98-48

In : Proceedings of the International Workshop on Advanced Black-Box Techniques for Nonlinear Modeling. Leuven, Belgium, Jul. 1998, pp. 55-61.

Bayesian Learning and the Fokker-Planck Machine

Herman Verrelst, Johan Suykens, Joos Vandewalle, Bart De Moor
Electrical Engineering Department, Katholieke Universiteit Leuven, Belgium

***Abstract*-In this paper we discuss the links between simulated annealing, Langevin optimisation, Markov Chain Monte Carlo methods and ensemble learning on the one hand and the Fokker-Planck machine on the other hand. The FP Machine was introduced as an alternative method for global optimization. Where most optimisation methods follow point-trajectories through search space (either deterministic or stochastic), the FP Machine explicitly describes the density evolution of Markov processes. We will show that if the cost function that is inserted in to FP Machine, is derived from the Bayesian posterior distribution, the FP Machine can be seen as an alternative and extended method for ensemble learning, with its use in Bayesian marginalisation and global optimisation.**

I. Introduction

A lot of attention has gone to Bayesian approaches to the training of neural networks over the last years. Thinking in terms of a (posterior) distribution $p(\mathrm{w}|\mathrm{D})$ (based on data D and prior knowledge) over the possible parameter-choices w, has shown its useful application in many regression and classification problems [4, 14].

One of the challenges in Bayesian training methods, is the multi-modality (i.e. multiple local maxima) of the posterior distribution $p(\mathrm{w}|\mathrm{D})$ or equivalently the local minima in the cost-function $U(\mathrm{w}) = -\log(p(\mathrm{w}|\mathrm{D}))$ derived from this posterior. Sometimes theoretical considerations of the specific problem at hand, show that the effect of the multiple modes is not significant and can therefore be neglected. One focuses on *the* mode and bases the Bayesian inference/analysis entirely on local single Gaussian approximations [14, 15, 3]. Looking at it from the other side, it is known that in many cases the multi-modality of the posterior can have serious influence. So one has to look for methods that correctly incorporate or eliminate this influence, as for example Markov Chain Monte Carlo (MCMC) methods, that have the drawback of being computationally more expensive [16, 17].

In this paper we describe the principle of the Fokker-Planck Machine (FPM) [21, 22, 23] and show how this method can be seen as a combination of the ensemble learning approach and advanced stochastic methods to search the parameter-space. The FPM is defined on the Langevin Markov process, but instead of simulating the point-trajectory of the process, the FPM tracks the position-probability by dynamically approximating it with a (multi-modal) Gaussian mixture model. As in the study of MCMC methods, it can be shown that the probability density evolves towards a Gibbs-distribution, which can be tailored to be just the posterior distribution $p(\mathrm{w}|\mathrm{D})$ when the temperature of the Markov process goes to $T = 1$.

This paper is organised as follows. In Section II we give a small account on the basic concepts behind Bayesian techniques: the likelihood function and the incorporation of prior knowledge via Bayes rule. In Section III two possible ways to make use of the resulting posterior distribution over parameter space, namely numerical optimisation and marginalisation are given. We briefly describe the most important concepts behind annealing algorithms, ensemble learning and MCMC methods, to be able to relate them to the Fokker-Planck Machine. In Section IV the basic algorithm of the FPM is given together with two main technical improvements, that allow an efficient application of the FPM to high-dimensional distributions. In Section V we discuss a bit more in depth the links between the FPM and the existing methods and we conclude in Section VI.

II. Likelihood and Bayes rule

In this section we briefly introduce the basic concepts behind the maximum likelihood approach, its extension to prior knowledge via Bayes' rule to produce the posterior distribution over search space.

A. Setting

Say we are given a data set consisting of N_D input-output pairs

$$\mathrm{D} = \{(\mathrm{x}_k, \mathrm{y}_k)\}_{k=1}^{N_D}$$

where $\mathrm{x} \in \mathbb{R}^m$ and $\mathrm{y} \in \mathbb{R}^l$. We now try to model the functional dependence $\mathrm{y} = f(\mathrm{x})$ that we assume to exist, by another function $g(\mathrm{x}|\mathrm{w})$. The notation says, that g belongs to a model class of functions depending on x that are parameterised by the parameter vector w. As an example, consider the model class of one-hidden-layer feedforward neural networks, with n_h

Bayesian Learning and the Fokker-Planck Machine

Herman Verrelst, Johan Suykens, Joos Vandewalle, Bart De Moor
Electrical Engineering Department, Katholieke Universiteit Leuven, Belgium

Abstract-In this paper we discuss the links between simulated annealing, Langevin optimisation, Markov Chain Monte Carlo methods and ensemble learning on the one hand and the Fokker-Planck machine on the other hand. The FP Machine was introduced as an alternative method for global optimization. Where most optimisation methods follow point-trajectories through search space (either deterministic or stochastic), the FP Machine explicitly describes the density evolution of Markov processes. We will show that if the cost function that is inserted in to FP Machine, is derived from the Bayesian posterior distribution, the FP Machine can be seen as an alternative and extended method for ensemble learning, with its use in Bayesian marginalisation and global optimisation.

