Bayesian Inference, Quantum field theory and Geometry

David S. Berman
based on work with
Jonathan Heckman and Marc Klinger

to appear arXiv:210X:xxxxx where

$$p(X) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}(\frac{(X-\mu)}{\sigma})^2)\ , \quad p(\mathrm{x}) = ?$$

$$p(\mu) = \frac{1}{\sqrt{2\pi}0.1}\exp(-\frac{1}{2}(\frac{2-\mu}{0.1})^2$$

$$p(\sigma) = \frac{1}{\sqrt{2\pi}0.01}\exp(-\frac{1}{2}(\frac{0.5-\sigma)}{0.01})^2$$

Outline

- Motivation
- Bayes
- Bayesian Updating
- Information Geometry
- The geometry of Bayesian updating
- The Exact renormalisation Group equation and updating
- Discussion

Overview

We want to connect ideas of statistical inference with quantum field theory and in particular focus on something called Bayesian up-dating.

In what follows, the objects of study are probability distributions (in fact often we will consider a family of distributions called a model with some set of parameters on which the distribution depends). Learning is using the data we collect to alter these probablility distributions to conincide as much as possible to some true underlying distribution.

This is quite generaland include: Neural Networks, Gaussian Random Processes, simple regression etc. etc.

Notation and fundamentals

A generic probability distribution for a set of random variables $S = \{X_1, ..., X_n\}$ will be denoted by a P subscript S.

$$P_S(s) = P_{X_1,...,X_n}(x_1, ..., x_n) \tag{1}$$

A joint probability distribution can be marginalized or conditionalized:

$$\text{Let } A \subseteq S \text{ with } A' = S \setminus A. \text{ Then } P_A(a) = \int_{A'} da' P_S(a, a') \tag{2}$$

$$\text{Let } A \subseteq S \text{ with } A' = S \setminus A. \text{ Then } P_{A|A'}(a \mid a') = \frac{P_S(a, a')}{P_{A'}(a')} \tag{3}$$

From the definition of the conditional distribution we have the following string of equalities:

$$P_S(a, a') = P_{A|A'}(a \mid a')P_{A'}(a') = P_{A'|A}(a' \mid a)P_A(a) \tag{4}$$

This is Bayes' Rule.

The Bayesian problem involves two kinds of random variables: Data, denoted by Y, and Parameters of the model denotes by Θ. Denote the true underlying joint probability distribution over the probability space (Θ, Y) with the letter T. That is:

$$T_{\Theta,Y}(\theta, y) \tag{5}$$

We will have a model of the true distribution denoted by

$$M_{\Theta,Y}(\theta, y)\,. \tag{6}$$

Lets have an example before we get lost in abstraction: Take our model of the data in question (eg. my publication date) to be gaussian with mean μ and variance σ but I don't know μ and σ perfectly. These are the parameters of my model θ. Based on experience, I declare a prior distribution for μ and θ and thus state $M(\theta)$.

Statistical Inference

The Bayesian approach is to use Bayes' Rule to contruct a new distribution on the parameters of the model, $M_{\Theta}(\theta, posterior)$, based on the data available y.
This is called the Posterior Distribution, given by:

$$M_{\Theta|Y}(\theta \mid y; post) = \frac{M_{Y|\Theta}(y \mid \theta) M_{\Theta}(\theta)}{\int d\theta M_{Y|\Theta}(y \mid \theta) M_{\Theta}(\theta)} \tag{7}$$

So the so called posterior distribution on the parameters of our model is what we have leant given our date y

Bayesian Updating

We will consider the inclusion of data iteratively and apply Bayes theorem everytime we get more data to change $M_{\Theta,Y}(\theta, y)$. At iteration t the model distribution will be indexed by the iteration number t:

$$M_{\Theta,Y}(\theta, y; t)\,. \tag{8}$$

To make this updating scheme dynamic, we employ the additional identification that the Posterior Distribution in iteration t of the update becomes the Prior Distribution in iteration $t+1$. Explicitly:

$$M_{\Theta}(\theta; t+1) = M_{\Theta|Y}(\theta \mid y; t) \tag{9}$$

KL divergence and the Fisher Metric

As we do updates using this how do we approach the true distribution. A measure of the promixity of two distributions is given by the Kullback-Leibler divergence:

$$D_{KL}(P_1(x)||P_2(x)) = \int dx\, P_1(x) \log\left(\frac{P_1(x)}{P_2(x)}.\right) \tag{10}$$

