_n|f_n)$ with posterior $p(h|D) \propto p(h)p(y_T|x_T, h)$.

Next, consider a control device

$$p(y|x, h) = g(x, y, \xi) \tag{265}$$

depending on some parameters ξ which are *unknown but fixed for all x*. Thus, the parameters ξ represent the unknown state of Nature h and the control device can be said to be a (indirect) measurement device for ξ. Such a situation is shown in Fig.16. Because hereby prior information is implemented by choosing a model with specific structure this may be called a structural prior.

The number of parameters ξ may be very large. Consider, for example, a control device producing y according to a gaussian density with known variance and mean similar but probably not equal to $t(x)$. Assuming now the true but unknown mean $\xi(x) = h(x)$ is an empirical realization of a gaussian process with mean $t(x)$ and covariance K^{-1} would result in an quadratic error term $(1/2) < h - t|K|h - t >$. Thus, in this interpretation the template function $t(x)$ characterizes an average control device. We have already discussed that an approximative control device $t(x)$ is often provided by human experts. For example they may specify what constituents define an image to be a face. Looking for $h(x)$ in the neighborhood of $t(x)$ approximate control devices $t(x)$ can than be refined by training data.

Summarizing, a measurement device can be seen as control device with unknown but constant parameters with a priori information specifying the amount of control or knowledge we have about that device.

B.3. Quantum Mechanics

B.3.1. Density operators

The formalism can also be applied to quantum mechanical problems and can be used to solve inverse quantum mechanical problems, i.e., problems where empirical, (e.g., scattering) data are used to determine the Hamiltonian of a system. In quantum mechanics a system is specified by a density operator ρ describing the state of the system. An observable x is represented by a hermitian operator $\mathbf{O}_x$[3] and its eigenvalues are the possible measurements results y. As a measurement in quantum mechanics changes ρ, repeated measurement under the same ρ requires a new preparation of ρ before each measurement.

In particular, the probability of measuring value y for observable x represented by an hermitian operator $\mathbf{O}_x$ under density operator ρ is given by

$$p(y|x,h) = p(y|\mathbf{O}_x,\rho) = \mathrm{Tr}(\mathbf{P}_y\rho) = \sum_j^{n_y} < y_j|\,\rho\,|y_j >, \tag{266}$$

where and

$$\mathbf{P}_y = \sum_j^{n_y} |y_j >< y_j| \tag{267}$$

is the projector (with $|y_j >< y_j|$ denoting the dyadic product) into the subspace of (orthonormalized) eigenfunctions y_j of the (hermitian) operator $\mathbf{O}_x$ with eigenvalue y

$$\mathbf{O}_x|y_j >= y|y_j >, \quad \langle y_j|y_i'\rangle = \delta_{ji}\delta(y-y'). \tag{268}$$

The indices distinguish n_y different eigenvectors with equal eigenvalues y. For nondegenerate eigenvalues $n_y = 1$. The density operator characterizing the system is a hermitian, positive definite operator with eigenfunction decomposition (the sum to be replaced by an integral for non discrete cases)

$$\rho = \sum_i p(i)\,|\varphi_i >< \varphi_i| \tag{269}$$

with $0 \le p_i \le 1$ and $\sum_i p(i) = 1$ and orthonormal $\langle\varphi_i|\varphi_j\rangle = \delta_{ij}$. In case the φ_i are the eigenstates y_i of $\mathbf{O}_x$ the $p(i)$ become $p(y|\mathbf{O}_x,\rho)$. A density operator which is a projector,

$$\rho^2 = \rho = |\varphi >< \varphi|, \tag{270}$$

consists of only one term and is called a pure (quantum mechanical) state. For a pure state

$$p(y|x,h) = \mathrm{Tr}\Big(\sum_j^{n_y} |y_j >< y_j|\varphi >< \varphi|\Big)$$

$$= \sum_j^{n_y} \langle\varphi|y_j\rangle\,\langle y_j|\varphi\rangle = \sum_j^{n_y} |\,\langle y_j|\varphi\rangle\,|^2 = \sum_j^{n_y} |\varphi(y_j)|^2. \tag{271}$$

[3] We used $\mathbf{O}_x$ instead of $\mathbf{x}$ to denote the operator representing a general observable, because $\mathbf{x}$ is used in quantum mechanics to denote the multiplication operator in coordinate space.

Pure states already show all (non classical) quantum phenomena while statistical mixtures with $\rho^2 \ne \rho$ just add a classical (non quantum mechanical) averaging to quantum mechanical systems.

Consider a system with $\rho(t_0)$ prepared at time t_0. To calculate the density operator $\rho(t)$ of that system at a later time $t > t_0$ of measurement it is necessary to study the time dependence of the system. The time development of a quantum mechanical system is given by a unitary evolution operator $\mathbf{U}$ (unitary means $\mathbf{U}^{-1}$ $= \mathbf{U}^\dagger$ with $\mathbf{U}^\dagger$ denoting the hermitian conjugate of $\mathbf{U}$)

$$|\varphi(t)> = \mathbf{U}(t,t_0)|\varphi(t_0)>, \tag{272}$$

leading for density operators to

$$\rho(t) = \mathbf{U}(t,t_0)\rho(t_0)\mathbf{U}^\dagger(t,t_0). \tag{273}$$

The evolution operator is usually expressed by a Hamiltonian operator $\mathbf{H}$

$$U(t,t_0) = e^{-i(t-t_0)\mathbf{H}}, \tag{274}$$

for time–independent Hamiltonian $\mathbf{H}$ (and disregarding a factor $\hbar$) or (at least formally) by

$$U(t,t_0) = Te^{-i\int_{t_0}^{t} \mathbf{H}(t)\,dt}, \tag{275}$$

for time–dependent Hamiltonian $\mathbf{H}(t)$ with T denoting the time–ordering operator (see for example [31,51,77]).

B.3.2. Quantum statistics

Stationary density operators

$$\rho = \sum_{ij} p_{ij} |E_{ij}><E_{ij}| \tag{276}$$

with $\rho(t) = \rho(t_0)$ are built from (orthonormalized) eigenstates $|E_{ij}>$ of the Hamiltonian

$$\mathbf{H}\,|E_{ij}> = E_i\,|E_{ij}>, \tag{277}$$

the index j distinguishing eigenstates with equal eigenvalue. For example, a canonical ensemble at temperature $1/\beta$ for a Hamiltonian $\mathbf{H}(\xi)$ depending on parameters ξ is given by the density operator [2]

$$\rho(\mathbf{H}(\xi)) = \frac{e^{-\beta\mathbf{H}(\xi)}}{Z_\xi} = \frac{e^{-\beta\mathbf{H}(\xi)}}{\mathrm{Tr}\, e^{-\beta\mathbf{H}(\xi)}} = \frac{\sum_i e^{-\beta E_i(\xi)} \sum_j^{n_{E_i}} |E_{ij}><E_{ij}|}{\sum_i \sum_j^{n_{E_i}} e^{-\beta E_i(\xi)}} \tag{278}$$

yielding

$$p(y|x,h) = p(y|\mathbf{O}_x,\rho(\mathbf{H}(\xi))) = \frac{\sum_j^{n_y} \left\langle y_j \left| e^{-\beta\mathbf{H}(\xi)} \right| y_j \right\rangle}{\sum_y \sum_j^{n_y} \left\langle y_j \left| e^{-\beta\mathbf{H}(\xi)} \right| y_j \right\rangle}. \tag{279}$$

Here ρ has been expressed by eigenfunctions of $\mathbf{H}$ by expanding the exponential and inserting the eigenfunction expansion

$$\mathbf{H} = \sum_i \sum_j^{n_{E_i}} E_i |E_{ij} >< E_{ij}|. \tag{280}$$

(Replace the sum by an integral if $\mathbf{H}$ has a continuous spectrum.)

A Bayesian approach now uses the likelihood (279) to update a given prior density $p(\xi)$ to a new posterior density $p(\xi|D)$ after new data D became available. Because the measurement of a quantum mechanical system changes ρ, repeated data for the same ρ requires the repeated preparation of the canonical ensemble before each measurement. If learning about ξ is intended the canonical ensembles must hereby correspond to fixed (but unknown) ξ. This can simply mean waiting long enough between two measurements until a given system with fixed but unknown hamiltonian (described by ξ) is again in thermal equilibrium.

