ICES REPORT 15-04

February 2015

Bayes is Optimal

by

Tan Bui-Thanh and Omar Ghattas

The Institute for Computational Engineering and Sciences
The University of Texas at Austin
Austin, Texas 78712

Reference: Tan Bui-Thanh and Omar Ghattas, "Bayes is Optimal," ICES REPORT 15-04, The Institute for Computational Engineering and Sciences, The University of Texas at Austin, February 2015.

Bayes is Optimal!

Tan Bui-Thanh
Department of Aerospace Engineering and Engineering Mechanics
Institute for Computational Engineering & Sciences
The University of Texas at Austin, Austin, TX 78712, USA.

Omar Ghattas
Institute for Computational Engineering & Sciences
Jackson School of Geosciences
Department of Mechanical Engineering
The University of Texas at Austin, Austin, TX 78712, USA

December 27, 2013

In this short note we construct a convex optimization problem whose first order optimality condition is exactly the Bayes' formula and whose unique solution is precisely the posterior distribution. In fact, the solution of our optimization problem includes the usual Bayes' posterior as a special case and it is therefore more general. We provide the construction, and hence a generalized Bayes' formula, for both finite and infinite dimensional settings. We shall show that the our posterior distribution, and the Bayes' one as a special case, is optimal in the sense that it is the unique minimizer of an objective function. We provide the detailed and constructive derivation of the objective function using information theory and optimization technique. In particular, the objective is the compromise of two quantities: 1) the relative entropy between the posterior and the prior, and 2) the mean squared error between the computer model and the observation data. As shall be shown, our posterior minimizes these two quantities simultaneously.

1 Finite dimensional case

In this note we exclusively tackle statistical inverse problems in which the task at hand is to combine prior distribution and observation data to come up with a "better" distribution of some parameter $\mathbf{m}$ in a parameter space $\mathcal{M}$. In other words, we have (or are provided with) some prior knowledge and we would like to update our knowledge as soon as the (noise-corrupted) data is available. The Bayes' framework provides a solution to such a problem. In this framework, we seek a statistical description of all possible parameters that conform to some prior knowledge and at the same time are consistent with the observation data. These parameters are distributed according to the so-called posterior distribution. We ask ourselves if the Bayes paradigm is the best approach, if so, in which sense? In this report, we provide an answer to this question.

To begin, let us introduce some notations. We denote by $\mathcal{G}(\mathbf{m})$ the computer prediction/model of some finite dimensional output of interest. For simplicity, our computer model is assumed to be exact, i.e. it is adequate, and the noise is Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{C})$ with $\mathbf{C}$ denoting the covariance matrix. Under additive noise assumption, the observation data is given by

$$\mathbf{d} := \mathcal{G}(\mathbf{m}) + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$. Note that additive Gaussian noise is not a limitation. In fact it is simply used to motivate our approach. As will be seen, the resulting formulation will be valid for any likelihood model. We assume that the prior is distributed by $\pi_{\text{prior}}(\mathbf{m})$ and we seek an updated distribution $\pi(\mathbf{m})$ to incorporate information from the prior and the data. Bayes' formula is one way to accomplish this task. It says that the posterior distribution (the updated distribution) is given by

$$\pi_{\text{post}}(\mathbf{m}) := \pi_{\text{post}}(\mathbf{m}|\mathbf{d}) := \frac{\pi_{\text{like}}(\mathbf{d}|\mathbf{m}) \times \pi_{\text{prior}}(\mathbf{m})}{\int_{\mathcal{M}} \pi_{\text{like}}(\mathbf{d}|\mathbf{m}) \times \pi_{\text{prior}}(\mathbf{m})\ d\mathbf{m}}, \tag{2}$$

where, for the above assumption, the likelihood is

$$\pi_{\text{like}}(\mathbf{d}|\mathbf{m}) = \exp\left(-\frac{1}{2}\left\|\mathcal{G}(m) - \mathbf{d}\right\|_{\mathbf{C}}^2\right).$$

That is, the posterior probability density is proportional to the product of the likelihood and the prior probability density. To see whether Bayes' formula (2) is the best in which sense, we approach the problem of updating knowledge in a radical approach, namely, combining information theory and optimization technique. From now on to the end of the report, we still call the updated probability density $\pi(\mathbf{m})$ as the posterior though it *is not* from the usual Bayes' formula.