I. Introduction

A lot of attention has gone to Bayesian approaches to the training of neural networks over the last years. Thinking in terms of a (posterior) distribution $p(\mathbf{w}|\mathbf{D})$ (based on data D and prior knowledge) over the possible parameter-choices $\mathbf{w}$, has shown its useful application in many regression and classification problems [4, 14].

One of the challenges in Bayesian training methods, is the multi-modality (i.e. multiple local maxima) of the posterior distribution $p(\mathbf{w}|\mathbf{D})$ or equivalently the local minima in the cost-function $U(\mathbf{w}) = -\log(p(\mathbf{w}|\mathbf{D}))$ derived from this posterior. Sometimes theoretical considerations of the specific problem at hand, show that the effect of the multiple modes is not significant and can therefore be neglected. One focuses on *the* mode and bases the Bayesian inference/analysis entirely on local single Gaussian approximations [14, 15, 3]. Looking at it from the other side, it is known that in many cases the multi-modality of the posterior can have serious influence. So one has to look for methods that correctly incorporate or eliminate this influence, as for example Markov Chain Monte Carlo (MCMC) methods, that have the drawback of being computationally more expensive [16, 17].

In this paper we describe the principle of the Fokker-Planck Machine (FPM) [21, 22, 23] and show how this method can be seen as a combination of the ensemble learning approach and advanced stochastic methods to search the parameter-space. The FPM is defined on the Langevin Markov process, but instead of simulating the point-trajectory of the process, the FPM tracks the position-probability by dynamically approximating it with a (multi-modal) Gaussian mixture model. As in the study of MCMC methods, it can be shown that the probability density evolves towards a Gibbs-distribution, which can be tailored to be just the posterior distribution $p(\mathbf{w}|\mathbf{D})$ when the temperature of the Markov process goes to $T = 1$.

This paper is organised as follows. In Section II we give a small account on the basic concepts behind Bayesian techniques: the likelihood function and the incorporation of prior knowledge via Bayes rule. In Section III two possible ways to make use of the resulting posterior distribution over parameter space, namely numerical optimisation and marginalisation are given. We briefly describe the most important concepts behind annealing algorithms, ensemble learning and MCMC methods, to be able to relate them to the Fokker-Planck Machine. In Section IV the basic algorithm of the FPM is given together with two main technical improvements, that allow an efficient application of the FPM to high-dimensional distributions. In Section V we discuss a bit more in depth the links between the FPM and the existing methods and we conclude in Section VI.

II. Likelihood and Bayes rule

In this section we briefly introduce the basic concepts behind the maximum likelihood approach, its extension to prior knowledge via Bayes' rule to produce the posterior distribution over search space.

A. Setting

Say we are given a data set consisting of N_D input-output pairs

$$\mathbf{D} = \{(\mathbf{x}_k, \mathbf{y}_k)\}_{k=1}^{N_D}$$

where $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^l$. We now try to model the functional dependence $\mathbf{y} = f(\mathbf{x})$ that we assume to exist, by another function $g(\mathbf{x}|\mathbf{w})$. The notation says, that g belongs to a model class of functions depending on $\mathbf{x}$ that are parameterised by the parameter vector $\mathbf{w}$. As an example, consider the model class of one-hidden-layer feedforward neural networks, with n_h

hidden neurons with the tanh(.) as activation function

$$y(\mathbf{x}|\mathbf{w}) = \sum_{j=1}^{n_h} w_j \tanh(\sum_{i=1}^{m} v_{ij} x_i + \beta_j), \qquad (1)$$

parameterised by the vector $\mathbf{w}$ consisting of the network's weights w_j and v_{ij} and bias terms β_j.
The problem statement in (system) identification is then: when given only the data $\mathbf{D}$ and only partial or no knowledge about $f(.)$, find $\mathbf{w}^*$, the parameter vector that gives the function $g(.)$ that is 'closest' to $f(.)$. (We make abstraction of the norm in which this 'closeness' is defined.)

B. Likelihood

Assume we are also given that there is noise on the data. This means that even knowing the 'correct' parameter vector $\mathbf{w}^*$, the relation

$$\mathbf{y}_k = g(\mathbf{x}_k|\mathbf{w}^*) \qquad (2)$$

is generally not satisfied for any $k = 1 \ldots N$.
A way to overcome this is to allow additive corrections ϵ_k^y and ϵ_k^x to the 'real' output vector $\hat{\mathbf{y}}$ and the input vector $\hat{\mathbf{x}}$ at time k, such that (2) is satisfied. This method of correcting measurements such that they would fit in a model, does not require the model to be the 'correct' one, so one could equally put an arbitrary parameter vector $\mathbf{w}$ in equation (2). It is assumed however that the corrections needed will be bigger (and thus less likely given our noise assumptions), the further $\mathbf{w}$ is from $\mathbf{w}^*$.
In order to describe how big and thus how (un-)likely the corrections are, the likelihood L is then defined as the probability of actually observing the corrections needed. If for example we know that the possible errors are independent (between $\mathbf{x}$ and $\mathbf{y}$ and between different times k), that they are also identically distributed over the different observation times k (together the so-called 'i.i.d' property) and that the probability distributions of the errors are $p_x(\epsilon^x)$ and $p_y(\epsilon^x)$, the likelihood becomes

$$L(\mathbf{w}) = p(\mathbf{D}|\mathbf{w}) = \prod_{k=1}^{N_D} p_x(\epsilon_k^x) \prod_{k=1}^{N_D} p_y(\epsilon_k^y).$$

The higher the likelihood, the better the data fit into the model with its assumed parameter vector. As the likelihood is dependent on the choice of the parameters, we can use it to find that parameter vector for which the corrections needed are smallest, i.e. for which the data-fit is best. This is called the maximum likelihood method.