B.3.3. Quantum mechanical scattering

As a second example we prepare a pure state $|z(t_0)\!>$ at time t_0 with $\mathbf{z}|z\!>= z|z\!>$. This corresponds to the density operator at time t

$$\rho(t) = |z(t)\!><\!z(t)| = \mathbf{U}(t,t_0)|z(t_0)\!><\!z(t_0)|\mathbf{U}^\dagger(t,t_0). \tag{281}$$

Measuring then at time t a non degenerated eigenvalue y of observable $\mathbf{y}$ has probability

$$\begin{aligned} p(y|x,h) = p(y|\mathbf{O}_x,\rho(t)) &= |z(t,y)|^2 =< y(t)|z(t)\!><\!z(t)|y(t) > \\ &= |\,\langle y(t)|z(t)\rangle\,|^2 = |\,\langle y(t)|\mathbf{U}(t,t_0)z(t_0)\rangle\,|^2. \end{aligned} \tag{282}$$

In scattering theory one takes the limit $t \to \infty$ and $t_0 \to -\infty$ and chooses the asymptotic states $\lim_{t\to\infty} y(t)$ and $\lim_{t_0\to-\infty} z(t_0)$ to converge (in the weak sense) to ('free) states y^0 or z^0, respectively, fulfilling

$$\mathbf{H}_{in}|z_j^0 >= z^0|z_j^0 >, \quad \mathbf{H}_{out}|y_j^0 >= y^0|y_j^0 > . \tag{283}$$

(We will skip the degeneration index j assuming there exist non–degenerate (i.e., unique) states $|z^0 >$ and $< y^0|$. In case the eigenvalues z^0 or y^0 are degenerated the states can be made unique by measuring additional commuting observables commuting also with $\mathbf{H}_{in}$ or $\mathbf{H}_{out}$, respectively. For non-unique states, see below.) One obtains

$$p(y|x,h) = \lim_{t\to\infty} p(y|\mathbf{O}_x,\rho(t)) = \lim_{t\to\infty} |\,\langle y(t)|z(t)\rangle\,|^2 = \lim_{\substack{t\to\infty \\ t_0\to-\infty}} |\,\langle y(t)|\mathbf{U}(t,t_0)z(t_0)\rangle\,|^2, \tag{284}$$

or, inserting the free in (initial, preparation) states $|z^0>$ and out (final, measured) states $<y^0|$,

$$\lim_{\substack{t\to\infty\\ t_0\to-\infty}} p(y|\mathbf{O}_x,\rho(t,z(t_0))) = \left|\left\langle y^0(0)\middle|\mathbf{S}z^0(0)\right\rangle\right|^2 \tag{285}$$

which defines the scattering operator

$$\mathbf{S} = \lim_{\substack{t\to\infty\\ t_0\to-\infty}} \mathbf{U}^\dagger_{out}(t,0)\mathbf{U}(t,t_0)\mathbf{U}_{in}(t_0,0), \tag{286}$$

with (for time–independent $\mathbf{H}_{in}$, $\mathbf{H}_{out}$)

$$\mathbf{U}_{in}(t,t_0) = e^{-i(t-t_0)\mathbf{H}_{in}}, \quad \mathbf{U}_{out}(t,t_0) = e^{-i(t-t_0)\mathbf{H}_{out}}. \tag{287}$$

Often only a partial measurement is performed which does not allow to identify a unique final state y^0 (or y). Then there is a set A of y^0 which can not be distinguished by the measurement. Also, often the preparation is not a pure state but a mixture of states $z^0 \in B$ (or $z \in B$, possibly also with varying preparation observables $\mathbf{z}$ or $\mathbf{H}_{in}$) with probability $p(z^0)$. In that case one has to sum over out states in A and average over in states in B

$$\sum_{y\in A}\sum_{h\in B} p(y|x,h)p(h(z^0))$$

$$= \sum_{y^0\in A}\sum_{z^0\in B} p(z^0)p(y|\mathbf{O}_x,\rho) = \sum_{y^0\in A}\sum_{z^0\in B} p(z^0)\left|\left\langle y^0(0)\middle|\mathbf{S}(\xi)z^0(0)\right\rangle\right|^2. \tag{288}$$

Eq.288 links quantum mechanics to the Bayesian framework and allows to determine a system Hamiltonian $\mathbf{H}(\xi)$ from scattering experiments given a prior $p(\xi)$ over parameters ξ.

B.3.4. Disordered systems

A general density operator ρ may be constructed as a mixture of components $\rho(\mathbf{H}(\xi))$ which are stationary with respect different $\mathbf{H}(\xi)$. Consider, for example, a situation where the preparation of a system with stationary density with respect to a constant (but possibly unknown) $\mathbf{H}(\xi)$ is not possible before each measurement. Assuming that at least a constant (but possibly unknown) mixture can be prepared, density operators representing states of Nature $\rho(h)$ have the form

$$\rho(h) = \int d\vartheta\, p(\vartheta|h)\,\rho(\mathbf{H}(\vartheta)), \tag{289}$$

so that (see Section B.2.4)

$$p(y|x,h) = \int d\vartheta\, p(\vartheta|h)\, p(y|x,h,\vartheta) = \int d\vartheta\, p(\vartheta|h)\, p(y|\mathbf{O}_x,\rho(\mathbf{H}(\vartheta)). \tag{290}$$

There are many cases where it is easier to specify a system in a conditional (disordered) form (289) than to give directly the joint density $p(y,\vartheta|x,h)$. However, as discussed in Section A.4, technical complications arise because Eq.(289) requires calculation of possibly ϑ–dependent normalizations Z_ϑ of $\rho(\mathbf{H}(\vartheta))$ for all ϑ. In such situations where the likelihood is defined as a mixture of conditional densities $p(y|x,h,\vartheta)$ and where we want to emphazise the need to deal with ϑ–dependent Z_ϑ we will also speak of a *disordered system*. Note however, that written in its eigenbasis the density operator for disordered systems takes again the form of Eq.(269), and vice versa a general ρ of that form is a mixture of components $|\varphi_i><\varphi_i|$. If such a formulation has been found, however, one already has $Z_i = \mathrm{Tr}(|\varphi_i><\varphi_i|) = 1$ for orthonormalized φ_i.

We remark that most problems studied in textbooks of quantum mechanics (or, analogously, of quantum field theory) assume a given ρ and aim in calculating $p(y|\mathbf{O}_x,\rho)$. Hereby ρ can, for example, be a pure (quantum mechanical) stationary state (bound state problems), a pure (quantum mechanical) non–stationary state (scattering with completely determined initial state), a stationary mixture (e.g., a system at finite temperature), a mixture of conditional $\rho(\vartheta)$ (disordered system), or, equivalently but differently specified, a general non–stationary mixture (e.g., scattering with not completely observed initial states). For such problems with fixed ρ no learning can occur. To allow learning a space of possible ρ together with a prior density $p(\rho)$ must be specified which is updated to obtain a posterior $p(\rho|D)$ after new data D have been received. The following table shows possible forms of density operators with $\varphi_i(\mathbf{H})$ denoting orthonormalized eigenfunctions of $\mathbf{H}$

	ρ
stationary pure state	$\vert\varphi(\mathbf{H})><\varphi(\mathbf{H})\vert$
general pure state	$\vert\varphi><\varphi\vert$
stationary mixture state	$\sum_i p(i\vert\mathbf{H})\,\vert\varphi_i(\mathbf{H})><\varphi_i(\mathbf{H})\vert$
disordered mixture state	$\int d\vartheta \sum_i p(\vartheta)p(i\vert\vartheta)\,\vert\varphi_i(\vartheta)><\varphi_i(\vartheta)\vert$
general mixture state	$\sum_i p(i)\,\vert\varphi_i><\varphi_i\vert$

B.3.5. Path integrals

The path integral approach provides an alternative to the operator formalism for quantum mechanics or quantum field theory, respectively [25,60,77,48]. For example, a density operator of a canonical ensemble (278) can be expressed as a path or functional integral

$$< y|\,\rho\,|y> = \frac{1}{Z}\int d\pi \int d\phi\, \phi_0(y)\phi_\beta^*(y) e^{\int_0^\beta d\tau \int dx^3 \left(i\pi\frac{\partial\phi}{\partial\tau} - H(\pi,\phi)\right)}, \tag{291}$$

with

$$Z = \int d\pi \int_{\phi_0=\phi_\beta} d\phi\; e^{\int_0^\beta d\tau \int dx^3 \left(i\pi\frac{\partial\phi}{\partial\tau} - H(\pi,\phi)\right)}. \tag{292}$$

Hereby $H(\pi, \phi)$ is a classical function depending on classical fields ϕ and π, and $\int d\pi\, d\phi$ denotes a functional integral over functions $\pi(x, \tau)$ and $\phi(x, \tau)$. The function H corresponds to a Hamiltonian

$$\mathbf{H} = \int dx\, H(\hat{\pi}, \hat{\phi}), \tag{293}$$

expressed in terms of field operators $\hat{\phi}(x, \tau)$ and their canonical conjugates $\hat{\pi}(x, \tau)$. For H which are quadratic in π the π–integral is gaussian and can be performed analytically. For more details see for example [35].