For the case that the distributions are "near" we can expand the KL divergence

For two models that are parameterically close we can take the Hessian of the KL divergence in the parameters of the model η^i to define the Fisher information metric:

$$g_{ij} = \frac{\partial^2}{\partial \eta^i \partial \eta^j} D_{KL}(P_1(\eta) \parallel P_2(\eta_0)) \tag{11}$$

This metric describes the proximity of distributions and , from Chentsov's theorem it is the unique information metric for statistical models that gives "sufficient statistics" in the Fisher sense.

Note, for a Gaussian probability distribution with moduli $\sigma, \vec{x_0}$:

$$p(\vec{x}; \vec{x}_0, \sigma) = (\pi\sigma^2)^{-(N)/2} exp\left(-\sum_i \frac{(x^i - x_0^i)^2}{\sigma^2}\right)$$

the Fisher information metric is given by:

$$ds^2 = 2\frac{d\sigma^2 + d\vec{x}^0 \cdot d\vec{x}^0}{\sigma^2}$$

which is the Poincare Patch for AdS.

Back to updating

The idea now is to examine updating by looking at a time dependent KL divergence and then take the large data limit so that we can move from difference equations for $M(\theta, t)$ to differential equations.
After some work, we find the following from the update equation:

$$\frac{\partial}{\partial t} \ln(M_{\Theta}(\theta; t)) = D_{KL}(T_{Y|\Theta}(y \mid \theta^{*}) \parallel M_{Y}(y; t)) \qquad (12)$$

Solve with simple example, take the true distribution to be normal, fix the variance but take data to find the mean.Thus

$$M_{Y,\Theta}(y, \alpha(t), \sigma; t) = \mathcal{N}(\alpha(t), \sigma^{2})(y) \qquad (13)$$

and

$$T_{Y,\Theta}(y, \mu, \sigma; t) = \mathcal{N}(\mu, \sigma^{2})(y) \qquad (14)$$

The dynamical equation becomes:

$$\frac{\partial}{\partial t}\left((\alpha(t)-\mu)^2\right) = -\frac{1}{\sigma^2}\left((\alpha(t)-\mu)^2\right) \tag{15}$$

Which we solve by:

$$\alpha(t) = \sqrt{A}\exp(-\frac{1}{2}\frac{t}{\sigma^2}) + \mu\,. \tag{16}$$

This is how the mean of the model as a function of "time" approaches the true mean μ.
A crucial part that made this calculation work so easily is the idea of Bayesian self-conjugacey. This is the where the posterior and prior distributions are in the same model and only the parameters shift. Practically much of Bayes works because the Gaussian is self-conjugate. Just as in QFT where we do Gaussian integrals.

Distance Weighted Probability Models

Define a Probability Distribution as:

$$P_\Theta(\theta \mid \alpha) \propto \exp(-D_\alpha(\theta)) \tag{17}$$

with $D_\alpha(\theta)$ is a distance between θ and α, some reference state. This is very physical, think of:
The Boltzmann Weight: Probability is exponentially weighted as the difference between the energy of a state and some ground state energy:

$$P_E(\epsilon \mid \epsilon_0) \propto \exp(-(\epsilon - \epsilon_0)) \tag{18}$$

e.g. the distance function is on the space of energetics.