C. Bayes rule

The likelihood function $L = p(\mathbf{D}|\mathbf{w})$ is solely based on the data. Sometimes we have other knowledge to determine or to favour some of the parameters in the model. For example in the use of neural networks for classification tasks, big weights bring the neurons in saturation, which is often associated with overfitting. Therefore small weights are favoured. A way to express this (a priori) knowledge is through the definition of a prior distribution $p(\mathbf{w})$ on the parameters in the model. The prior and the likelihood can be combined with Bayes rule

$$p(\mathbf{D}, \mathbf{w}) = p(\mathbf{D}|\mathbf{w})\ p(\mathbf{w}) = p(\mathbf{w}|\mathbf{D})\ p(\mathbf{D})$$

to give the posterior probability (i.e. goodness) of the parameters

$$\begin{aligned} p(\mathbf{w}|\mathbf{D}) &= \frac{p(\mathbf{D}|\mathbf{w})\ p(\mathbf{w})}{p(\mathbf{D})} \\ &= \frac{p(\mathbf{D}|\mathbf{w})\ p(\mathbf{w})}{\int_{\mathbb{R}^d} p(\mathbf{D}|\mathbf{w})\ p(\mathbf{w}) d\mathbf{w}} \qquad (3) \\ &= C \times L(\mathbf{w}) \times \text{prior}. \end{aligned}$$

If there is no prior knowledge at hand, we can define the prior to be uniform over search space. The posterior is than the normalised likelihood function. In the sequel we will make abstraction of the specific choice for the prior and will only speak about the posterior.

III. Uses of the posterior distribution

A. Optimisation

The optimal set of parameters $\mathbf{w}$ maximises the a posteriori probability (MAP) or equivalently minimises the negative logarithm of the posterior $U(\mathbf{w}) = -\log p(\mathbf{w}|\mathbf{D})$,

$$\begin{aligned} \mathbf{w}^* &= \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{D}) \\ &= \arg\min_{\mathbf{w}} U(\mathbf{w}). \qquad (4) \end{aligned}$$

This is the typical relationship between the statistical setting on the one hand in which we have notions like probability, likelihood, priors and posteriors and the numerical minimisation setting in which we have the basic notion of the cost-function (see also Figure 1).

Typical gradient numerical optimisation methods are steepest gradient descent - mostly implemented as the backpropagation algorithm -, (Quasi-)Newton methods and more advanced methods like Levenberg-

Marquardt and conjugate gradient [6, 10].

Scheme 1 **Metropolis Hastings**

A. initialise temperature $T = T_0$, state $\mathbf{w} = \mathbf{w}_0$ and cost $U = U(\mathbf{w}_0)$
B. generate new candidate ξ and the corresponding U_ξ
C. **if** $\exp(-\frac{U_\xi - U}{T}) > \varrho \in (0,1)$ with $p(\varrho) = 1$ (i.e. uniform) **then** $\mathbf{w} = \xi$ and $\mathbf{w} = f_\xi$
D. decrease T
E. **if** $T > T_{\text{end}}$ **then** go to B

Examples of stochastic optimisation methods are simulated annealing [12, 18], where steps in search space are defined by a random walk Markov Process. This is mostly implemented as the Metropolis-Hastings algorithm, where the probability of accepting the candidate steps is dictated by the difference in cost function and a time-varying temperature T, see also Scheme 1. In Langevin annealing the random walk process is 'corrected' by gradient information, such that the search process is more efficient. Langevin annealing will be described in more detail in Section IV.

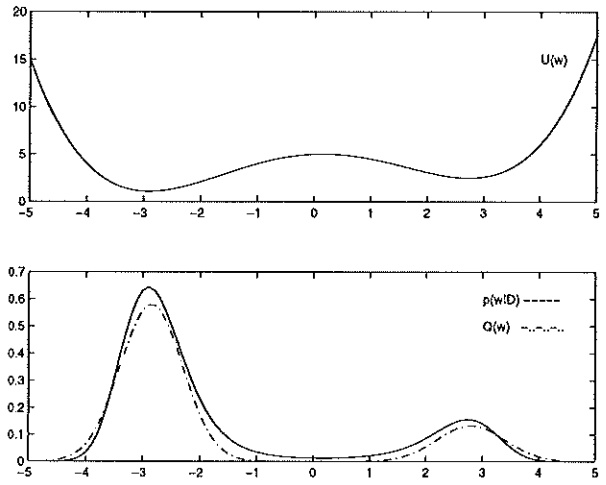


Figure 1: (a) Example of a bi-modal, one-dimensional cost-function $U(\mathbf{w}) = -\log(p(\mathbf{w}|\mathbf{D}))$ and (b) the approximation of the posterior by $Q(\mathbf{w};\theta)$, a sum of two Gaussians.

