# Bayesian Inference Featuring Entropic Priors

Tilman Neumann

*Tilman.Neumann@lycos.de*

*2006-11-15 First draft*
*2007-07-02 Revision 1.0*
*2007-09-13 Revision 1.1*

Final version published in *Proceedings of 27th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, American Institute of Physics, vol. 954 (2007), pp. 283–292*

**Abstract.** The subject of this work is the parametric inference problem, i.e. how to infer from data on the parameters of the data likelihood of a random process whose parametric form is known a priori. The assumption that Bayes' theorem has to be used to add new data samples reduces the problem to the question of how to specify a prior before having seen any data. For this subproblem three theorems are stated. The first one is that Jaynes' Maximum Entropy Principle requires at least a constraint on the expected data likelihood entropy, which gives entropic priors without the need of further axioms. Second I show that maximizing Shannon entropy under an expected data likelihood entropy constraint is equivalent to maximizing relative entropy and therefore reparametrization invariant for continuous-valued data likelihoods. Third, I propose that in the state of absolute ignorance of the data likelihood entropy, one should choose the hyperparameter $\alpha$ of an entropic prior such that the change of expected data likelihood entropy is maximized. Among other beautiful properties, this principle is equivalent to the maximization of the mean-squared entropy error and invariant against any reparametrizations of the data likelihood. Altogether we get a Bayesian inference procedure that incorporates special prior knowledge if available but has also a sound solution if not, and leaves no hyperparameters unspecified.

**Keywords:** parametric inference, Bayesian inference, Maximum Entropy Principle, entropic prior, reparametrization invariance, non-informative prior, least-informative prior, expected entropy change maximization, stable inference
**PACS:** 02.50.Cw, 02.50.Tt, 05.20.-y

## 1. THE PARAMETRIC INFERENCE PROBLEM

If we know the parametric form of the data likelihood function $L(x|\theta)$ of a random process with random variable(s) $X$ and have observed $n$ data samples $\{x_1,...,x_n\} = x^n \in X^n$, then what can we say about the parameters $\theta$ after having seen the data?

## 2. FROM BAYES TO ENTROPIC PRIORS

Let's assume that the right way to update our degrees of belief in some $\theta$'s being the real parameters $\theta^*$ from new data samples is Bayes' theorem. Emphasizing its quality as an "update rule" this theorem can be stated as

$$P(\theta|x^n) \propto P(\theta|x^{n_1})L(x^{n_2}|\theta) \tag{1}$$

with $n_1 + n_2 = n$ and $n_1, n_2 \geq 0$. Then, all our knowledge (and uncertainty) about the parameters $\theta$ after having seen data $x^n$ is represented by the posterior density $P(\theta|x^n)$.

Adopting this point, the parametric inference problem reduces to the question of how to determine a prior $P(\theta) \equiv P(\theta|x^0)$ expressing the state of mind of the "reasoner" before having seen any data. In order to avoid confusions with priors that already incorporate knowledge derived from data, in the following I shall call these priors "first priors".

Note that a first prior might contain real information, although it represents a state of mind before having seen any data. Such information could stem from knowledge about the design of the random experiment, like that we know that a die is loaded and has its mass center near the six face. If we don't have such prior knowledge, a first prior will be equivalent to what is most often called a "non-informative" or "least-informative" prior.

The problem of how to assign priors dates back at least to the work of Laplace and Bernoulli [1] and is still an active research area. Founded in 1957 by Edwin T. Jaynes, the approach that has probably received most attention is the Principle of Maximum Entropy [2] [3] [4] [5]. Let's restrict the discussion in this section to the case of discrete-valued data likelihoods; then the ME principle claims that among all possible densities $P(\theta)$ satisfying a couple of constraints given as expectation values (and a normalization constraint), we should choose the one that maximizes the Shannon entropy

$$S_P = -\int_\Theta P(\theta) \ln P(\theta) d\theta \tag{2}$$

Note that doubts on the uniqueness of the entropy measure raised e.g. by Uffink [6] have been countered by Caticha and Giffin [7].