Calculating the functional integral in saddle point approximation yields the classical field equations. Such a saddle point approximation can be combined with the saddle point approximation for the h–integral which will be discussed in Section B.5 [67].

B.4. Bayes' rule for complete data

B.4.1. Posterior probabilities and likelihoods

Typically, model states h are defined by giving their data generating probabilities or likelihoods $p(D|h)$. The posterior probabilities $p(h|f) = p(h|D)$ we are interested in are related to the likelihoods $p(D|h)$ by Bayes' rule

$$p(h|f) = p(h|D) = \frac{p(D|h)p(h)}{p(D)} = \frac{p(D|h)p(h)}{\int dh\, p(D|h)p(h)}. \tag{294}$$

Restricting this Bayesian inversion for the moment to training data one finds

$$p(h|D) = \frac{p(D_T|h)p(h|D_0)}{\int dh\, p(D_T|h)p(h|D_0)} = \frac{\prod_i p(y_i|x_i, h)p(h|D_0)}{\int dh\, \prod_i p(y_i|x_i, h)p(h|D_0)}, \tag{295}$$

where the training data term has been written as product over all training data $D_T = (x_T, y_T) = \{(x_i, y_i)|1 \le i \le n\}$ (x_T, y_T denoting the vectors of x_i, y_i),

$$\frac{p(D_T|h)}{p(x_T)} = p(y_T|x_T, h) = \prod_i p(y_i|x_i, h). \tag{296}$$

Hence the effective distribution under the posterior state of knowledge f becomes the quotient of two correlation functions

$$p(y|x, f) = \frac{\left\langle p(y|x, h) \prod_i p(y_i|x_i, h)\right\rangle_P}{\left\langle \prod_i p(y_i|x_i, h)\right\rangle_P}, \tag{297}$$

where $\langle g(h)\rangle_P$ denotes the prior average $\int dh\, p(h|D_0)\, g(h)$.

In the following prior data D_0 will be treated analogously to training data D_T. We also assume enough prior data so that all information leading to nonuniform $p(h|D)$ is contained in D. Then, $p(h)$ is uniform (possibly improper, i.e., non-normalizable) and thus h–independent, so that $p(h, D) =$

$p(y_D|x_D,h)p(h)p(x_D) \propto p(y_D|x_D,h)p(x_D)$. We call such data $D = D_T \cup D_0$ *complete.*

The (training and prior data) generating probability or (complete) likelihood $p(D|h)$ is related to the posterior probability according to Eq.(294)

$$p(h|D) = p(h|y_D,x_D) = \frac{p(y_D|x_D,h)p(h|x_D)}{p(D|x_D)} = \frac{p(y_D|x_D,h)p(h|x_D)}{\int dh\, p(y_D|x_D,h)p(h|x_D)}, \qquad (298)$$

which becomes for complete data or uniform $p(h|x_D)$

$$p(h|D) = \frac{p(y_D|x_D,h)}{\int dh\, p(y_D|x_D,h)} \propto p(y_D|x_D,h). \qquad (299)$$

B.4.2. Posterior and likelihood energies

Let us now introduce posterior energy $E(h|f)$, posterior temperature $1/\beta_P$, with corresponding free energy $F(\mathcal{F}_0|f)$ and partition sum $Z(\mathcal{F}_0|f)$ to write the posterior

$$p(h|f) = p(h|D) = \frac{Z(h|D)}{Z(\mathcal{H}|D)} = e^{-\beta_P\left(E(h|D)-F(\mathcal{F}_0|D)\right)}, \qquad (300)$$

with

$$Z(h|D) = e^{-\beta_P E(h|D)}, \quad Z(\mathcal{H}|D) = \int dh\, e^{-\beta_P E(h|D)} = e^{-\beta_P F(\mathcal{H}|D)}. \qquad (301)$$

Analogously, the x_i–conditional training likelihood $p(y_i|x_i,h)$ becomes in terms of (conditional) training likelihood energy $E(y_i|x_i,h)$ and (conditional) training likelihood temperature $1/\beta_T$

$$p(y_i|x_i,h) = e^{-\beta_T \sum_i \left(E(y_i|x_i,h)-F(\mathcal{Y}|x_i,h)\right)} = \frac{e^{-\beta_T \sum_i E(y_i|x_i,h)}}{\prod_i Z(\mathcal{Y}|x_i,h)}, \qquad (302)$$

with free energy

$$F(\mathcal{Y}|x_i,h) = -\frac{1}{\beta_L} Z(\mathcal{Y}|x_i,h), \qquad (303)$$

and normalization factor $Z(\mathcal{Y}|x_i,h)$ over responses $y_i \in \mathcal{Y}$ for given h and x_i

$$Z(\mathcal{Y}|x_i,h) = \int_{\mathcal{Y}} dy_i\, e^{-\beta_L E(y_i|x_i,h)}. \qquad (304)$$

The complete conditional likelihood $p(y_D|x_D,h)$ reads in terms of (complete, conditional) likelihood energy $E(y_D|x_D,h)$ and (complete, conditional) likelihood temperature $1/\beta_L$

$$p(y_D|x_D,h) = e^{-\beta_L(E(y_D|x_D,h)-F(\mathcal{Y}|x_D,h))} = \frac{e^{-\beta_L E(Y|x_D h)}}{Z(\mathcal{Y}|h)}, \qquad (305)$$

with $E(y_D|x_D,h) = E_T + E_0$ and $E_T = E_T(y_T|x_T,h)$, $E_0 = E(y_0|x_0,h,D_T)$, free energy

$$F(\mathcal{Y}|x_D,h) = -\frac{1}{\beta_L} Z(\mathcal{Y}|x_D,h), \tag{306}$$

and normalization factor over data y_D for given state h

$$Z(\mathcal{Y}|x_D,h) = \int_{\mathcal{Y}} dy_D \, e^{-\beta_L E(y_D|x_D,h)}. \tag{307}$$

B.5. Saddle point approximation

B.5.1. Maximum a posteriori approximation

An exact analytical solution of the full integral in $r(a,f)$ is most times not possible and approximations have to be used. We also remark that for functions $h(x)$ of continuous variables x, like in field theory in physics, the integral $\int dh$ is a functional (or path) integral. Such integrals have typically to be defined by perturbation theory, starting from a well defined, e.g. gaussian, case. Alternatively, the integral can also be discretized, like it is done in lattice field theory [48] and like we will do for numerical calculations.

All approximations can only calculate a finite number of solvable terms. A solvable term can correspond to a solvable infinite sum or, e.g., a gaussian, integral. Low temperature approximations restrict the evaluation to the most important terms (thus they replace integration by maximization), high temperature approximations start from a case where all terms are equal (e.g., a cumulant expansion starts with the mean), while Monte Carlo integration evaluates a random sample of terms (so that under certain conditions the sampled sum converges to the true result). We will discuss in the following the maximum posterior approximation, which is a special variant of a saddle point approximation [14,12].

The h–integral within the risk $r(a,f)$ (247) involves two h–dependent factors

$$\int dh \, p(h|f)p(y|x,h) = \int dh \, e^{-\beta_P(E(h|f)-F(\mathcal{H}))} e^{-\beta_{x,h}(E(y|x,h)-F(\mathcal{Y}|x,h))} \tag{308}$$

with posterior temperature $1/\beta_P$ and posterior energy $E(h|f) = E(h|D)$, which will in the following be written simply $E(h)$. For a Taylor expansion of the energy $E(h)$ with respect to h around a minimum h^* of $E(h)$, the first order terms vanish and the Hessian H, i.e., the matrix of second derivatives is positive definite. Hence,

$$E(h) = e^{\langle h-h^*|\nabla\rangle} E(h^*) \tag{309}$$

$$= E(h^*) + \frac{1}{2}\langle \Delta h \,|H|\, \Delta h\rangle + \sum_{n=3}^{\infty} \frac{1}{n!} \langle \Delta h|\nabla\rangle^n \, E(h)\big|_{h=h^*} \,,$$

with H the positive definite Hessian of $E(h)$ at h^*

$$H(h) = \left. \frac{\partial^2 E(h)}{\partial h(x,y)\partial h(x',y')} \right|_{h=h^*}, \tag{310}$$

and ∇ acting on $E(h)$ but not on $\Delta h = h - h^*$.