B. Marginalisation

The posterior distribution can also be used for what is called marginalisation over parameter space. In this last approach, the posterior is seen as a weighting that emphasises those model parameters that optimally fit both data and prior knowledge. The model itself is then not used with a single selected set of model parameters. On the contrary, all parameter combinations are implicitly taken into account in the marginalisation approach. If we for example have a model $\mathbf{y} = g(\mathbf{x}|\mathbf{w})$ for an input-to-output mapping and a posterior distribution $p(\mathbf{w}|\mathbf{D})$, based on training data, the Bayesian model prediction for a 'new' input $\mathbf{x}$ would be the average output

$$\overline{\mathbf{y}} = <\mathbf{y}(\mathbf{x}|\mathbf{w})>_{\mathbf{w}} = \int_{\mathbb{R}^d} g(\mathbf{x}|\mathbf{w}) p(\mathbf{w}|\mathbf{D}) d\mathbf{w}. \quad (5)$$

B.1. Markov Chain Monte Carlo

In most cases this integral can not be calculated analytically, so it must be approximated somehow. Due to the high dimensionality of the posterior, numerical integration based on a regular grid in search space is not feasible. The Monte Carlo approach is to specify (5) as the mean of a random process [16, 19]. An often used random process is a Markov Chain process. By carefully designing the transition probability of such processes (taking into account properties like detailed balance and ergodicity), their equilibrium distribution can be tailored to be the posterior, as studied by Neal in [16, 17]. Also the Metropolis-Hastings algorithm can be used to define such a Markov Chain. From the theory of stochastic differential equations we know that for any temperature T the state probability of this process wil evolve to the Gibbs distribution $\frac{1}{Z}\exp[-U(\mathbf{w})/T]$, with $Z = \int_{\mathbb{R}^d}\exp[-U(\mathbf{w})/T]d\mathbf{w}$. In simulated annealing the smoothing effect of the temperature T is important: for high temperatures, the approximation will be facilitated as many modes will be smoothed out and by decreasing the temperature, the modes will be more and more expressed. For MCMC-purposes we are only interested in $T = 1$ where the equilibrium distribution coincides with the true posterior $p(\mathbf{w}|\mathbf{D})$. The state-points of the random process $\mathbf{w}_k$ will then be samples from this posterior and so

$$\overline{\mathbf{y}} \approx \frac{1}{N}\sum_{k=1}^{k=N} g(\mathbf{x}|\mathbf{w}_k).$$

B.2. Ensemble learning

Another way to approximate $p(\mathbf{w}|\mathbf{D})$ is via parametric density estimation. Here the aim is now to build an explicit density approximation to the posterior via a parametric model (illustrated in Figure 1,b). The integral (5) is then taken over this approximation or via a sample derived from this approximation.
In Laplace's method the posterior is approximated by a Gaussian centred at a mode of $p(\mathbf{w}|\mathbf{D})$ - which can for example be found by a local optimisation algorithm - where the covariance matrix of the Gaussian is derived from the curvature of the posterior at the mode. This method was studied in depth by MacKay [15] and

Bishop [4]. The more general approach was introduced by Hinton and Van Camp in [11]. They discussed the idea of approximating $p(\mathbf{w}|\mathbf{D})$ by an ensemble $Q(\mathbf{w};\theta)$ (a probability density parameterised by θ) and by optimising the quality of this approximation. A classically used measure for the 'distance' between two densities $Q(\mathbf{w};\theta)$ and $p(\mathbf{w}|\mathbf{D})$ is the Kullback-Leibler divergence (or cross-entropy)

$$\begin{aligned} K(Q,p) &= \int_{\mathbb{R}^d} Q(\mathbf{w};\theta) \log \tfrac{Q(\mathbf{w};\theta)}{p(\mathbf{w}|\mathbf{D})} d\mathbf{w} \\ &= \int_{\mathbb{R}^d} Q(\mathbf{w};\theta) \log Q(\mathbf{w};\theta) d\mathbf{w} \\ &\quad - \int_{\mathbb{R}^d} Q(\mathbf{w};\theta) \log p(\mathbf{w}|\mathbf{D}) d\mathbf{w} \\ &= S(\theta) + \int_{\mathbb{R}^d} Q(\mathbf{w};\theta) U(\mathbf{w}) d\mathbf{w}, \end{aligned} \tag{6}$$

with $S(\theta)$ the entropy of the approximating ensemble. Hinton and van Camp considered a single Gaussian approximation with diagonal covariance matrix (i.e. a separable Gaussian). Barber and Bishop [3] discuss the extension to single Gaussians with full covariance matrices. The major reason for using only a single Gaussian, is the problem of calculating the entropy $S(\theta)$ of $Q(\mathbf{w};\theta)$ and gradients of the KL-divergence.
In Section IV we will show how the Fokker-Planck Machine makes use of a mixture of Gaussians to approximate the posterior, how we can use information geometry to analytically calculate the entropy of the Gaussian mixture and how the Fokker-Planck equation based on Langevin equation can provide a learning rule for the parameters θ of the mixture.