A particular proposal for least-informative priors are the so-called "entropic priors" [8] [9] [10] [11] [12] [13], that have first been derived by Skilling [8] from the axioms of Maximum Entropy and an additional "quantification" argument. Entropic priors $P_\alpha(\theta) \equiv P(\theta)$ owe their name to the shape of their density

$$P_\alpha(\theta) \propto e^{\alpha S_L(\theta)} \tag{3}$$

where

$$S_L(\theta) = -\int_X L(x|\theta) \ln L(x|\theta) dx \tag{4}$$

is the entropy of the data likelihood function with parameters $\theta$.

In 2004, Caticha and Preuss recognized that in order to solve problems with repeatable experiments, entropic priors need a constraint on the expected data likelihood entropy $\langle S_L \rangle$ (see [12], page 5), which is defined as

$$\langle S_L \rangle = \int_\Theta S_L(\theta) P(\theta) d\theta \tag{5}$$

However, application of Jaynes' Maximum Entropy Principle with a constraint like $\langle S_L \rangle = \bar{S}$ always gives an entropic prior! Thus, such a constraint has the same power as Skilling's quantification argument, and defines the assumption made in entropic priors.

**Theorem 1 (Entropic Priors)** *Application of the Maximum Entropy Principle requires at least a constraint on the expected data likelihood entropy like $\langle S_L \rangle = \bar{S}$. The result is*

*an entropic prior. The information contained in a "pure" entropic prior is an expectation about the data likelihood entropy and nothing else.*[1]

## 3. CONTINUOUS-VALUED DATA LIKELIHOODS

It is a well-known fact that for continuous-valued data likelihoods, equation 2 is not invariant under reparametrizations of the data likelihood; consequently, mere reparametrizations might yield different inference results. To overcome this problem, we have to introduce a measure $m(\theta)$ in the log that transforms as $P(\theta)$ does. This gives the relative entropy:[2]

$$S_{P|m} = -\int_{\Theta} P(\theta) \ln \frac{P(\theta)}{m(\theta)} d\theta \tag{6}$$

Note that reparametrization invariance would be guaranteed by any $m(\theta)$, and that the ME principle doesn't give us a hint which one to use. However, we can formulate a couple of desirable properties:

- **Jaynes' Argument:** "Except for a constant factor, the measure $m(\theta)$ is also the prior describing 'complete ignorance' of $\theta$." (Jaynes in [14], page 377)
- **Axiomatic Consistency:** Even for continuous-valued data likelihoods, the resulting prior still has to obey the restrictions imposed by Skillings axioms.
- **Limit Argument:** For continuous-valued data likelihoods that can be derived as some limit of a discrete-valued likelihood (e.g. hypergeometric and binomial), the following two solutions should be equivalent: First, solving the variational problem with the simple Shannon-entropy for the discrete-valued data likelihood and getting the limit of the solution; and second, solving the variational problem with the relative entropy for the continuous-valued likelihood.

First proposed by Rodriguez [9], a popular approach for entropic priors is that we should maximize the relative entropy under a normalization and an expected data likelihood entropy constraint, giving the solution

$$P_{\alpha}(\theta) \propto m(\theta) e^{\alpha S_L(\theta)} \tag{7}$$

where

$$m(\theta) \propto \sqrt{det g_{ij}(\theta)} \tag{8}$$

is Jeffreys' prior which is based on the Fisher information matrix.

It is easy to see that the solution proposed by equations 7 and 8 doesn't satisfy the first two demands defined above. But did we already put in all the information we have to find

---

[1] I call an entropic prior "pure" if there are no more constraints than one on the expected data likelihood entropy and one for the normalization. Further constraints may lead to more complex expressions than equation 3; on the other hand, as we will see in section 5.1, the expressions may simplify as well if the data likelihood entropy takes on a logarithmic form.