We begin by observing that there are two pieces of information available: the prior distribution $\pi_{\text{prior}}(\mathbf{m})$ and the data $\mathbf{d}$. The former completely depends on our prior knowledge and/or our prior belief while the latter is controlled by the ability to obtain the data.

From the information theory point of view, we should elicit the prior so that its discrepancy relative to the (unknown) posterior is as small as possible. Turning this argument around, if we believe that our prior is meaningful, the information gained in the posterior that we seek should not be large. The relative lost or gain between two probability densities is precisely captured by the relative entropy, also known as the Kullback-Leibler (KL) distance among many other names,

$$D_{\text{KL}}(\pi(\mathbf{m})|\pi_{\text{prior}}(\mathbf{m})) := \int_{\mathcal{M}} \pi(\mathbf{m}) \log\left(\frac{\pi(\mathbf{m})}{\pi_{\text{prior}}(\mathbf{m})}\right) d\mathbf{m}. \tag{3}$$

Clearly when the posterior $\pi(\mathbf{m})$ is identical to the prior $\pi_{\text{prior}}(\mathbf{m})$ the KL distance is zero, i.e, the prior is perfect. On the other hand, when the posterior departs from the prior $\pi_{\text{prior}}(\mathbf{m})$, i.e., there is discrepancy between our prior elicitation and the actual posterior, the KL distance is positive.

The other piece of information is the observation data $\mathbf{d}$. In general, we like to match the data as well as we can. Since the parameter is assumed to be distributed by the posterior, one way to approximately match the data well is to look for a posterior that minimizes the mean squared error between the computer prediction and the data, i.e.,

$$\min_{\pi(\mathbf{m})} \frac{1}{2} \int_{\mathcal{M}} \pi(\mathbf{m}) \left\|\mathcal{G}(\mathbf{m}) - \mathbf{d}\right\|_{\mathbf{C}}^2 d\mathbf{m} = -\int_{\mathcal{M}} \pi(\mathbf{m}) \log\left(\pi_{\text{like}}(\mathbf{d}|\mathbf{m})\right) d\mathbf{m}. \tag{4}$$

It should be pointed out that, in the context of deterministic inverse problem, $\frac{1}{2}\left\|\mathcal{G}(\mathbf{m}) - \mathbf{d}\right\|_{\mathbf{C}}^2$ is known as the misfit between the computer prediction and the data. The mean squared error is simply the average of the misfit over the posterior distribution. One also notes that the last term does not depend on the particular form of the data (1), and hence the likelihood model. To the end of the report we shall work with this general form of the likelihood.

At this point we see that there is a competition between the prior knowledge and the information from the data in the process of constructing the posterior. On the one hand the posterior should follow the prior, if we believe that the prior is the best possible within our subjective capability, so that the discrepancy in the prior modeling is minimized. On the other hand the posterior should be constructed such that that the computer prediction matches the data well in the mean squared sense. Note that the data is limited in general and only a few directions (regions) in the parameter

space are typically well-informed by the data. As a result, the posterior should compromise these two sources of information such that it captures the limited directions provided by the data while acting like the prior in the other directions. One way to construct such a posterior is to minimize the KL distance (3) and the mean squared error simultaneously, i.e.,

$$\min_{\pi(\mathbf{m})\geq 0} \mathcal{J} := \alpha_1 \int_{\mathcal{M}} \pi(\mathbf{m}) \log\left(\frac{\pi(\mathbf{m})}{\pi_{\text{prior}}(\mathbf{m})}\right) d\mathbf{m} - \alpha_2 \int_{\mathcal{M}} \pi(\mathbf{m}) \log\left(\pi_{\text{like}}(\mathbf{d}|\mathbf{m})\right) d\mathbf{m} \tag{5}$$

subject to

$$\int_{\mathcal{M}} \pi(\mathbf{m})\, d\mathbf{m} = 1, \tag{6}$$

where we have introduced two positive weights α_1 and α_2 to give us the freedom in making the prior more important than the data and vice versa. We note that our optimization formulation (5)–(6) is meaningful since it is a convex optimization problem with respect to $\pi(\mathbf{m})$. As a result, it has a unique solution and our next task is to find it.