Now assume $p(y|x,h)$ is slowly varying at the stationary point (i.e., it has a β_P–independent Taylor expansion at the stationary point at h^*) and approximate it by its value $p(y|x,h^*)$ at h^*. Then the second order term results in a gaussian with mean h^* and the Hessian of $E(h)$ at h^* as inverse covariance matrix H. Diagonalizing the positive definite H by an orthogonal transformation O

$$H = O^T D O, \tag{311}$$

changing the integration variables from (q–dimensional) h to g according to

$$\Delta h = \sqrt{\beta_P} (\sqrt{D} O)^{-1} g \tag{312}$$

with Jacobian

$$J = \det \left(\frac{\partial h(x)}{\partial g(y)} \right) = \beta_P^{-\frac{q}{2}} \det H^{-\frac{1}{2}} \tag{313}$$

the q–dimensional gaussian integral can be performed analytically

$$\int dh \, e^{-\frac{\beta_P}{2} \langle \Delta h | H | \Delta h \rangle} = J (2\pi)^{\frac{q}{2}} = (\det H)^{-\frac{1}{2}} \left(\frac{2\pi}{\beta_P} \right)^{\frac{q}{2}}. \tag{314}$$

Thus in case we restrict to the contribution of only the lowest minimum of E^*, i.e., assuming other local minima of E to be sufficiently larger than E^* for a given β_P, we have for β_P large enough (and no other stationary points contribute)

$$\int dh \, p(h|f) \, p(y|x,h) \approx (\det H)^{-\frac{1}{2}} \left(\frac{2\pi}{\beta_P} \right)^{\frac{q}{2}} e^{\beta_P F(\mathcal{H})} e^{-\beta_P E(h^*)} p(y|x,h^*). \tag{315}$$

The factors beside $p(y|x,h^*)$ are h, x, y and a–independent and can therefore be skipped from the risk in case only one minimum of $E(h)$ contributes to the h–integral. Moreover, evaluating $e^{-\beta_P F(\mathcal{H})} = Z(\mathcal{H}) = \int dh \, e^{-\beta_P E(h|f)}$ also by saddle point approximation one finds for β_P large enough

$$\frac{1}{Z(\mathcal{H})} = e^{\beta_P F(\mathcal{H})} = \int dh \, e^{-\beta_P E(h|f)} \approx (\det H)^{\frac{1}{2}} \left(\frac{2\pi}{\beta_P} \right)^{-\frac{q}{2}} e^{\beta_P E(h^*)} \tag{316}$$

so the factors in Eq.(315) cancel yielding

$$\int dh \, p(h|f) \, p(y|x,h) = p(y|x,h^*) + \mathcal{O}(\frac{1}{\beta_P}). \tag{317}$$

(For the justification of the symbol $\mathcal{O}(\frac{1}{\beta_P})$ see Section B.5.3.)

A maximum posterior approximation of the h integral in $r(a,f)$ minimizes instead of the expected risk $r(a,f)$ of (247) the risk

$$r(a,h^*) = \int dx \int dy\, p(x) p(y|x,h^*) l(x,y,a) \tag{318}$$

for

$$h^* = \text{argmax}_{h\in\mathcal{H}} p(h|f) = \text{argmin}_{h\in\mathcal{H}} E(h|f), \tag{319}$$

which assumes slowly varying $p(y|x,h)$ at h^*. Hence, integration is replaced by minimization of $E(h|f)$ (maximization of $p(h|f)$).

For multiple minima of $E(h)$ which are well enough separated the volume factor $|\det H|^{-\frac{1}{2}}$ can be used as weighting factor for the contributions of different minima. Well enough separated, means that the gaussian approximations around the different maxima do not considerably overlap. Note however, that in this case the (nontrivial) calculation of the Hessian is required. If this is not possible, the ratio $\det H_1/\det H_2 = \det H_1 H_2^{-1}$ of the determinant of two Hessians H_1, H_2 can be approximated by expanding the logarithm according to

$$\ln(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} x^n \tag{320}$$

around the identity matrix I in $\Delta = H_1 H_2^{-1} - I$

$$\det H_1 H_2^{-1} = e^{\mathrm{Tr}\ln(I+\Delta)}. \tag{321}$$

A graphical algorithm for expansion of a determinant is given by the polymer expansion [48].

For the case of minima with overlapping contributions see [6,47,13].

B.5.2. Complete maximum likelihood approximation

In this Section the saddle point approximation will be discussed in the limit of small (conditional) likelihood temperatures $1/\beta_L$ instead of posterior temperatures $1/\beta_P$. It should be stressed, that in contrast to a (conditional) maximum likelihood approximation which maximizes the likelihood $p(y_T|x_T,h)$ $= \prod_i p(y_i|x_i,h)$ for training data we consider here the complete (conditional) likelihood $p(y_D|x_D,h) = p(y_T \cup y_P|x_D,h)$ containing training and prior data. The two approximations are equivalent only for uniform $p(h|D_0)$.

To be a bit more general, we also include continuous hidden variables θ and discrete hidden variables i. In the following the integral over θ will be treated in saddle point approximation analogous to the h–integral while the (finite) sum over i will be treated exactly. Then, one finds for complete data, i.e., uniform, possibly improper, $p(h)$ for the predictive density

$$p(y|x,D) = \int dh\, p(h|D)\, p(y|x,h) \tag{322}$$

$$= \frac{\int dh\, p(y_D|x_D,h)\, p(y|x,h)}{\int dh\, p(y_D|x_D,h)} \tag{323}$$

$$= \frac{\int dh\, \int d\theta\, \sum_i p(y_D,\theta,i|x_D,h) p(y|x,h)}{\int dh\, \int d\theta\, \sum_i p(y_D,\theta,i|x_D,h)} \tag{324}$$

$$= \frac{1}{Z_L}\int dh\, \int d\theta\, \sum_i p(y_D,\theta,i|x_D,h) p(y|x,h) \tag{325}$$

introducing $Z_L = \int dh \int d\theta \sum_i p(y_D,\theta,i|x_D,h)$. Decomposing $\theta = (\theta_x,\theta_y)$ and $i = (i_x,i_y)$ in an input and an output noise component (see Section B.2.4), this yields

$$p(y|x,D) = \frac{1}{Z_L}\int dh\, \int d\theta\, \sum_i p(\theta_x,i_x|x_D) p(y_D|\theta_y,i_y) \times p(\theta_y,i_y|x_D,h,\theta_x,i_x) p(y|x,h), \tag{326}$$

The likelihood may be written in terms of (the complete, θ_x–, i_x–conditional) likelihood energy $E(\theta_y,i_y|\theta_x,i_x,h)$ and corresponding likelihood temperature $1/\beta_L$

$$p(\theta_y,i_y|x_D,h,\theta_x,i_x) = e^{-\beta_L\left(E(\theta_y,i_y|\theta_x,i_x,h) - F(\Theta_y,\mathcal{I}_y|\theta_x,i_x,h)\right)}, \tag{327}$$

with $F(\Theta_y,\mathcal{I}_y|\theta_x,i_x,h)$ the free energy corresponding to normalization over θ_y and i_y. Notice, that for θ–dependent free energy its derivative contributes to the stationary equation for θ_x. For gaussians energies the free energy yields contributions proportional to $\beta^{-1}\ln\beta$ which for large β is small compared to the energy. Furthermore, an asymptotic justification of a saddle point approximation in that case would require the introduction of $\tilde\beta$ multiplying the logarithm of the sum over i, i.e., $\sum_i e^{-\beta(E-F)} \to e^{\tilde\beta \ln \sum_i e^{-\beta(E-F)}}$. Approximating the h– and θ–integrals by a gaussian centered at a stationary point the contributions from numerator and denominator Z_L cancel yielding, like Eq.(317)

$$\int dh\, p(h|D)\, p(y|x,h) \approx p(y|x,h^*). \tag{328}$$

If the normalization $Z(\Theta_y,\mathcal{I}_y|\theta_x,i_x,h)$ over θ_y and i_y is i_x–, θ_x–, h–independent then $h^* = h^*(\theta^*)$ with

$$h^* = \mathrm{argmax}_h \sum_i p(\theta_x^*,i_x|x_D) p(y_D|\theta_y^*,i_y) e^{-\beta_L E(\theta_y^*,i_y|\theta_x^*,i_x,h)}, \tag{329}$$

$$\theta^* = \mathrm{argmax}_\theta \sum_i p(\theta_x,i_x|x_D) p(y_D|\theta_y,i_y) e^{-\beta_L E(\theta_y,i_y|\theta_x,i_x,h^*)}, \tag{330}$$

which are solutions of the coupled stationary equations

$$0 = \frac{\partial}{\partial h}\sum_i p(\theta_x,i_x|x_D) p(y_D|\theta_y,i_y) e^{-\beta_L E(\theta_y,i_y|\theta_x,i_x,h)}, \tag{331}$$

$$0 = \frac{\partial}{\partial \theta} \sum_i p(\theta_x, i_x|x_D) p(y_D|\theta_y, i_y) e^{-\beta_L E(\theta_y, i_y|\theta_x, i_x, h)}. \tag{332}$$

If i_x–, θ_x–, or h–dependent, normalization terms $Z(\Theta_y, \mathcal{I}_y|\theta_x, i_x, h)$ also appear in the stationarity equations.