[2] Relative entropy is often written with an opposite sign. I follow the notation of [12].

the right entropy functional and/or $m(\theta)$? I think we can do better than just demanding reparametrization invariance. What we really want is: *Maximizing our entropy functional under a normalization constraint and an expected data likelihood entropy constraint should give the same result, no matter which parametrization we choose.* Let's have a look at the Lagrangian describing this using the simple Shannon entropy:

$$\begin{aligned}
\mathscr{L} &= (-\int_\Theta P(\theta)\ln P(\theta)d\theta) + \alpha(-\int_\Theta P(\theta)\int_X L(x|\theta)\ln L(x|\theta)dxd\theta - \bar{S}) \\
&+ \lambda(\int_\Theta P(\theta)d\theta - 1) \\
&= (-\int_\Theta P(\theta)(\ln P(\theta) + \alpha\int_X L(x|\theta)\ln L(x|\theta)dx)d\theta) - \alpha\bar{S} + \lambda(\int_\Theta P(\theta)d\theta - 1) \\
&= (-\int_\Theta P(\theta) ln\frac{P(\theta)}{m_\alpha(\theta)}d\theta) + \lambda(\int_\Theta P(\theta)d\theta - 1) - \alpha\bar{S} \qquad (9)
\end{aligned}$$

where

$$m_\alpha(\theta) \propto e^{-\alpha\int_X L(x|\theta)\ln L(x|\theta)dx} \equiv e^{\alpha S_L(\theta)} \qquad (10)$$

Since the $\alpha\bar{S}$-term cancels when expression 9 is maximized, we see:

**Theorem 2 (Reparametrization Invariance)** *Maximizing the simple Shannon entropy under an expected data likelihood entropy constraint is equivalent to maximizing relative entropy with the underlying measure $m(\theta)$ given by $m_\alpha(\theta)$, and therefore reparametrization invariant. Consequently, both the inference procedure and the resulting prior $P_\alpha(\theta)$ are exactly the same for discrete- and continuous-valued data likelihoods.*

The argument above shows in my opinion that it is a tautology to have both an $m(\theta)$ given by equation 8 and the entropic term as in equation 7. One of the two is enough! Nevertheless, my personal belief is that this solution is only slightly wrong, because Jeffreys' prior is quite similar to the entropic term, and therefore the two terms are almost linear dependent conditions in the solution of the variational problem.[3]

## 4. THE HYPERPARAMETER $\alpha$

In this section we are going to see how we can deal with the hyperparameter $\alpha$ still present in the "generic" entropic priors derived so far. Note that like the internal energy of an ideal gas, the expected data likelihood entropy $\langle S_L\rangle$ of an entropic prior is a function of (inverse) temperature and nothing else, i.e. $\langle S_L\rangle \equiv \langle S_L\rangle(\alpha)$ (see figure 1 for a typical example). Therefore, if we have a concrete expectation like $\langle S_L\rangle = \bar{S} = 0.35$, then we just pick the $\alpha$ that realizes the expected $\langle S_L\rangle = \bar{S}$, and our prior is fixed, i.e. contains no more variables than the $\theta$s.

---

[3] Actually, I think that Jeffreys' prior is a second-order approximation to the entropic prior, the small error being caused by merely asymptotically valid expansions, possibly those in [15], page 13, and [16], page 4.