In order to solve the optimization problem (5)–(6) we follow the Lagrangian formalism. In particular, consider the following Lagrangian

$$\mathcal{L} := \alpha_1 \int_{\mathcal{M}} \pi(\mathbf{m}) \log\left(\frac{\pi(\mathbf{m})}{\pi_{\text{prior}}(\mathbf{m})}\right) d\mathbf{m} - \alpha_2 \int_{\mathcal{M}} \pi(\mathbf{m}) \log\left(\pi_{\text{like}}(\mathbf{d}|\mathbf{m})\right) d\mathbf{m} + \gamma\left(\int_{\mathcal{M}} \pi(\mathbf{m})\, d\mathbf{m} - 1\right).$$

Note that we have ignored the constraint $\pi(\mathbf{m}) \geq 0$ since, as shall be shown, the posterior will be automatically nonnegative. Taking the first variation of the Lagrangian with respect to γ and arguing that it must be zero for all variations we recover the constraint (6). Now taking the first variation of the Lagrangian with respect to $\pi(\mathbf{m})$ in the direction $\tilde{\pi}(\mathbf{m})$ and arguing that it must be zero for all $\tilde{\pi}(\mathbf{m})$ we obtain

$$\pi(\mathbf{m}) = \exp\left(-\frac{\alpha_1 + \gamma}{\alpha_1}\right) \pi_{\text{like}}(\mathbf{d}|\mathbf{m})^{\frac{\alpha_2}{\alpha_1}}\, \pi_{\text{prior}}(\mathbf{m}),$$

which after substituting into the normalized constraint (6), becomes

$$\pi(\mathbf{m}) = \frac{\pi_{\text{like}}(\mathbf{d}|\mathbf{m})^{\frac{\alpha_2}{\alpha_1}}\, \pi_{\text{prior}}(\mathbf{m})}{\int_{\mathcal{M}} \pi_{\text{like}}(\mathbf{d}|\mathbf{m})^{\frac{\alpha_2}{\alpha_1}}\, \pi_{\text{prior}}(\mathbf{m})\, d\mathbf{m}}. \tag{7}$$

We define the unique solution (7) of our optimization problem (5)–(6) as the posterior distribution. It should be emphasized that we have not postulated any particular form of the prior or the data (likelihood) in order to obtain the posterior (7). Instead our posterior is simply a solution of the optimization problem in which we try to compromise the information from the data and the prior knowledge. Yet, it looks similar to the Bayes' formula (2)! In fact when $\alpha_1 = \alpha_2$, i.e. the prior and the likelihood are equally important, we rediscover the Bayes' formula! In general our posterior (7) is different from the Bayes' posterior. Indeed it is more general since it includes Bayes' formula as a special case.

Remark 1. *It might seem that our posterior is unusual at the first sight, it is in fact the Bayes' formula for a statistical inverse problem whose log likelihood is equal to the usual Bayes' log likelihood times* α_2/α_1.

Remark 2. *If we substitute the optimal posterior* (7) *into the cost function* (5)*, the minimum cost, after a simple algebra manipulation, is given by*

$$\min_{\pi_{post}(\mathbf{m})} \mathcal{J} = -\alpha_1 \log\left[\int_{\mathcal{M}} \pi_{like}(\mathbf{d}|\mathbf{m})^{\frac{\alpha_2}{\alpha_1}}\, \pi_{prior}(\mathbf{m})\, d\mathbf{m}\right] \tag{8}$$

.