In the case of n training data with likelihood energy being the sum

$$\sum_j^n E_j(y_j|x_j, h) = n \frac{1}{n} \sum_j^n E_j(y_j|x_j, h) = n \, \langle E_j \rangle_{\text{training}} \tag{333}$$

one can choose

$$\beta_L = n, \tag{334}$$

provided additional prior information ensures that the expectation $(1/n) \sum_i E_i = n \, \langle E_j \rangle_{\text{training}}$ converges for large n. In that case the saddle point approximation is a large n approximation.

B.5.3. Loop expansion

One can go beyond a maximum posterior approximation and include higher order contributions. We will discuss in the following the expansion in posterior temperatures. For that purpose, we will write a full expectation over $p = p(x|y, h)$ as a sum of gaussian expectations, i.e., symbolically,

$$\langle \, p \, \rangle_{\text{full}} = a^* \left\langle p \, e^R \right\rangle_{\text{gauss}} = a^* \left\langle p \, \left(1 + R + \frac{R^2}{2} + \cdots \right) \right\rangle_{\text{gauss}}, \tag{335}$$

where $a^* = e^{-\beta_P E(h^*)} (2\pi)^{\frac{d}{2}} / (JZ)$ is a maximum a posteriori result. The expansion of the exponential e^R corresponds to a moment expansion analogous to Eqs.(185,192). This expectation can be expressed in terms of moments of h by expanding p and the remaining term R around h^*. This will result in an expansion in powers of $1/\beta_P$. Expanding also Z in the denominator common prefactors cancel

$$\langle \, p \, \rangle_{\text{full}} = \frac{\left\langle p \, \left(1 + R + \frac{R^2}{2} + \cdots \right) \right\rangle_{\text{gauss}}}{\left\langle 1 + R + \frac{R^2}{2} + \cdots \right\rangle_{\text{gauss}}}. \tag{336}$$

Hence, we begin by expanding $p(y|x, h)$ around a stationary point h^*

$$p(y|x, h) = e^{\langle \Delta h | \nabla \rangle} p(y|x, h) \big|_{h=h^*} = \sum_{n=0}^{\infty} \frac{1}{n!} \langle \Delta h | \nabla \rangle^n \, p(y|x, h) \big|_{h=h^*}. \tag{337}$$

Changing for $H = O^T DO$ the integration variables from h to g with $\Delta h = h - h^*$ $= \sqrt{\beta_P}(\sqrt{D}O)^{-1}g$ and Jacobian J of Eq.(313), we find

$$\int dh\, p(h|f)\, p(y|x,h)$$

$$= \frac{1}{Z}\int dh\, \left(e^{\langle\Delta h|\nabla\rangle}p(y|x,h^*)\right) e^{-\beta_P\left(e^{\langle\Delta h|\nabla\rangle}E(h^*)\right)}$$

$$= \frac{1}{Z}e^{-\beta_P E(h^*)}\int dh\, \left(\sum_{m=0}^{\infty}\frac{1}{m!}\langle\Delta h|\nabla\rangle^m p(y|x,h^*)\right)$$

$$\times e^{-\beta_P\left(\frac{1}{2}\langle\Delta h|H|\Delta h\rangle + \sum_{n=3}^{\infty}\frac{1}{n!}\langle\Delta h|\nabla\rangle^n E(h^*)\right)}$$

$$= \frac{1}{Z}e^{-\beta_P E(h^*)}\int dh\, e^{-\frac{\beta_P}{2}\langle\Delta h|H|\Delta h\rangle}\left(\sum_{m=0}^{\infty}\frac{1}{m!}\langle\Delta h|\nabla\rangle^m p(y|x,h^*)\right)$$

$$\times\sum_{k=0}^{\infty}\frac{(-\beta_P)^k}{k!}\left(\sum_{n=3}^{\infty}\frac{1}{n!}\langle\Delta h|\nabla\rangle^n E(h^*)\right)^k$$

$$= \frac{1}{J\,Z}e^{-\beta_P E(h^*)}\int dg\, e^{-\frac{\langle g|g\rangle}{2}}\left(\sum_{m=0}^{\infty}\frac{1}{m!\beta_P^{\frac{m}{2}}}\left\langle(\sqrt{D}O)^{-1}g|\nabla\right\rangle^m p(y|x,h^*)\right)$$

$$\times\sum_{k=0}^{\infty}\frac{1}{k!}\left(-\sum_{n=3}^{\infty}\frac{1}{n!\beta_P^{\frac{n}{2}-1}}\left\langle(\sqrt{D}O)^{-1}g|\nabla\right\rangle^n E(h^*)\right)^k,$$

$$= \frac{1}{J\,Z}e^{-\beta_P E(h^*)}\int dg\, e^{-\frac{\langle g|g\rangle}{2}}\sum_{m=0}^{\infty}\sum_{k=0}^{\infty}\sum_{n_1=3}^{\infty}\cdots\sum_{n_k=3}^{\infty}\frac{(-1)^k}{k!m!\prod_l^k n_l!}\frac{1}{\beta_P^{\frac{m}{2}+\sum_l^k\frac{n_l}{2}-k}}$$

$$\times\left(\left\langle(\sqrt{D}O)^{-1}g|\nabla\right\rangle^m p(y|x,h^*)\right)\prod_l^k\left(\left\langle(\sqrt{D}O)^{-1}g|\nabla\right\rangle^{n_l}E(h^*)\right), \tag{338}$$

with normalization factor

$$J\,Z = e^{-\beta_P F(\mathcal{H})}(\frac{\beta_P}{2\pi})^{\frac{q}{2}}(\det H)^{\frac{1}{2}}. \tag{339}$$

The normalization integral Z can be treated analogously in saddle point approximation leading in first order to the cancellation of the prefactor. The individual terms are moments of a multidimensional gaussian and can be evaluated using Wick's theorem (182). Because the gaussian has mean zero odd moments vanish and Eq.(338) is an expansion in $1/\beta_P$, also known as loop expansion. Only if not expanding around a saddle point linear terms would survive. Higher order terms are usually represented as Feynman diagrams [31,25,58,60,51,77,32,42,10,48]. Surprisingly, the presence of the denominator $e^{-\beta_P F(\mathcal{H})} = Z(\mathcal{H})$ leads to a simplification. It can be shown that it cancels exactly the so called vacuum diagrams. Further simplifications arise for expanding the cumulant generating functional $W = \ln Z$ where only connected diagrams contribute or its Legendre transform Γ where only amputated and one–particle irreducible diagrams have to be considered. Γ is the generating function of so called (proper) vertex functions. Equations connecting moments, cumulants, and vertex functions of different order can be obtained (e.g., Ward–identities, equations of motion or Dyson–Schwinger equations). Solving Eq.(225), for example, corresponds to summing up infinite subclasses of diagrams. The name loop expansion stems from the fact that expansion of Z (which does not contain $p(y|x,h)$) or Γ in powers of β_P^{-1} is equivalent to an expansion in the number of loops of the corresponding Feynman diagrams.

B.5.4. Stationarity equations

A specific state h is given by a parameter vector ξ determining $p(y|x,h)$, $\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}$. Thus, the stationarity equation to be solved for an a–posteriori approximation is of the form

$$\frac{\partial E(h)}{\partial \xi_m} = 0, \quad \forall m, \tag{340}$$

where $E(h)$ is an exponent to be minimized.