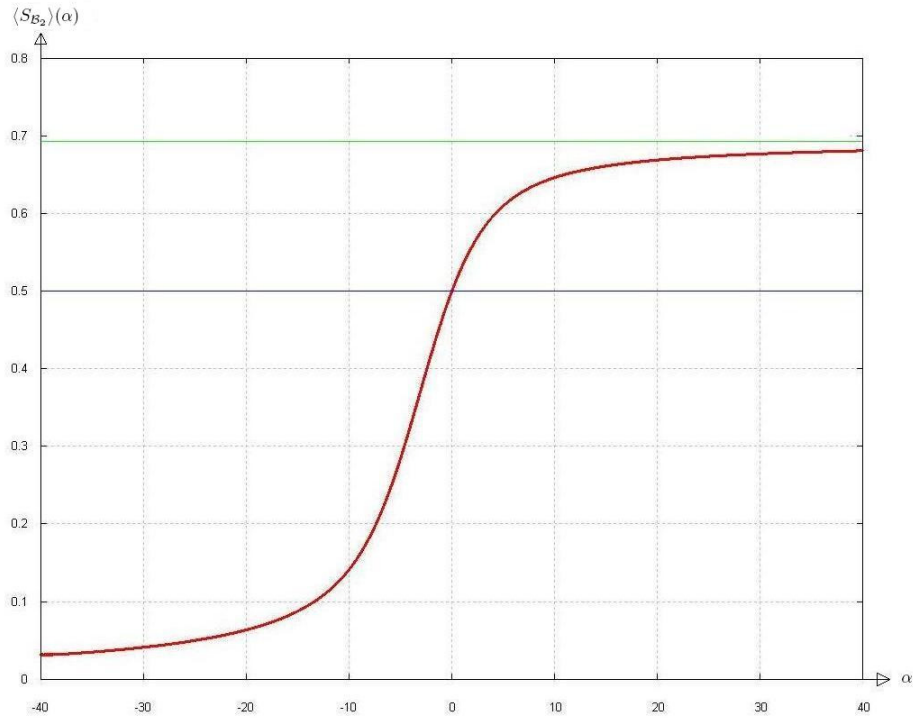


**FIGURE 1.** Expected entropy as a function of $\alpha$ given the entropic prior for 2-class discrete data

The more difficult case is when we have no idea at all about $\bar{S}$, which is when first priors become equivalent to least-informative priors. How shall we deal with $\alpha$? The most popular approach currently is to treat $\alpha$ as a nuisance parameter and eliminate it via outmarginalization [12] [17] [18]. Nevertheless, I see a couple of problems with this:

- **Technical Problems:** The outmarginalization procedure requires a prior $P(\alpha)$. In my eyes, this just means that the problem of determining the least-informative prior is moved to another, less "visible" place. Furthermore, we have even less intuition on how to determine a prior on a hyperparameter than on $\theta$. Consequently, the attempts to specify $P(\alpha)$ I've seen so far don't convince me very much: Strauss, Wolf and Wolpert (see [17], page 115) simply assumed a "reasonable" flat prior, which is obviously not a well-founded argument. Preuss and Caticha [12] advocate for an entropic prior on $\alpha$, but then they get another hyperparameter (e.g. $\beta$) they have to deal with. Rodriguez [19] proposed an infinite progression of entropic priors, but how can we compute this?
- **Axiomatic Consistency:** The result of an outmarginalization will usually not match anymore the entropic form required by Skilling's axioms.
- **Epistemic Argument:** We are doing Bayesian inference, because we are interested in the uncertainties of all possible $\theta$s being the real $\theta^*$s, based upon the state of knowledge we are in after having seen certain data. But we are not interested at all in the distribution of a hyperparameter. Therefore in this case I'm missing the motivation for a Bayesian treatment.

Which alternatives do we have? Skilling advanced the view that $\alpha$ could not be fixed a priori. (see [8], page 51). Nevertheless, my opinion is that the desiderata formulated above can only be satisfied by a point estimate, and that *though no particular value is correct for any possible data likelihood, there might exist a rule to determine $\alpha$ for a given data likelihood*.

In order to progress into that direction, let's again have a look at figure 1 and recall that any choice of $\alpha$ corresponds to a particular expectation $\langle S_L \rangle$. If we don't have a clue about the real value of the data likelihood entropy, we would surely not want to expect a zero ("the die always gives the same number") or maximal value ("the die is fair"), which would be examples of great prior knowledge. But which of the "moderate" $\alpha$-values makes most sense, and what is this sense? My opinion is:

**Theorem 3 (Least-Informative Priors)** *If we are implicitly making an assumption on the data likelihood entropy anyway, although we don't know which one to expect, the least biased choice is the $\alpha^*$ that maximizes the assumption error, i.e. the entropy variance. This is the exact meaning of "least-informativity".*