IV. The Fokker-Planck Machine

The Fokker-Planck Machine was introduced by Suykens et al. [21, 22, 23] as an alternative optimisation method for the unconstrained optimisation problem

$$\min_{\mathbf{w}\in\mathbb{R}^d} U(\mathbf{w}), \tag{7}$$

with $U(\mathbf{w})$ a twice differentiable cost function.
It was originally defined on the recursive stochastic algorithm (also studied in [7, 8, 13]), associated with the following Langevin-type Markov diffusion

$$d\mathbf{w}(t) = -\nabla U[\mathbf{w}(t)]\, dt + \tau(t)\, dv(t), \tag{8}$$

with state vector $\mathbf{w} \in \mathbb{R}^d$, $v \in \mathbb{R}^d$ a Wiener process, noise intensity $\tau(t)$ and cooling schedule $\tau^2(t) = \tau_0/\log(t)$ (for t large) with τ_0 a fixed positive real constant. This has been called continuous simulated annealing in [7, 8]. In the context of (global) optimization this method can be interpreted to follow a point trajectory through search space, corresponding to a steepest descent method with additive noise.
In [21, 22, 23] it has been interpreted as a special case of the nonlinear stochastic differential equation

$$d\mathbf{w} = f(\mathbf{w},t)\, dt + \tau(\mathbf{w},t)\, dv \tag{9}$$

with state vector $\mathbf{w} \in \mathbb{R}^d$, $v \in \mathbb{R}^{d_v}$ a Wiener process and $f : \mathbb{R}^d \times \mathbb{R} \mapsto \mathbb{R}^d$, $\tau : \mathbb{R}^d \times \mathbb{R} \mapsto \mathbb{R}^{d\times d_v}$. The conditional transition density $p(\mathbf{w},t|\mathbf{w}_0,t_0)$ satisfies the Fokker-Planck equation [5, 9, 25, 26]

$$\frac{\partial p}{\partial t} = -\sum_i \frac{\partial}{\partial w_i}(p f_i) + \frac{1}{2}\sum_i\sum_j \frac{\partial^2}{\partial w_i \partial w_j}(p\beta_{ij}) \tag{10}$$

where $\beta(\mathbf{w},t) = \tau(\mathbf{w},t)\tau(\mathbf{w},t)^T$ and $p(\mathbf{w},t|\mathbf{w}_0,t_0)$ denotes the probability density of being in state $\mathbf{w}$ at time t, given the process is in state $\mathbf{w}_0$ at time t_0. Continuous simulated annealing is a special case of (3) with $f(\mathbf{w},t) = -\nabla U(\mathbf{w})$, $d_v = d$ and $\tau(\mathbf{w},t)$ the diagonal matrix $\tau(t)I_d$ leading to the Fokker-Planck equation

$$\frac{\partial p}{\partial t} = \sum_i \frac{\partial U}{\partial w_i}\frac{\partial p}{\partial w_i} + \sum_i \frac{\partial^2 U}{\partial w_i^2} p + \frac{1}{2}\sigma^2(t) \sum_i \frac{\partial^2 p}{\partial w_i^2}. \tag{11}$$

with $p(\mathbf{w},t|\mathbf{w}_0,t_0)$ the conditional transition probability density and $\frac{\sigma^2}{2} = T$.
Instead of simulating the Markov Chain explicitly, the Fokker-Planck Machine now tracks the evolution through time of the state (transition) probability. (The term 'transition' can be dropped as in the limit of $t \to \infty$, the state probability gets independent of the initialisation.) This tracking is done by approximating the state probability by a Gaussian mixture (a Radial Basis Function RBF neural network),

$$Q(\mathbf{w};\theta) = \sum_{i=1}^{n_h} \pi_i(t)\, N(\mathbf{w}|\mu_i(t),\Sigma_i(t)), \tag{12}$$

$$\begin{aligned} \text{with} \quad \sum_{i=1}^{n_h} \pi_i(t) &= 1, \\ N(\mathbf{w}|\mu,\Sigma) &= \tfrac{(2\pi)^{\frac{d}{2}}}{|\Sigma|^{1/2}} \exp[-\tfrac{1}{2}(\mathbf{w}-\mu)^T\Sigma^{-1}(\mathbf{w}-\mu)] \end{aligned}$$

and n_h the number of Gaussians in the mixture and with the paramters of the RBF network (denoted by θ): π_i, $\mu_i \in \mathbb{R}^d$ and $\Sigma_i \in \mathbb{R}^{d\times d}$, respectively the weights, the centres and the covariance matrices of the Gaussians. The basic idea of the FP Machine is to project the FP partial differential equation onto ordinary differential equations on the RBF parameters and to update the parameters of the RBF network, according to this projection.
It has been proposed then to do this by evaluating the Fokker-Planck equation in $Q(\mathbf{w};\theta)$ at N points, which yields a constrained linear set of equations

