An Information Theoretic Interpretation of Variational Inference based on the MDL Principle and the Bits-Back Coding Scheme

Ghassen Jerfel

April 2017

As we will see during this talk, the Bayesian and information-theoretic views of variational inference provide complementary and mutually beneficial perspectives to the same problems with two different languages.

More specifically, based on the paper by Honkela and Valpola [4], we will provide an interpretation of variational inference based on the MDL principle as a theoretical framework for model evaluation using code length and on Bits-back as a practical efficient coding scheme which builds on the MDL principle.

From this perspective, we can better understand the shortcomings of practical variational inference and the over-pruning phenomenon of variational auto-encoders.

1 Variational Inference: The Bayesian Interpretation

In statistical Bayesian learning, parameter values can have distributions too. Accordingly, for a given statistical model, we start with a prior distribution of the parameters that rationally incorporates some knowledge about the world [2]. [1]

Using the Bayes Rule [Eq 1], every observed evidence updates the parameter distribution resulting in a posterior distribution $p(\theta \mid X)$ which contains all of the parameter information. We can then make predictions by evaluating all possible parameter values as weighed by their posterior probabilities [Eq 2] (marginalization) .

$$p(\theta \mid X) = \frac{p(X \mid \theta)p(\theta)}{p(X)} \tag{1}$$

This use of distributions is insensitive to over-fitting and provides a natural method to compare and select models since we can easily rerun and evaluate the posteriors.

$$p(X) = \int_{\theta} p(X \mid \theta)p(\theta)d\theta \tag{2}$$

However, the posterior distribution is rarely tractable[2] and requires an approximation such as maximum-a-posteriori (MAP) estimates or MCMC sampling. The most scalable and currently standard approximation is variational inference [5].

Variational inference proposes a variational or approximate distribution (or family thereof) over the parameters. It then proceeds to minimize the KL divergence (a popular distribution metric) between the true posterior and the approximate posterior. Eventually, VI returns the set of variational parameters that best approximate the true posterior and one can use the variational distribution as an alternative to the true posterior distribution for testing and model selection [5].

The optimization objective is the KL divergence or its equivalent negative Evidence Lower BOund

[1] Whether this has to be an informative or a non-informative prior or if to depend on the data (empirical Bayes) is up for debate

[2] Logistic regression, use of neural networks or even a mixture of Gaussians present intractable posteriors

as can be seen in the following derivation:

$$\begin{aligned}
\log p(x) &= \log \int p(x,\theta) d\theta \\
&= \log \int \frac{q(\theta) p(x,\theta)}{q(\theta)} d\theta \\
&= \log E_q[\frac{p(x,\theta)}{q(\theta)}] \\
&>= E_q[\log \frac{p(x,\theta)}{q(\theta)}] = L[q;p,x] \\
&= E_q[\log p(x,\theta)] - E[q(\theta)]
\end{aligned}$$

$$\begin{aligned}
ELBO = L[q;p,x] &= E_q[\log p(x|\theta)p(\theta)] - E_q[\log q(\theta)] \\
&= E_q[\log p(x|\theta)p(\theta)] + H[q(\theta)] \\
&= E_q[\log p(x|\theta)] - KL(q(\theta)||p(\theta))
\end{aligned}$$

However, Variational Inference is problematic from a purely Bayesian point-of-view:

A: The ELBO minimize an "ad hoc" distance measure between an approximate posterior and the true posterior. It just so happens that the cost function is a convenient lower bound to the model evidence $\log p(x) \geq L(x)$

B: It is not clear, from a Bayesian perspective, how the approximate marginalization over the approximate posterior relates to other methods since the approximation doesn't correspond to a proper loss function (but rather a lower-bound).

C: In order to minimize the expected loss of a log score function $\log \frac{q}{p}$ we should use

$$E_p[\log p - \log q] = KL(p \mid\mid q)$$

rather than [3]

$$KL(q \mid\mid p) = -E_q[\log q - \log p]$$

[3] However, since KL is asymmetric, there is no relation between the two different forms.

2 Minimum Description Length

Von Neumann is claimed to have said, "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk."

Based on the Minimum Message Length by Wallace [8], Minimum Description Length [7] was developed by Risanen in 1979 as an information theoretic formalization of **Occam's razor** and provides a criterion that can be seen as a **tractable approximation to the Kolmogorov Complexity** [4] [8].

The MDL principle states that the best model of some data is the one that minimizes the combined cost of describing the data and the cost of describing the misfit between the model and the data.

The description in this case is a binary code that the receiver can use to reconstruct the original data using the communicated model:

- First part of the code describes the model: the number of bits it takes to describe the parameters.
- Second part of the code describes the model misfit: the number of bits it takes to describe the discrepancy between the correct output and the training output.

Using this two-part code, the receiver can train the model and then use the misfit part to correct the training output and reach the true output.

This principle offers a **natural trade-off** between model complexity and descriptive adequacy (model fit or good compression) and links the resolution of this trade-off to the problem of finding the most efficient use of memory: the shortest code.

The original framework focuses on evaluating code length and doesn't address how to construct the

[4] Kolmogorov Complexity is uncomputable and language-dependent

codes. However, we know from Shannon's coding theory [6] that data from a discrete distribution $p(X)$ cannot be represented with less than $E[-\log_2(p(X))]$ bits or $E[-\ln(p(X))]$ nats.

Therefore, for any probability distribution, it is possible to construct a code such that the length in bits is equal to $-\log_2 p(X)$. The reverse is also true.
Searching for an efficient code thus reduces to searching for a good probability distribution or model.