This choice has amazing properties, for example the following equivalence:

$$\begin{aligned}
&\langle S_L^2 \rangle(\alpha) - \langle S_L \rangle(\alpha)^2 = \frac{\int_\Theta S_L(\theta)^2 e^{\alpha S_L(\theta)} d\theta}{\int_\Theta e^{\alpha S_L(\theta)} d\theta} - \left(\frac{\int_\Theta S_L(\theta) e^{\alpha S_L(\theta)} d\theta}{\int_\Theta e^{\alpha S_L(\theta)} d\theta}\right)^2 \\
&= \frac{\left(\frac{\partial}{\partial \alpha} \int_\Theta S_L(\theta) e^{\alpha S_L(\theta)} d\theta\right)\left(\int_\Theta e^{\alpha S_L(\theta)} d\theta\right) - \left(\int_\Theta S_L(\theta) e^{\alpha S_L(\theta)} d\theta\right)\left(\frac{\partial}{\partial \alpha} \int_\Theta e^{\alpha S_L(\theta)} d\theta\right)}{\left(\int_\Theta e^{\alpha S_L(\theta)} d\theta\right)^2} \\
&= \frac{\partial}{\partial \alpha} \frac{\int_\Theta S_L(\theta) e^{\alpha S_L(\theta)} d\theta}{\int_\Theta e^{\alpha S_L(\theta)} d\theta} = \frac{\partial}{\partial \alpha} \langle S_L \rangle(\alpha)
\end{aligned} \tag{11}$$

Thus, choosing the $\alpha^*$ that maximizes the data likelihood entropy variance of an entropic prior is equivalent to choosing the turning point of $\langle S_L \rangle(\alpha)$,

$$\alpha^* = \arg \max_{\{\alpha\}} \left(\frac{\partial \langle S_L \rangle(\alpha)}{\partial \alpha}\right) \tag{12}$$

Therefore, theorem 3 could as well be called a "Maximum Entropy Change Principle".

Further interesting properties of this principle are

- It is *absolutely invariant against reparametrizations* of the probabilistic model, because the principle itself is based solely on the data likelihood function. If the likelihood function is reparametrized, the functional form of the condition for the turning point of $\langle S_L \rangle(\alpha)$ changes correspondingly so that the resulting $\alpha^*$ always stays the same.
- It realizes a *stable inference* solution as proposed by Tikochinsky [20]: The $\alpha$ to choose is the one that maximizes the change in expected entropy; the other way round, it is the choice where a small error in our assumption (say, the expected entropy differs from the real one) has the least effect on the shape of the resulting density and estimators derived from it. Since the quantity whose stability is guaranteed is an entropy, we could call that solution "entropy-stable".

- That the proposed choice *maximizes the mean-squared entropy error* is in agreement with [21] that the quadratic loss is the unique loss function consistent with the entropy measure.
- The principle has many meaningful transformations; for example it may be rewritten as the demand to set the expected entropy skew to zero.

Note last not least that maximization of expected entropy change has been proposed in nonequilibrium thermodynamics by Q.A. Wang [22] [23] before, and that we could state the choice of $\alpha$ as well in a Bayesian style with the prior $P(\alpha)$ being a delta function.

# 5. APPLICATION OF THE MAXIMUM ENTROPY CHANGE PRINCIPLE

I applied the proposed rule to the normal data likelihood function as well as to data likelihoods for discrete-valued random processes. Computations usually consist of two components: First an approximation of the expected data likelihood entropy and/or some higher central moments for a given $\alpha$, and second a kind of Newton-step procedure to find the $\alpha^*$ where the entropy change is maximized.