In the present paper we use equally weighted quadratic error terms for training data assuming gaussian $p(y|x,h)$ according (254) with equal variance (and therefore equal normalization) for all x and mean specified by a regression function $h(x)$. Then the stationarity equation is obtained by setting the functional derivatives

$$\frac{\partial E(h)}{\partial h(x)} = 0, \quad \forall x \in \mathcal{X}_R. \tag{341}$$

In the more general case, like in density estimation where one wants to allow for non–gaussian densities $p(y|h)$ (then also the data terms $\ln p(y_i|h)$ in the error functional become non–quadratic), arbitrary $p(y|x,h)$ can be used, as long as they fulfil the normalization conditions

$$\frac{\partial E(h)}{\partial p(y|x,h)} = 0, \quad \forall x \in \mathcal{X}_R, y \in \mathcal{Y} \quad \text{with} \int dy\, p(y|x,h) = 1, \forall x \in \mathcal{X}_R, h \in \mathcal{H}. \tag{342}$$

In that case the posterior probability can be maximized with respect to $p(y|x,h)$ using standard techniques of constraint optimization [19,3,7]. The normalization conditions can be implemented by δ–functions

$$p(h|f) \propto e^{-\beta\left(\sum_i E(y_i|x_i,h) + E(h|D_0)\right)}\, \delta\!\left(\int dy\, e^{-\beta E(y|x,h)+c(x)} - 1\right), \tag{343}$$

where $p(h|D_0) \propto e^{-\beta E(h|D_0)}$, $p(x|y,h) \propto e^{-\beta E(y|x,h)}$, $c(x) = -\ln Z(\mathcal{Y}|x,h) = (1/\beta)F(\mathcal{Y}|x,h)$, h–independent factors have been skipped, and we have written $\beta_L = \beta$. Using the Fourier representation of the δ–function

$$\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dk\, e^{ikx} = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} dk\, e^{-kx}, \tag{344}$$

and performing the Λ integral by a saddle point approximation (which is exact for the delta function) yields

$$p(h|f) \propto e^{-\beta\left(\sum_i E(y_i|x_i,h)+E(h|D_0)\right)+\int dx\, \Lambda^*(x)\left(1-\int dy\, e^{-\beta E(y|x,h)+c(x)}\right)}. \tag{345}$$

Here the Lagrange parameter function $\Lambda^*(x)$ denotes the stationary value of $\Lambda(x)$, i.e., a solution of the saddle point condition $\partial P(h|f)/\partial\Lambda(x) = 0$. It is easy to see that this stationarity condition which Λ has to fulfil at the end of the optimization procedure is equivalent to the normalization constraint. There exist many standard iteration procedures to perform the maximization of (345).

Alternatively, the normalization constraint can be implemented by writing

$$p(y|x,h) = \frac{g(x,y,h)}{Z(\mathcal{Y}|x,h)} \tag{346}$$

and solve for a stationary point $g(x,y,h^*)$

$$\frac{\partial E\left(\frac{g(x,y,h)}{Z(\mathcal{Y}|x,h)}\right)}{\partial g(x,y,h)} = 0, \quad \forall x \in \mathcal{X}_R, y \in \mathcal{Y}. \tag{347}$$

This is equivalent to the insertion of the δ–functions in Eq.(343) and using always the stationary value $\Lambda^*(x)$ during iteration.

B.6. Approximation problems

In the most common kinds of problems the hypothesis space is identified with the model space $\mathcal{A} = \mathcal{H}$. Let an *approximation problem* be defined as a situation with $\mathcal{A} = \mathcal{H}$ and log–loss

$$b(x)l(x,y,a{=}h) + c(x,y) = -\ln p(y|x,h) = \beta E(y|x,h) \tag{348}$$

with a–independent coefficients $c(x,y)$ and $b(x) > 0$. Notice that the probability $p(y|x,h)$ is that governing the production of test data and is in principle not related to the posterior $p(h|D)$ or likelihood probabilities $p(y_D|x_D,h)$. Often, however, training data are also assumed to be produced according to the same $p(y|x,h)$. For log–loss it is easy to see, by using Jensen's inequality, that

$$a^* = \text{argmin}_{a\in\mathcal{A}} r(a,h^*) = h^*, \tag{349}$$

and thus

$$p(y|x,a^*) = p(y|x,h^*). \tag{350}$$

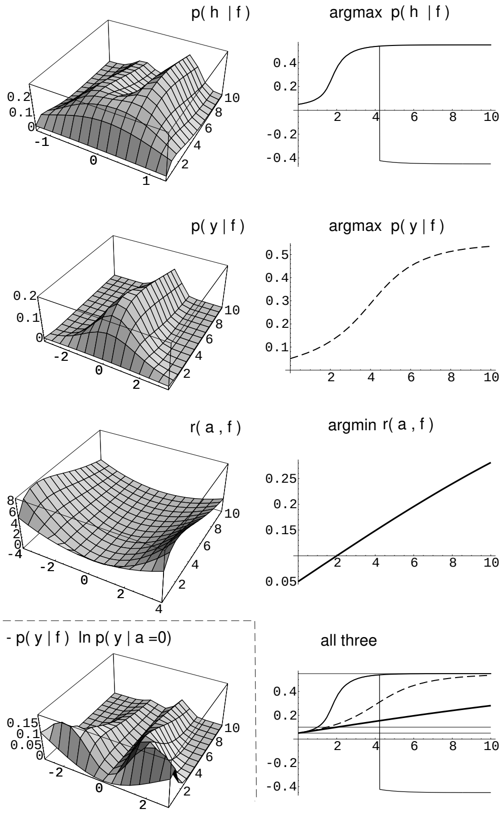

Figure 17.

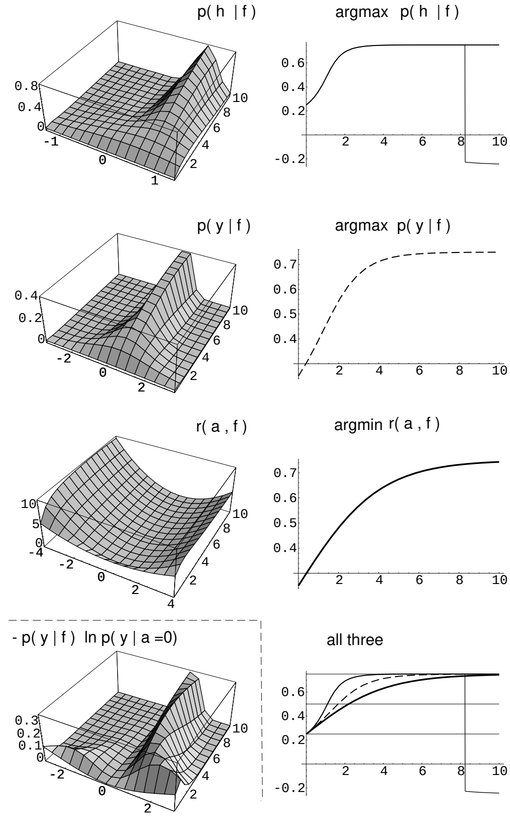

Figure 18.

The figures compare the maximum posterior approximation (top row) with the mode of the posterior (second row) and a full Bayesian risk minimization (third row) for the model defined by Eqs.(353,354,355). (Fig.17 for $d = 0.1$ and Fig.18 for $d = 0.5$.) Row 1. Left: $p(h|f)$ (unnormalized) for $-1.3 \le h \le 1.3$ and $0 \le \beta \le 10$. Right: $\text{argmax}_{h\in\mathcal{H}} p(h|f)$ for $0 \le \beta \le 10$. The second local maximum which appears for larger β (low temperature) is also shown. Row 2. Left: $p(y|f)$ (unnormalized) for $-3.3 \le y \le 3.3$ and $0 \le \beta \le 10$. Right: $\text{argmax}_{y\in\mathcal{Y}} p(y|f)$ for $0 \le \beta \le 10$. Row 3: Left: $r(a,f)$ for $-4 \le a \le 4$ and $0 \le \beta \le 10$. Right: $\text{argmin}_{a\in\mathcal{A}} r(a,f)$ for $0 \le \beta \le 10$. Row 3: Left: The contributions of $-3 \le y \le 3$ to the full risk for $a = 0$ and $0 \le \beta \le 10$. Right: All three approximations within the same diagram.

Thus, for approximation problems the maximum posterior approximation h^* gives already a consistent solution $a = h^*$. In this case the energy function of the posterior probability $p(h|f)$ to be minimized is also called *error* function, and the maximum a posteriori approximation is equivalent to empirical risk minimization [68–70]. In the typical case of gaussian states $p(y|x,h)$ the log–loss is up to a constant proportional to the squared error $(y - a(x))^2$.