We take this idea and construct a distance wieghted prior distriubtion using the KL divergence as compared to some reference distribution, $P_{Y|\Theta}(y \mid \theta)$:

$$M_{\Theta}(\theta) = f(\theta) e^{-D_{KL}(M_Y(y) \| P_{Y|\Theta}(y|\theta))} \tag{19}$$

working with this prior, the model has a greater possibility for self-conjugacey and we can apply the flow equations on a broader set of problems.
But what is $f(\theta)$? After some work, one sees:

$$f(\theta) = \sqrt{\mathrm{detg}(\theta)}, \tag{20}$$

ie. the measure given the Fisher information.

Using the above parameterisation we revist the normal distribution with non-fixed variance. Instead of using the KL divergence, expand to get the update equation in terms of the Fisher metric. This becomes:

$$\frac{d(g_{ij})}{dt} = -2R[g]_{ij} \tag{21}$$

with $R[g]_{ij}$ the Ricci tensor of the Fisher metric.
This we recognise as the Ricci flow equation.
Cavaet: I used the detailed form of the Fisher metric for the normal distribution to derive the above equation. I do not know about its generality but it is suggestive.

One can tackle multivariate normal distributions in the exactly same way, or even have an infinite number of normals to produce a Gaussian Random Process.
The one constructs a metric functional over the infinte dimensional space of the GRP. The updating would then evolve the GRP Kernel according to the update equation as new data arrives.
As we have seen this can be related to Neural Nets in some limit.

What about neural net training?
Often one describes the training for neural nets using the neural tangent Kernel gradient equation. We describe a neural net as a function, $f(x;\theta)$ with input, x and weights, θ.
Then learning is described by:

$$\frac{df(x_1,\theta)}{dt} = K(x_1,x_2;\theta)\partial_g C(g,y)\Big|_{g=f(x_2,\theta(t))} \tag{22}$$

Where the function $C(f(x;\theta),y)$ is the cost function, usually taken to be least squares.
$K(x_1,x_2;\theta)$ is the neural tangent Kernel given by:

$$K(x_1,x_2;\theta) = \partial_{\theta^i} f(x_1;\theta)\partial_{\theta_i} f(x_2;\theta) \tag{23}$$

This bares resemblence to the Bayes update equation but it is certainly not the same. However, it is suggestive, if instead of using the least squares cost one took the the cost function to be the exponential of the KL divergence and if $f(x, \theta)$ takes on a particular form then the update equations may be related. (Note for a Restricted Boltzman Machine-NN, the KL divergence is the cost function).
Note, that taking the "NTK scaling" for a NN of infinite width, reproduces the results of the Bayes answer as one ends up on the maximum a posteriori estimate given a Gaussian prior on functions- the result of [Jacot, Gabriel and Hongler].

QFT

We have a probability distribution for fields given by:

$$P_\Phi[\phi] = \frac{1}{Z} e^{-S_E[\phi]} \tag{24}$$

and a partition function as follows

$$Z = \int \mathcal{D}\phi e^{-S_E[\phi]} \tag{25}$$

to calculate moments (correlation functions) it is useful to define a generating function:

$$Z[\mathcal{J}] = \int \mathcal{D}\phi e^{-S_E[\phi]} e^{\sum\limits_{\Phi} \int J_\phi \wedge \phi} \tag{26}$$

We may also define the effective field:

$$\varphi(x) = \langle \phi(x) \rangle = \frac{\delta \ln(Z[\mathcal{J}])}{\delta J_\phi(x)} \tag{27}$$

Schematically, the generating functional has the following form:

$$Z[\mathcal{J}] = e^{\sum_\Phi \frac{1}{2} \int d^d x d^d x' J_\phi(x) G_\phi(x-x') J_\phi(x')} \tag{28}$$

Where $G_\phi(x - x')$ is the Green's Function and then we can write

$$\varphi(x) = \frac{\delta \ln(Z[\mathcal{J}])}{\delta J_\phi(x)} = \int d^d x' G_\phi(x - x') J_\phi(x') \tag{29}$$

Often, we choose some energy scale, e.g. by defining a set of fields with a given range of momenta – $A = \{\phi \in \Phi \mid p_\phi^2 > k^2\}$). Then, we integrate only over fields in that set. This defines an effective Generating Functional at scale given by k:

$$Z_k[\mathcal{J}] = \int \mathcal{D}_A\phi e^{-S_E[\phi]} e^{\sum_\Phi \int J_\phi \wedge \phi} \tag{30}$$