## 5.1. Normal data

For a normal data likelihood $\mathcal{N}(x|\mu,\sigma) \equiv L(x|\theta)$ with

$$\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{13}$$

the data likelihood entropy $S_{\mathcal{N}}(\mu,\sigma) \equiv S_L(\theta)$ is given by

$$S_{\mathcal{N}}(\mu,\sigma) = \ln \sqrt{2\pi}\sigma \tag{14}$$

The entropic prior with hyperparameter $\alpha$ therefore has the form

$$P_\alpha(\mu,\sigma) \propto e^{\alpha \ln \sqrt{2\pi}\sigma} \tag{15}$$

Though $\langle S_{\mathcal{N}} \rangle(\alpha)$ gives $\infty$, its derivative towards $\alpha$ can be computed. Maximizing the change of $\langle S_{\mathcal{N}} \rangle(\alpha)$ with respect to $\alpha$ gives $\alpha^* = -1$ and with this, the entropic prior resolves to

$$P(\mu,\sigma) \propto \frac{1}{\sigma} \tag{16}$$

This is the result preferred by Jeffreys although it contradicts his "general rule" (see [24], page 1345).

## 5.2. Discrete random processes

Now we consider random processes that can only take on a finite number of values $X \in \{X_1, ..., X_k\}$ with probabilities $p_1, ..., p_k, \sum_{i=1}^{k} p_i = 1$. A single random experiment with such a likelihood function is called a Bernoulli trial for $k = 2$, and a Bernoulli scheme for any $k \geq 2$; I will refer to the corresponding distributions in general as Bernoulli likelihoods $\mathscr{B}_k$. If we combine several Bernoulli experiments ignoring their order, we get binomial or multinomial data likelihoods. Inference from such data likelihoods has turned out to be a tough problem [25] [26], so let's see what the new approach delivers:

The entropy of a Bernoulli likelihood is given by the standard Shannon entropy

$$S_{\mathscr{B}_k}(p_1, ..., p_{k-1}) = -\sum_{i=1}^{k} p_i \ln p_i \tag{17}$$

The entropic prior with hyperparameter $\alpha$ is

$$P_\alpha(p_1, ..., p_{k-1}) \propto e^{\alpha S_{\mathscr{B}_k}(p_1, ..., p_{k-1})} \tag{18}$$

and the expected data likelihood entropy

$$\langle S_{\mathscr{B}_k} \rangle(\alpha) = \int ... \int_{\mathbf{P}} S_{\mathscr{B}_k}(p_1, ..., p_{k-1}) P_\alpha(p_1, ..., p_{k-1}) d\mathbf{p} \tag{19}$$

where we have to integrate over the $(k-1)$-dimensional simplex
$\mathbf{P} = \{(p_1, ..., p_{k-1}) \mid p_1 = 0..1, p_2 = 0..(1-p_1), ..., p_{k-1} = 0..(1-p_1-...-p_{k-2})\}$.

The required computations are pretty time-consuming, because the expected entropy integrals have the tendency to converge very slowly. Already for $k = 3$, it was quite necessary to use special numerical integration techniques. The best method I applied is an Adaptive Quadrature algorithm [27] using Gauss-Legendre polynomials and some analytical simplifications. The results are given in table 1.

**TABLE 1.** $\alpha^*$-choice for the entropic prior for k-class discrete data

| k | $\alpha^*$ | $\langle S_{\mathscr{B}_k} \rangle(\alpha^*)$ |
|---|---|---|
| 2 | -3.118356848554... | 0.3685467... |
| 3 | -4.772026959... | 0.5676038... |
| 4 | -6.0437688... | 0.7127738... |
| 5 | -7.104524... | 0.8308972... |
| 6 | -8.03223... | 0.932265... |
| 7 | -8.9216... | 1.01628... |

The entropic prior for 2-class discrete data resembles much Jeffreys' prior $P_J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$ but is proper. In fact, we can develop the entropic prior into a power series around $\theta_0 = 1/2$ to which Jeffreys' prior is a second order approximation.

As indicated by table 1, the $\alpha^*$-values keep the expected data likelihood entropies near to $\frac{\ln k}{2}$ for any $k$.

# 6. FINAL REMARKS

The major points of this work are expressed in the three theorems. The first claims that Jaynes' Maximum Entropy Principle requires necessarily a constraint $\langle S_L \rangle = \overline{S}$ on the expected entropy of the considered data likelihood $L(x|\theta)$. Practically, such a constraint means that we have an idea about the complexity of the random problem. If we apply the ME principle with such a constraint, we will always get an entropic prior $P_\alpha(\theta)$.