A similar result holds for the full Bayesian approach. Choosing actions a

from the space $\mathcal{A} = \mathcal{F}$ of states of knowledge f instead from the model space $\mathcal{H}$ of states of Nature h ($\mathcal{F}$ is the convex hull of $\mathcal{H}$), and using log–loss (348) gives

$$a^* = \text{argmin}_{a\in\mathcal{A}} r(a, f) = f, \tag{351}$$

and thus

$$p(y|x, a^*) = p(y|x, f). \tag{352}$$

Other choices are the squared error $(y - a(x))^2$ for which the mean of y over the predictive distribution $p(y|x, f)$ is optimal, absolute error loss $|y - a(x)|$ which leads to the median, and zero–one loss $1 - \delta_{y,a(x)}$ resulting in the mode to be optimal. For gaussian distributions mean, median, and mode coincide.

For non–approximation problems calculating the maximum posterior solution h^* (or the state of knowledge f, respectively) has to be followed by the second step of finding the optimal a^* which minimizes $r(a, h^*)$ (or $r(a, f)$) for given h^* (or f) [44]. For squared error but non–gaussian $p(y|x, h^*)$, for example, the second minimization step would require the calculation of the mean of y under $p(y|x, h^*)$.

There are typical non–approximation situations. Complexity costs, for example, which penalize a complex a (e.g. due to computational resources or to facilitate comprehensibility) are typical for non–approximation problems. Only when there are reasons to believe that also the real state of Nature h fulfils such complexity constraints this results in an approximation problem. This means, one has to distinguish between complexity costs and knowledge about true simplicity in nature.

Related is the distinction between a generative and a reconstruction model. Consider the face detection task of Example 2 in Section 2.1. Hereby, a reconstruction model $p(y|x, h)$ would specify a probability of face vs. non–face given an image, while a generative model $p(x|y, \tilde{h})$ specifies the probability of an image provided it represents a face or non–face. The roles of x and y as independent and dependent variables are herein exchanged. Looking for an optimal $y = a(x)$ can be formulated as approximation problem for the reconstruction model. However, the probability of representing a face does depend on all non–faces which can appear. Hence changing the set of non–faces the reconstruction model has to be adapted. The specification of the inverse generative model is often easier. Here the generative model for faces $p(x|y, \tilde{h})$ does not have to be changed if the class of non–faces changes. Modeling physical processes of the mapping from objects to images also falls in this class. But a generative model does not define an approximation problem for the function a which returns an answer $y = a(x)$ for a given x and not vice versa. Then a maximum posteriori approximation requires a second minimization.

Finally, Figs.17,18 show the relations between the optimal a for a full Bayesian risk $r(a, f)$ and a maximum posterior approximation (or empirical risk minimization) for a simple situation with only one x–value, i.e., $\mathcal{X} = \{x\}$, so x

and x_D can be skipped from the notation. Assumed are gaussian model states parameterized by their mean $h \in \mathcal{H} = I\!R$

$$p(y|h) = \sqrt{\frac{\beta}{2\pi}} e^{-\frac{\beta}{2}(h-y)^2}, \tag{353}$$

real numbers as possible actions $a \in \mathcal{A} = \mathcal{H} = I\!R$ with log–loss

$$l(y,a) = (y-a)^2 = -\frac{2}{\beta} \ln p(y|h=a) + \frac{1}{\beta} \ln \frac{\beta}{2\pi}. \tag{354}$$

Data y_D of the form d AND (b OR c) with $b = 1$, $c = -1$ have been represented by a gaussian mixture with equally weighted components

$$p(y_D|f) = \frac{\beta}{2\pi} e^{-\beta(h-d)^2/2} \left(e^{-\frac{\beta}{2}(h-1)^2} + e^{-\frac{\beta}{2}(h+1)^2} \right). \tag{355}$$

This means for complete data, i.e., uniform prior $p(h)$

$$p(h|f) = \frac{p(y_D|h)}{\int dh\, p(y_D|h)} = \frac{p(y_D|h)}{\frac{1}{2}\sqrt{\frac{\beta}{\pi}} \left(e^{-\frac{\beta}{4}(d-1)^2} + e^{-\frac{\beta}{4}(d+1)^2} \right)}. \tag{356}$$

The expected risk reads

$$r(a,f) = \int dy\, p(y|f)\, l(y,a), \tag{357}$$

with predictive distribution

$$p(y|f) = \int dh\, p(y|h)\, p(h|f). \tag{358}$$

Shown are the β–dependency of 1. the maximum posterior solutions $h^* = \text{argmax}_{h\in\mathcal{H}} p(h|f)$ which for the given approximation situation corresponds to an empirical risk minimization $a^* = \text{argmin}_{a\in\mathcal{A}} E(a)$ for error $E(a) = -\ln p(a|f)$, 2. the mode of the predictive distribution $\text{argmax}_{y\in\mathcal{Y}} p(y|f)$, and 3. the full Bayesian solution $a^*_{Bayes} = \text{argmin}_{a\in\mathcal{A}} r(a,f)$. In Fig.17 the 'data' term $d = 0.1$ is nearly in the middle between the two alternative templates $b = 1$ and $c = -1$, while in Fig.18 the 'data' term $d = 0.5$ is much nearer to $b = 1$. Compared to Fig.17 the second local maximum of the posterior probability $p(h|f)$ appears at larger β (or lower temperature) in Fig.18, where also the maximum posterior approximation is better. Indeed it is well known from physics that mean field (maximum posterior) approximations break down near phase transitions. On the other hand, using an adapted β' unequal to the true β an improved solution can be obtained by a maximum posterior approximation.

Acknowledgements

The work was started during a research stay at the Massachusetts Institute of Technology. The author was supported by a Postdoctoral Fellowship (Le 1014/1–1) from the Deutsche Forschungsgemeinschaft and a NSF/CISE Postdoctoral Fellowship. He also wants to thank Federico Girosi, Gernot Münster, Tomaso Poggio, Joerg Uhlig, Achim Weiguny, and Christian Wieczerkowski, for stimulating discussions.

References

[1] Aleksander, I, & Morton, H.: Neurons and Symbols. Chapman& Hall, 1993.
[2] Balian, R.: *From Microphysics to Macrophysics.* (Vol. I), Springer–Verlag, 1991.
[3] Bazaraa, M.S., Sherali, H.D., & Shetty, C.M.: *Nonlinear Programming.* (2nd Ed.) New York: Wiley, 1993.
[4] Beck. C. & Schlögl F.: *Thermodynamics of chaotic systems.* Cambridge University Press.
[5] Berger, J.O.: *Statistical Decision Theory and Bayesian Analysis.* Springer–Verlag, 1980.
[6] Berry, M.V.: *Proc. Phys. Soc.*(London) **89**, 479, 1966.
[7] Bertsekas, D.P.: *Nonlinear Programming.* Belmont, MA: Athena Scientific, 1995.
[8] Binder, K.(ed.): *The Monte Carlo Method in Condensed Matter Physics.* Springer–Verlag, 1992.
[9] Binder, K. & Heermann, D.W.: *Monte Carlo Simulation in Statistical Physics.* (Corr. 2nd Ed.) Springer–Verlag, 1992.
[10] Binney, J.J., Dowrick, N.J., Fisher, A.J., & Newman, M.E.J.: *The Theory of Critical Phenomena.* Oxford: Clarendon Press, 1992.
[11] Bishop C.M.: *Neural Networks for Pattern Recognition.* Oxford: Clarendon Press, 1995.
[12] Bleistein, N. & Handelsman, N.: *Asymptotic Expansions of Integrals.* (Originally published in 1975 by Holt, Rinehart and Winston, New York) New York: Dover, 1986.
[13] Connor J.N.L. & Marcus, R.A.: *J. Chem. Phys.* **55**, 5636, 1971.
[14] De Bruijn, N.G.: *Asymptotic Methods in Analysis.* (Originally published in 1958 by the North–Holland Publishing Co., Amsterdam) New York: Dover, 1981.
[15] Deco, G. and Obradovic, D.: *An Information–Theoretic Approach to Neural Computing.* Springer–Verlag, 1996.
[16] Doob, J.L.: *Stochastic Processes.* (New edition 1990) Wiley, New York, 1953.
[17] Efetov, K.B.: Supersymmetry and theory of disordered metals. *Advances in Physics*, **32** (1), 53–127, 1983.
[18] Fischer, K.H. & Hertz J.A.: *Spin Glasses.* Cambridge University Press, 1991.
[19] Fletcher, R.: *Practical Methods of Optimization.*(2nd Ed.) Wiley, 1987.
[20] Gardiner, C.W.: *Handbook of Stochastic Methods.* (2nd Ed.), Springer–Verlag, 1990.
[21] Gelman, A, Carlin, J.B,. Stern, H.S., & Rubin, D.B.: *Bayesian Data Analysis.* Chapman& Hall, 1995.
[22] Geman, S. & Geman, D: Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **6** (7), 721–741, 1984.
[23] Girosi, F., Jones, M., & Poggio, T.: Regularization Theory and Neural Networks Architectures. *Neural Computation* **7** (2), 219–269, 1995.
[24] Girosi, F.,: An equivalence between sparse approximation and support vector machines. A.I. Memo No.1606, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1997. (To appear in *Neural Computation*)