2 Extension to infinite dimension

When the dimension of the parameter space $\mathcal{M}$ is infinite we no longer have the probability densities $\pi_{\text{prior}}(\mathbf{m})$ and $\pi_{\text{post}}(\mathbf{m})$ (with respect to the usual Lebesgue measure). In that case, we have to resort to using the prior and posterior distributions which are valid for both finite and infinite dimensional settings. Let us define $\mu(\mathbf{m})$ and $\nu(\mathbf{m})$ as the prior and the posterior distributions, respectively. Recall that in finite dimensional case, we have the following relations

$$d\mu := \mu(d\mathbf{m}) = \pi_{\text{prior}}(\mathbf{m})\, d\mathbf{m}, \quad d\nu := \nu(d\mathbf{m}) = \pi_{\text{post}}(\mathbf{m})\, d\mathbf{m}.$$

The KL distance now becomes

$$D_{\text{KL}}(\nu(\mathbf{m})|\mu(\mathbf{m})) := \int_{\mathcal{M}} \log\left(\frac{d\nu}{d\mu}\right) d\nu,$$

and similarly the mean squared error of the computer prediction provided that the parameter distributed by the posterior is

$$\frac{1}{2}\int_{\mathcal{M}} \|\mathcal{G}(\mathbf{m}) - \mathbf{d}\|_{\mathbf{C}}^2\, d\nu = -\int_{\mathcal{M}} \log(\pi_{\text{like}}(\mathbf{d}|\mathbf{m}))\, d\nu.$$

Consequently, the optimization problem in this case reads

$$\min_{\nu(\mathbf{m})} \mathcal{J} := \alpha_1 \int_{\mathcal{M}} \log\left(\frac{d\nu}{d\mu}\right) d\nu - \alpha_2 \int_{\mathcal{M}} \log(\pi_{\text{like}}(\mathbf{d}|\mathbf{m}))\, d\nu \tag{9}$$

subject to

$$\int_{\mathcal{M}} d\nu = 1. \tag{10}$$

The Lagrangian in this case has the following form

$$\mathcal{L} = \alpha_1 \int_{\mathcal{M}} \log\left(\frac{d\nu}{d\mu}\right) d\nu - \alpha_2 \int_{\mathcal{M}} \log(\pi_{\text{like}}(\mathbf{d}|\mathbf{m}))\, d\nu + \gamma \int_{\mathcal{M}} d\nu - \gamma,$$

which is clearly a linear function in the posterior measure ν. Thus, at stationary point of the Lagrangian we must have

$$\alpha_1 \log\left(\frac{d\nu}{d\mu}\right) - \alpha_2 \log(\pi_{\text{like}}(\mathbf{d}|\mathbf{m})) + \gamma = 0,$$

which is equivalent to

$$\frac{d\nu}{d\mu} = \exp\left(-\frac{\gamma}{\alpha_1}\right) \pi_{\text{like}}(\mathbf{d}|\mathbf{m})^{\frac{\alpha_2}{\alpha_1}},$$

and using the normalized constraint (10) we conclude that the optimal posterior distribution is such that its Radon-Nikodym derivative with respect to the prior distribution is proportional to the likelihood, i.e.,

$$\frac{d\nu}{d\mu} = \frac{\pi_{\text{like}}(\mathbf{d}|\mathbf{m})^{\frac{\alpha_2}{\alpha_1}}}{\int_{\mathcal{M}} \pi_{\text{like}}(\mathbf{d}|\mathbf{m})^{\frac{\alpha_2}{\alpha_1}}\, d\mu}. \tag{11}$$

Clearly, (11) becomes (7) in the finite dimensional setting.

Similar to the finite dimensional case, if we substitute the optimal posterior measure given by (11) into the cost function (9), then the smallest cost function is given by

$$\min_{\nu(\mathbf{m})} \mathcal{J} = -\alpha_1 \log\left[\int_{\mathcal{M}} \pi_{\text{like}}(\mathbf{d}|\mathbf{m})^{\frac{\alpha_2}{\alpha_1}}\, d\mu\right],$$

which is simplified to (8) when $\mathcal{M}$ is finite dimensional.

Remark 3. *Note that the cost objective* (9) *is nonnegative and linear in* ν *(which is nonnegative by definition of probability measure). As a result, the posterior distribution* ν *given by the Radon-Nikodym derivative* (11) *is a unique solution of the optimization problem* (9)–(10).