The second theorem shows that maximization of Shannon entropy under an expected data likelihood entropy constraint is equivalent to maximizing relative entropy without such a constraint, but with a particular choice of the underlying measure $m(\theta)$. Therefore, reparametrization invariance is guaranteed for continuous-valued data likelihoods by exactly the same formalism that was derived for discrete-valued data likelihoods.

The third and last theorem is a proposal for priors in the absence of any prior knowledge. It claims that in this case, we should choose the hyperparameter $\alpha$ such that the data likelihood entropy variance or equivalently, the change of expected data likelihood entropy against $\alpha$ is maximized. Among other interesting properties, this principle is completely invariant against reparametrizations of the data likelihood and could be called an *entropy-stable* inference solution. Furthermore, the principle shows us that Jose Bernardo's famous sentence "Non-informative priors do not exist" [28] is absolutely correct: Any entropic prior implies an assumption on the data likelihood entropy, and if we don't know which entropy to expect, all we can do is to minimize our assumption error by maximizing the entropy variance.

Putting all pieces together, we get a "universal" solution procedure (in the sense that it is applicable to any data likelihoods) for the parametric inference problem that leaves no hyperparameters unspecified. This procedure has the following components:

- You are given the parametric form of the data likelihood of a random problem.
- Determine the "first prior" using the ME principle with a constraint on the expected data likelihood entropy. The result is an entropic prior with hyperparameter $\alpha$. If you really have an expectation about the data likelihood entropy, choose the $\alpha$ that realizes it; else take the value that maximizes the data likelihood entropy variance (or equivalently, the change of data likelihood entropy).
- Use Bayes' theorem to update your degrees of belief from the first prior with data.
- Compute the desired parameter estimates from the posterior.

Concerning statistical inference, I think that the whole of the theory developed here is very self-consistent, explains some things that have not been explained yet, and gives encouraging results. An aspect I want to emphasize is that it prefers special ("subjective") knowledge if present. In fact, if we understand the ME principle as a mere tool to cast prior knowledge into a nicely shaped prior, then the approach presented here could help to close the gap between the positions of "objectivists" and "subjectivists". I'm now looking forward to use the results of this paper as a building block to tackle more complex problems, like non-parametric inference for high-dimensional real-world problems. That's where the right choice of priors will have the biggest impact.

Concerning physics, like Wang I believe that the Maximum Entropy Change Principle is a kind of natural law, and I wonder if there are further problems and theories to

which it might be applicable. Some potential candidates are already suggested by the near relationship between inference and thermodynamics through the ME principle, for example the problems of temperature fluctuations [10] [29] [30] or a thermodynamical uncertainty relation [31] [32]. Other interesting relationships are those with black hole thermodynamics [33] [34] and with the theories that can be derived by application of the ME principle with exotic probabilities [35].

## ACKNOWLEDGMENTS

I would like to thank Ariel Caticha, John Skilling and Matthew Brand for valuable comments, my mother Karin for her support, and my son Max for just being there.