Here, $\mathcal{D}_A\phi$ is indicating we integrate only over fields in the set A. From $Z_k[\mathcal{J}]$ we can define a scale dependent effective action:

$$\Gamma_k[\varphi] = -\ln(Z_k[\mathcal{J}[\varphi]]) \tag{31}$$

We can now ask, how does the effective action depend on the scale. This is the Exact renormalisation group equation:
Define the parameter $t = \ln(k)$, then Polchinski's Exact Renormalisation Group Equation is:

$$\frac{d\Gamma_k}{dt} = -\frac{1}{2}Tr(\frac{\delta^2\Gamma_k}{\delta\varphi\delta\varphi} - \frac{\delta\Gamma_k}{\delta\varphi}\frac{\delta\Gamma_k}{\delta\varphi}) \tag{32}$$

what is happening is that as we change the scale we integrate out over more fields giving us an alterted distribution on the remaining fields.

Now, lets consider this as a joint distribution over the data we have and the data we don't have. In the usual approach we have an energy scale that will provide the split. This is not the case now. Divide the set of all fields, Φ into two subsets $\Phi = Y \cup \Theta$ in a manner which depends on some continuous parameter t that we identify with the Bayesian iteration parameter. Y is the set of fields that form the observable data. We will write the full probability distribution as $P_{Y,\Theta}[y,\theta]$ to signify that we are thinking of it as a joint probability density. Similarly, the complete partition function for the theory will be written as $Z_{Y,\Theta}$.

Then we can write,

$$P_\Theta(t) = \int \mathcal{D}_Y \phi P_{Y,\Theta}[y,\theta] = \frac{1}{Z_{Y,\Theta}} \int \mathcal{D}_Y \phi e^{-S_E[\phi]} e^{\sum_\Phi \int J_\phi \wedge \phi} = \frac{Z_Y(t)}{Z_{Y,\Theta}} \tag{33}$$

and

$$P_{Y|\Theta}[y \mid \theta] = \frac{P_{Y,\Theta}[y,\theta]}{P_\Theta(t)} = \frac{e^{-\hat{S}_E}}{Z_{Y,\Theta}} \frac{Z_{Y,\Theta}}{Z_Y(t)} = \frac{e^{-\hat{S}_E}}{Z_Y(t)} \tag{34}$$

Now we can start to calculate KL divergences between $P_{Y,\Theta}[y,\theta]$ and $P_{\Theta}[\theta;t]$ and take the "time derivative" and follow the update equation from before where now the true distribution is the total joint distribution and the model distribution $M(\theta;t)$ is $P_{\Theta}[\theta;t]$. After various manipulations and some additional assumptions about the details of the distributions then, the update equation looks like:

$$\frac{d(-\ln(Z_Y(t)))}{dt} = -\frac{1}{2}Tr(\frac{\delta^2(-\ln(Z_Y(t)))}{\delta\theta\delta\theta} - \frac{\delta(-\ln(Z_Y(t)))}{\delta\theta}\frac{\delta(-\ln(Z_Y(t}{\delta\theta} \tag{35}$$

Discussion

Quantum Field theory intuition is a good starting point to understand learning. We parameterise our ignorance and see how probabilty distributions change as we have more data.
A key difference is that we do not have energy, so there is not the same notion of a Wilsonian effectve action. But for distance weighted probability models we have an ordering based on likelyhood given by the distance. The low energy effective action is then the action for the most likely configurations.
Much, much more to do...

What goes wrong?
In all of this there was an implicit assumption about the Hessian of the KL divergence being positive definite. This is true for the same models parametrically seperated eg the normal model.
However, importantly it is not true in general. If our model is not the same as the true distribution then we can have negative modes of the Hessian of the KL divergence. This leads to “unlearning” and perturbations drive the distribution away from the true distiubtion. We have examples of this.