$$A(\theta)\,\dot{\theta} = b(\theta) \quad \text{such that} \quad c^T\dot{\theta} = 0 \tag{13}$$

with $q = n_h + 2dn_h$ and $A \in \mathbb{R}^{N\times q}$, $b \in \mathbb{R}^N$ and $c \in \mathbb{R}^q$ [21, 23]. The constraint follows from the property $\sum_i \pi_i = 1$. It has been assumed that $N > q$. As

a solution of this overdetermined set of equations one finds an estimate for the time-derivative of the Gaussian mixture model's parameter vector. The matrix A and vector b contain information about the local geometry (gradient and diagonal elements of the Hessian) at each sampling point.
The Fokker-Planck machine method makes use then of a population of points, generated in search space according to the Gaussian mixture model (6). This is done in a sequence of generations, where the parameter vector of the mixture model is updated after each generation. The basic algorithm of the Fokker-Planck machine looks as follows:

1. *First generation: choose initial θ of Gaussian mixture distribution.*
2. *Generate N points according to $Q(\mathrm{w};\theta)$.*
3. *Calculate $\frac{\partial U}{\partial w_i}$, $\frac{\partial^2 U}{\partial w_i^2}$ at the N points.*
4. *Estimate $\dot{\theta}$ from the constrained linear least squares problem.*
5. *Next generation: compute $u(t+\Delta t) := \theta(t) + \dot{\theta}\,\Delta t$.*
6. *Update the noise intensity σ, according to the cooling schedule.*
7. *Go to 2, unless n_g is exceeded.*

Here σ_0, n_h, N, the number of generations n_g, the initial θ and the step size Δt serve as input parameters for the algorithm.

A. On-line estimation

In this Section we focus on the computation of $v \triangleq \dot{\theta}$ from the constrained linear least squares problem. The analytic solution in regularised linear least squares sense to

$$\min_v \|Av - b\|_2^2 + \|\Lambda^{1/2} v\|_2^2 \quad \text{such that} \quad c^T v = 0 \tag{14}$$

is given by

$$\begin{cases} v &= (A^T A + \Lambda)^{-1}(A^T b - \lambda c) \\ \lambda &= \frac{c^T (A^T A+\Lambda)^{-1} A^T b}{c^T (A^T A+\Lambda)^{-1} c} \end{cases} \tag{15}$$

where Λ is a diagonal matrix, which is often chosen in terms of fewer regularisation parameters δ_π, δ_β, δ_μ as $\Lambda = \mathrm{blockdiag}\{\delta_\pi I, \delta_\beta I, \delta_\mu I\}$. Because the matrix $A^T A + \Lambda$ is of dimension $q \times q$ this method is only applicable to lower dimensional cost function (typically $d < 40$ for Gaussian mixture distributions with a few hidden units).
In order to make the Fokker-Planck machine applicable to larger scale problems, an on-line learning scheme, which is based on a constrained regularised normalised LMS method, has been proposed in [23]. This method works with vector updates thereby avoiding the storage of huge matrices. It corresponds to the minimising solution of

$$\min_{v_{k+1}} \quad \alpha^2 \|v_{k+1} - v_k\|_2^2 + (b_k - a_k^T v_{k+1})^2 + \|\Lambda^{1/2} v_{k+1}\|_2^2 \quad \text{such that} \quad c^T v_{k+1} = 0 \tag{16}$$

with α a real constant, a_k the k-th row of A, b_k the k-th element of b and v_k the k-th iterate for the estimation of v.

B. Information geometry

The Gaussian mixture model can be written as a member of the exponential family $S = \{p(r;\Theta)\}$ with

$$p(r;\Theta) = \exp[\Theta^T r - \psi(\Theta)] \tag{17}$$

where r is a new vector random variable. For the scalar case ($n = 1$) of a random variable such a representation has been derived by Amari in [2]. In [23] we extended the results for $\mathrm{w} \in \mathbb{R}^d$. We denote the Gaussian mixture distribution as

$$Q(\mathrm{w};\theta) = \sum_{i=1}^{n_h} \frac{\pi_i \prod_{j=1}^{d} \beta_{ij}}{(2\pi)^{d/2}} \exp[-\sum_{j=1}^{d} \frac{\beta_{ij}^2}{2}(w_j - \mu_{ij})^2] \tag{18}$$

which is parameterised by the parameter vector θ. (We consider only diagonal covariance matrices.) According to Amari [2], this is considered as the marginal density of

$$p(\mathrm{w}, z;\theta) = \sum_{i=1}^{n_h} \frac{\pi_i\, \delta_i(z) \prod_{j}^{d} \beta_{ij}}{(2\pi)^{d/2}} \exp[-\sum_{j}^{d} \frac{\beta_{ij}^2}{2}(w_j - \mu_{ij})^2] \tag{19}$$

where z is a discrete random variable taking values on $\{1, 2, ..., n_h\}$ with probabilities $\pi_i = \mathrm{Prob}\{z = i\}$ for $i = 1, ..., n_h$ and $\sum_{i=1}^{n_h} \delta_i(z) = 1$. By calculating $\log p(\mathrm{w}, z)$ the exponential family