## REFERENCES

1. E. T. Jaynes, *IEEE Transactions on System Science and Cybernetics* **4**, 227–241 (1968).
2. E. T. Jaynes, *Physical Review* **106**, 620–630 (1957).
3. E. T. Jaynes, *Physical Review* **108**, 171–190 (1957).
4. W. T. Grandy Jr., "The three phases of statistical mechanics," in *Maximum Entropy and Bayesian Methods*, edited by J. Skilling, Kluwer, Dordrecht, 1989, pp. 73–91.
5. M. Mihelcic, Maximum entropy method and Bayesian probability theory, Technical Report Jül-3395, Forschungszentrum Jülich (1995).
6. J. Uffink, *Studies in History and Philosophy of Modern Physics* **26**, 223–261 (1995), URL `http://citeseer.ist.psu.edu/uffink97can.html`.
7. A. Caticha, and A. Giffin, Updating Probabilities, *arXiv:physics/0608185* (2006).
8. J. Skilling, "Classic Maximum Entropy," in *Maximum Entropy and Bayesian Methods*, edited by J. Skilling, Kluwer, Dordrecht, 1989, pp. 45–52.
9. C. C. Rodriguez, "The metrics induced by the Kullback number," in *Maximum Entropy and Bayesian Methods*, edited by J. Skilling, Kluwer, Dordrecht, 1989, pp. 415–422.
10. A. Caticha, "Maximum entropy, fluctuations and priors," in *Maximum Entropy and Bayesian Methods in Science and Engineering*, edited by A. Mohammad-Djafari, AIP, Melville, 2001, pp. 94–105.
11. A. Caticha, and R. Preuss, Entropic priors, *arXiv:physics/0312131* (2003).
12. A. Caticha, and R. Preuss, *Physical Review E* **70** (2004).
13. C. G. Chakrabarti, N. C. Das, and K. De, *Czechoslovak Journal of Physics* **52**, 911–918 (2002).
14. E. T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge (UK), 2003.
15. S. Amari, and H. Nagaoka, *Methods of Information Geometry*, American Mathematical Society, 2000.
16. V. Balasubramanian, Statistical inference, occam's razor and statistical mechanics on the space of probability distributions, *arXiv:cond-mat/9601030* (1996).
17. C. M. E. Strauss, D. H. Wolpert, and D. R. Wolf, "Alpha, evidence, and the entropic prior," in *Maximum Entropy and Bayesian Methods*, edited by A. Mohammad-Djafari, Kluwer, Dordrecht, 1993, pp. 113–120.
18. R. Fischer, W. von der Linden, and V. Dose, "On the importance of $\alpha$ marginalization in maximum entropy," in *Maximum Entropy and Bayesian Methods*, edited by K. Hanson, and R. Silver, Kluwer, Dordrecht, 1996, pp. 229–236.
19. C. C. Rodriguez, Entropic Priors for Discrete Probabilistic Networks and for Mixtures of Gaussian Models, *arXiv:physics/0201016* (2002).
20. Y. Tikochinsky, N. Z. Tishby, and R. D. Levine, *Physical Review A* **30**, 2638–2644 (1984).
21. Y. Tikochinsky, and R. D. Levine, *J. Math. Phys.* **25**, 2160–2168 (1984).
22. Q. A. Wang, Maximizing entropy change and least action principle for nonequilibrium systems, *arXiv:cond-mat/0312329* (2003).

23. Q. A. Wang, Action principle and Jaynes' guess method, *arXiv:cond-mat/0407515* (2004).
24. R. E. Kass, and L. Wasserman, *Journal of the American Statistical Association* **91**, 1343–1370 (1996).
25. P. Walley, *Journal of the Royal Statistical Society B* **58**, 3–57 (1996).
26. M. Zhu, and A. Y. Lu, *Journal of Statistics Education* **12** (2004).
27. A. Genz, "An Adaptive Numerical Integration Algorithm for Simplices," in *Computing in the 90s*, edited by N. Sherwani, E. de Doncker, and J. Kapenda, Springer, New York, 1991, pp. 279–292.
28. J. M. Bernardo, T. Z. Irony, and N. D. Singpurwalla, *Journal of Statistical Planning and Inference* **65**, 159–189 (1997).
29. G. D. J. Phillies, *American Journal of Physics* **52**, 629–632 (1984).
30. H. B. Prosper, *American Journal of Physics* **61**, 54–58 (1993).
31. Y. Alhassid, and R. D. Levine, *Chemical Physical Letters* **73**, 16–20 (1980).
32. F. Schlögl, *Journal of Physics and Chemistry of Solids* **49**, 679–683 (1988).
33. J. Baez, This week's finds in mathematical physics – week 111 (1997), URL `http://math.ucr.edu/home/baez/week111.html`.
34. R. M. Wald, Black holes and thermodynamics, *arXiv:gr-qc/9702022* (1997).
35. S. Youssef, Physics with exotic probability theory, *arXiv:hep-th/0110253* (2001).