[25] Glimm, J. & Jaffe, A.,: *Quantum Physics. A Functional Integral Point of View.* New York: Springer–Verlag, 1987.
[26] Golden, R.M.: *Mathematical Methods for Neural Network Analysis and Design.* Cambridge, MA: MIT Press, 1996.
[27] Hertz, J., Krogh, A., & Palmer, R.G.: *Introduction to the Theory of Neural Computation.* Addison–Wesley, 1991.
[28] Huber, P.J.: Robust Smoothing. In Launer, E. & Wilkinson G. (eds.) *Robustness in Statistics* New York: Academic Press, 1979.
[29] Huber, P.J.: *Robust Statistics.* New York: Wiley, 1981.
[30] Hull, J.: *Options, Futures, and other Derivative Securities.* Prentice Hall, 1989.
[31] Itzykson, C. & Zuber, J.–B.: *Quantum Field Theory.* New York: McGraw–Hill, 1985.
[32] Itzykson, C. & Drouffe, J–M.: *Statistical field theory.* Vol 1 & vol 2, Cambridge University Press, 1989.
[33] Jensen, F.V.: *An Introduction to Bayesian Networks.* New York: Springer, 1996.
[34] van Kampen, N.G.: *Stochastic Processes in Physics and Chemistry.* Amsterdam: North–Holland, 1992.
[35] Kapusta, J.L.: *Finite–temperature field theory* Cambridge University Press, 1989.
[36] Kirkpatrick, S., Gelatt Jr., C.D., & Vecchi, M.P.: Optimization by Simulated Annealing. *Science* **220**, 671–680, 1983.
[37] Klir, G.J. & Yuan, B.: *Fuzzy Sets and Fuzzy Logic.* Prentice Hall, 1995.
[38] Klir, G.J. & Yuan, B. (eds.): *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems.* World Scientific, 1996.
[39] Landau, L.D. & Lifshitz, E.M. *Statistical Physics* Part 1 (3rd edition) New York: Pergamon, 1980.
[40] Langton, C.G., Life at the Edge of Chaos. In: Langton, C.G., Taylor, C., Farmer, J.D., & Rasmussen, S. (eds.) *Artificial Life II.* Proceedings of the Workshop on Artificial Life held Feb., 1990 in Santa Fe, NM. Santa Fe Institute, Studies in the Sciences of Complexity, Proceedings Volume X. Reading, MA: Addison–Wesley, 1992.
[41] Lauritzen, S.L.: *Graphical Models.* Clarendon Press, Oxford, 1996.
[42] Le Bellac, M.: *Quantum and Statistical Field Theory.* Oxford: Clarendon Press, 1991.
[43] Lemm, J.C.: Inhomogeneous Random Phase Approximation for Nuclear and Atomic Reactions. *Annals of Physics* **244** (1), 136–200 & 201–238, 1995.
[44] Lemm, J.C. *Prior Information and Generalized Questions.* A.I.Memo No. 1598, C.B.C.L. Paper No. 141, Massachusetts Institute of Technology, 1996.
[45] Lifshits, M.A.: *Gaussian Random Functions.* Kluwer Academic Publ., 1995.
[46] Mezard, M., Parisi, G., & Virasoro, M.A.: *Spin Glass Theory and Beyond.* Singapore: World Scientific, 1987
[47] Miller, W.H.: *J. Chem. Phys* **53**, 3578, 1970.
[48] Montvay, I. & Münster, G.: *Quantum Fields on a Lattice.* Cambridge University Press, 1994.
[49] Mudry, Ch., Chamon, C., & Wen, X.–G.: Two–dimensional conformal field theory for disordered systems at criticality. *Nucl.Phys.* B **466** [FS], 383–443, 1996.
[50] Neal R.M.: *Bayesian Learning for Neural Networks.* Springer, 1996.
[51] Negele, J.W. & Orland, H. *Quantum Many–Particle Systems.* Frontiers In Physics Series (Vol. 68), Addison–Wesley, 1988.
[52] Michie, D., Spiegelhalter, D.J., & Taylor, C.C. (Eds.): *Machine Learning, Neural and Statistical Classification.* New York: Ellis Horwood, 1994.
[53] Parisi, G.: *Field Theory, Disorder and Simulations.* Singapore: World Scientific, 1992.
[54] Pearl, J.: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Francisco, CA: Morgan Kaufmann, 1988.

[55] Poggio, T., Torre, V., & Koch, C.: Computational Vision and Regularization Theory. *Nature* **317**, 314–319, 1985.

[56] Poggio, T. & Girosi, F.: Networks for Approximation and Learning. *Proceedings of the IEEE*, **78** (9), 1990.

[57] Poggio, T. & Girosi, F.: *Notes on PCA, Regularization, Sparsity and Support Vector Machines.* A.I.Memo in preparation, Massachusetts Institute of Technology, 1998.

[58] Pokorski, S.: *Gauge Field Theories.* Cambridge University Press, 1987.

[59] Ripley, B.D.: *Pattern Recognition and Neural Networks.* Cambridge University Press, 1996.

[60] Rivers R.J.: *Path integral methods in quantum field theory.* Cambridge University Press, 1987.

[61] Rockafellar, R.T.: *Convex Analysis.* Princeton University Press, 1970.

[62] Rose, K., Gurewitz, E., & Fox, G.C.: Statistical mechanics and phase transitions in clustering. *Phys. Rev. Lett.* **65**, 945–948, 1990.

[63] Schapire R.E., Freund, Y., Bartlett, P., & Lee, W.S.: Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods. Submitted for publication. Extended abstract in *Machine Learning: Proceedings of the Fourteenth International Conference*, 1997.

[64] Smola A.J. & Schölkopf, B,: From regularization operators to support vector kernels. In: Jordan, M.I., Kearns, M.J., & Solla S.A. (Eds.): *Advances in Neural Information Processing Systems*, vol.10. Cambridge, MA: MIT Press, 1998.

[65] Smola A.J., Schölkopf, B, & Müller, K–R.: The connection between regularization operators and support vector kernels. *Neural Networks* **11**, 637–649, 1998.

[66] Tikhonov, A.N.: Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.* **4**, 1035–1038, 1963.

[67] Uhlig, J.: Doctoral dissertation, Münster University (in preparation).

[68] Vapnik, V.N.: *Estimation of dependencies based on empirical data.* New York: Springer, 1982.

[69] Vapnik, V.N.: *The Nature of Statistical Learning Theory.* New York: Springer, 1995.

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

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

[72] Wolpert, D.H.: Bayesian backpropagation over I–O functions rather than weights. In *Advances in Neural Information Processing Systems 6.* Cowan, J.D., Tesauro, G., & Alspector, J.(Eds.), 200–207, Morgan Kaufmann, San Francisco, CA, 1994.

[73] Wolpert, D.H.: The Relationship Between PAC, the Statistical Physics Framework, the Bayesian Framework, and the VC Framework. In: Wolpert, D. (Ed.) The Mathematics of Generalization. SFI Proceedings Volume XX, Addison–Wesley, 1995.

[74] Wolpert, D.H.: The Lack of A Priori Distinctions between Learning Algorithms. *Neural Computation* **8** (7), 1341–1390.

[75] Yuille, A.L. & Kosowski, J.J.: Statistical Physics Algorithm That Converge. *Neural Computation* **6** (3), 341–356, 1994.

[76] Zhu, S.C. & Mumford, D.: Prior Learning and Gibbs Reaction–Diffusion. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **19** (11), 1236–1250, 1997.

[77] Zinn–Justin, J.: *Quantum Field Theory and Critical Phenomena.* Oxford Science Publications, 1989.