$$p(\mathrm{w}, z;\Theta) = \exp[\Theta^T r - \psi(\Theta)] \tag{20}$$

is obtained which is parameterised by the new parameter vector $\Theta = [\theta_{1j}; \theta_{2j}; \theta_{3i}; \theta_{4ji}; \theta_{5ji}]$ and depends on $r = [r_{1j}; r_{2j}; r_{3i}; r_{4ji}; r_{5ji}]$ with $r(\mathrm{w}, z)$. Furthermore, one considers the expectation of the random variable r with respect to the distribution $p(r;\Theta)$:

$$\Gamma(\Theta) = E_\Theta[r], \tag{21}$$

which is related to the function $\psi(\Theta)$ by

$$\Gamma(\Theta) = \frac{\partial}{\partial \Theta} \psi(\Theta). \tag{22}$$

The set $\mathcal{S}$ is a dually flat Riemannian manifold , where Θ and Γ play a special dual role [1, 2]. The entropy of the Gaussian mixture distribution can be calculated from Θ and Γ by

$$\begin{aligned} S(\theta) = H(\Theta) &= -\int Q(\mathbf{w};\theta)\log Q(\mathbf{w};\theta)\,d\mathbf{w} \\ &= \psi(\Theta) - \textstyle\sum_i \Theta_i \Gamma_i. \end{aligned} \tag{23}$$

It can be proven that for the case of an exponential family, a geodesic (the minimum length curve connecting two points) corresponds to a straight line. This means that simple line searches in the update direction, result in the shortest connection to the solution. Moreover, for a dually flat manifold it is proven that an invariant divergence

$$K(\Theta_{Q_1}, \Gamma_{Q_2}) = \psi(\Theta_{Q_1}) + \varphi(\Gamma_{Q_2}) - \sum_i \Theta_{Q_1,i}\Gamma_{Q_2,i} \tag{24}$$

is defined between two points $Q_1, Q_2 \in \mathcal{S}$, which corresponds to the Kullback-Leibler divergence

$$K(Q_1, Q_2) = E_{Q_1}\left[\log \frac{Q_1}{Q_1}\right]. \tag{25}$$

This KL-divergence can then be used in a very straightforward way in the line search in the direction of the update rule $\dot{\Theta}$.

V. The Fokker-Planck Machine as dynamic ensemble learning

As described in Section IV, the FPM tracks the evolution of the state probability of the Langevin Markov process. If the same conditions (detailed balance and ergodicity) hold that are needed for proper use in the MCMC-approach, then we will obtain an explicit approximation $Q(\mathbf{w};\theta)$ to the Gibbs distribution [20, 25]. As this Gibbs distribution is equal to the posterior distribution $p(\mathbf{w}|\mathbf{D})$ for $T = 1$, we can interpret the FPM as an alternative method for ensemble annealing.

$$Q(\mathbf{w};\theta) \approx p(\mathbf{w}, t \to \infty) \propto e^{-U(\mathbf{w})/T}.$$

An important aspect to the FPM is however, that it places this in a dynamical context. The update rule for the parameters w in the approximation is not based on the (static) Kullback-Leibler divergence, but on a annealing-like processes in which we have a time-varying temperature/noise level. As in the annealing optimisation algorithms, this temperature has a smoothing effect such that 1) in the initial phase, the approximation is easier and such that 2) by cooling, the local minima can be gradually taken into account. The Fokker-Planck Machine makes use of a mixture of Gaussians, so it can capture multiple modes.
Figure 2 visualises the conceptual links of the FPM - as an algorithm for the dynamical tracking of the state probability of a Markov process - with MCMC, annealing and ensemble learning. The fact that the temperature T appears explicitly in the expression for the equilibrium distribution motivates the term 'deterministic annealing', as also used by Ueda and Nakano in [24].

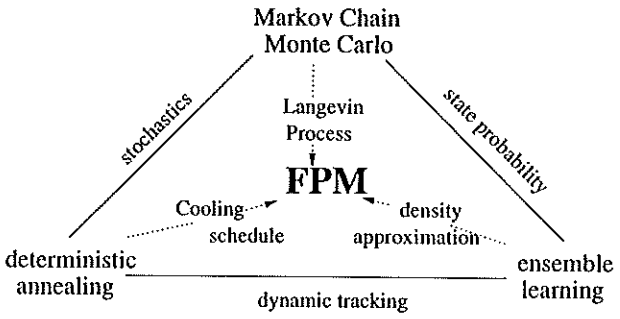


Figure 2: Conceptual links of MCMC, ensemble learning and (deterministic) annealing with FPM.

VI. Conclusions

In this paper we have given a small account on the basic concepts behind Bayesian techniques: the likelihood function, the incorporation of prior knowledge via Bayes rule and two possible ways to make use of the resulting posterior distribution over parameter space, namely numerical optimisation and marginalisation. In this context we described the (simulated and Langevin) annealing algorithm, ensemble learning and MCMC methods and their relation to the Fokker-Planck Machine. We have shown that the FPM can be seen to combine the notions of parametric density estimation, sophisticated stochastic search methods and smoothing temperature effects. We have shown how the FP equation can provide us with an update rule for the parameters of a Gaussian mixture distribution approximation to the posterior distribution. We also discussed ways in which the FPM can be made applicable to higher-dimensional problems and how information geometry provided a new parameterisation. In this parameterisation geodesics are straight lines, which means that simple line searches in the update direction, result in the shortest path to the solution.

Acknowledgements

H. Verrelst is researcher with the IWT, J. Suykens is post-doctoral researcher with the FWO-Flanders, J. Vandewalle is Full Professor at the KULeuven, Bart De Moor is a senior Research Associate with the F.W.O. (Fund for Scientific Research-Flanders). This research work was carried out at the ESAT lab and the ICNN of the KULeuven, in the following frameworks: the Belgian Programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Sci-

ence, Technology and Culture (IUAP P4-02 and IUAP P4-24), a Concerted Action Project MIPS of the Flemish Community, the FWO Research Communities: ICCoS and Advanced Numerical Methods for Mathematical Modelling and the FWO project G.0262.97 : Learning and Optimization: an Interdisciplinary Approach.

References

[1] Amari S.-i., *Differential-geometrical methods in statistics*, Lecture Notes in Statistics, Springer-Verlag, 1985.

[2] Amari S.-i., "Information geometry of the EM and em algorithms for neural networks," *Neural Networks*, Vol.8, No.9, pp.1379-1408, 1995.

[3] Barber D., Bishop C., "Ensemble Learning for Multi-Layer Networks," *Advances in Neural Information Processing Systems 10*, 1998.

[4] Bishop C., *Artificial Neural Networks for Pattern Recognition*, OUP, Oxford, 1996.

[5] Doob J.L., *Stochastic processes*, John Wiley & Sons, 1953.

[6] Fletcher R., *Practical methods of optimization*, second edition, Chichester and New York: John Wiley and Sons, 1987.

[7] Gelfand S.B., Mitter S.K., "Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$," *SIAM Journal on Control and Optimization*, Vol.29, No.5, pp.999-1018, 1991.

[8] Gelfand S.B., Mitter S.K., "Metropolis-type annealing algorithms for global optimization in $\mathbb{R}^d$," *SIAM Journal on Control and Optimization*, Vol.31, No.1, pp.111-131, 1993.

[9] Gihman I.I., Skorohod A.V., *The theory of stochastic processes I,II,III*, New York: Springer-Verlag, 1979.

[10] Gill P.E., Murray W., Wright M.H., *Practical Optimization*, London: Academic Press, 1981.

[11] Hinton G.E., van Camp D., "Keeping neural networks simple by minimizing the description length of the weights," *Proc. of COLT-93*, pp. 5-13, 1993.

[12] Kirkpatrick S., Gelatt C.D., Vecchi M., "Optimization by simulated annealing," *Science*, 220, pp.621-680, 1983.

[13] Kushner H.J., "Asymptotic global behavior for stochastic approximation and diffusions with slowly decreasing noise effects: global minimization via Monte Carlo," *SIAM Journal on Applied Mathematics*, 47, pp.169-185, 1987.

[14] MacKay D., *Information Theory Inference and Learning Algorithms*, Boook in preparation, http://wol.ra.phy.ca.ac.uk/mackay

[15] MacKay D., "A Practical Bayesian framework for back-propagation networks," Neural Computation, Vol. 4, No. 3, pp. 448-472, 1995.

[16] Neal R., "Probabilistic inference using Markov Chain Monte Carlo Methods," Tech. Report CRG-TR-93-1, Dept. of Comp. Sc., Univ. of Toronto, 1993.

[17] Neal R., "Probabilistic inference using Markov Chain Monte Carlo Methods," Tech. Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, 1993

[18] Otten R.H.J.M., van Ginneken L.P.P.P., *The Annealing Algorithm*, Kluwer Academic Publishers, 1989.

[19] Rubenstein R.Y., *Simulation and the Monte Carlo Method*, Wiley series in probability and mathematical statistics, John-Wiley, 1981.

[20] Soize C., *The Fokker-Planck Equation for Stochastic Dynamical Systems and Its Explicit Steady State Solutions*, Series on Advances in Mathematics for Applied Sciences - Vol. 17, World Scientific, 1994.

[21] Suykens J.A.K., Vandewalle J.P.L., De Moor B.L.R., *Artificial neural networks for modelling and control of non-linear systems*, Kluwer Acad. Publ., Boston, 1995.

[22] Suykens J.A.K., Verrelst H., Vandewalle J., "On-line learning Fokker-Planck machine," *Neural processing letters*, Vol.7, No.2, pp.81-89, 1998.

[23] Suykens J.A.K., Verrelst H., Cambre J., Vandewalle J., "Information geometry and the Fokker-Planck machine," submitted for publication.

[24] Ueda N., Nakano R., "Deterministic Annealing EM Algorithm," Neural Networks, Vol. 11, No. 2, pp. 271-282, 1998.

[25] van Kampen N.G., *Stochastic Processes in Physics and Chemistry*, North-Holland, 1981.

[26] Wong E., *Stochastic processes in information and dynamical systems*, McGraw-Hill, 1971.