More concretely, we can represent the expected description length as follows:

$$L(X) = E[-\log(P(\theta \mid H)) - \log(p(X \mid \theta, H))] \tag{3}$$

In the probabilistic setting, the offers a new interpretation to the two-code part:

- First, the data is modeled with parameters θ. We first encode the θ since it's the essence/structure of a datum.
- Based on θ, we next encode the modeling error or how the information from X deviates from the θ structure.

This establishes a connection to the Bayesian inference literature since minimizing this expected code length is equivalent to finding the MAP estimate of θ given the data X.

Notice that for the continuous case, a finite code can encode a distribution up to a quantization tolerance ϵ and the code length would then be:

$$L(X) = E[-\log(P(\theta \mid H)\epsilon_\theta^{|\theta|}) - \log(p(X \mid \theta, H))\epsilon_X^{|X|}]$$

3 Bits-back Coding

Coined by Hinton and van Camp Bits-back coding [3] suggests using **redundant codewords** (a set of alternative parameter values instead of single values) based on some **auxiliary information** (the distribution of those parameters).

This leads to an overall shorter code length since the receiver can re-run the same learning algorithm as the sender and recover the auxiliary information (the distribution), thus getting his "bits back". The coding length of the auxiliary information can thus be subtracted from the total code length leading to a more efficient coding scheme.

Hinton introduced Bits-back in order to add noise and decrease the information contained in the parameters. While this sounds like adding information (mean and variance of the noise), this code is no longer than the original since we don't need to transmit the noise in the MDL framework and we can thus subtract the noise code. ***(View 1)***

As a concrete example, let's consider a model with an arbitrary distribution over the parameters $q(\theta)$ (arbitrary). This distribution introduces redundant information (we now have multiple θ values). The original coding scheme would return the following total length:

$$E_{q(\theta)}[L(X)] = E_{q(\theta)}[L(\theta)] + E_{q(\theta)}[L(X \mid \theta)] = -\sum_{\theta} q(\theta) \log P(\theta) - \sum_{\theta} q(\theta) \log P(X \mid \theta) \tag{4}$$

This coding scheme is inefficient since the length is greater than the optimal $L(X)$ derived earlier. It's crucial to observe that the approximate posterior (retrieved from the learning algorithm which the receiver has access to) can transmit additional information up to Shannon entropy of $q(\theta)$: $H(\theta) = -\sum_{\theta} q(\theta) \log P(\theta)$.
In other words, $H_q(\theta \mid X)$ can be used to carry other information than the original X and should not be included in the total length of X when the receiver has access to $q(\theta \mid X)$. ***(View 2)***
(View 3): $q(\theta \mid X)$ has already been communicated successfully since we have access to both the values θ and X and we should subtract it from the true cost of communicating the model and the misfits.
(View 4): By running the same learning algorithm and observing the choices that were made (based on the transmitted values of X) and the results (values of θ) the receiver can realize $q(\theta \mid X)$ which is not relevant to the model communication.

The code length is thus:

$$
\begin{aligned}
L_{q(\theta)}(X) &= E_{q(\theta)}[L(X)] - H_q(\theta \mid X) \\
&= \sum_{\theta} q(\theta \mid X) \log \frac{q(\theta \mid X)}{p(\theta)p(X \mid \theta)} \\
&= \sum_{\theta} q(\theta \mid X) \log \frac{q(\theta \mid X)}{p(X, \theta)} \\
&= KL(q(\theta \mid X) \mid\mid p(X, \theta)) \\
&= \sum_{\theta} q(\theta \mid X) \log \frac{q(\theta \mid X)}{p(\theta \mid X)} - \log p(X) \\
&= KL(q(\theta \mid X) \mid\mid p(\theta \mid X)) - \log p(X)
\end{aligned}
$$

Notice that the second line of this equation is the same as the classic result obtained from the classic scheme [Eq. 3] which can be regarded as a special cases where $q(\theta \mid X)$ is a point-mass distribution at θ_0 that maximizes $L_{q(\theta)}(X)$.
Retrieving the minimum length code reduces to minimizing this KL divergence term between the coding distribution (approximate posterior) and posterior distribution of the parameters.
Additionally, this code length is optimal when $q(\theta \mid X) = p(\theta \mid X)$ which provides a theoretical insight into the optimality of variational inference but we can almost never reach that equality in practice.

4 Combining Both Perspectives

The MDL and Bits-back view provides a nice way to derive the cost function (the ELBO: $L_{q(\theta)}(X) = E_{q(\theta)}[\frac{(\theta|X)}{p(X,\theta)}]$ as the KL divergence between an approximate posterior and the true posterior in addition to explaining the gap in the lower-bound based on the KL divergence between $p(\theta \mid X)$ and $q(\theta \mid X)$ which can cause the sub-optimality/inefficiency of the code.
This solves the issues of using an *ad hoc* measure and the reversed KL that a pure Bayesian perspective would find problematic.

5 Understanding Overpruning in Variational Autoencoders

Combining these complementary views can shed new light on practical problems in constructing and training models. Variational autoencders are such models.
The Variational Lossy Autoencoder [1] model provides a Bits-back interpretation of how the overpruning in classic VAE models is not an optimization challenge but is reasonable and probably even necessary in the cases where the latent code is not need for reconstruction (local information can be used instead, for example).
In that case, the training of the model would place the posterior of the latent code/parameter at the prior to avoid incurring a KL regularization cost and read optimal code length (for the specific X). This interpretation allowed an easier integration of auto-regressive structures in the decoder model without fully ignoring the encoder/latent representation.
Based on this interpretation, annealing the KL loss is an *ad hoc* solution that should not have a theoretical foundation and should probably be avoided.

References

[1] Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.

[2] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

[3] Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13. ACM, 1993.

[4] Antti Honkela and Harri Valpola. Variational learning and bits-back coding: an information-theoretic view to bayesian learning. *IEEE Transactions on Neural Networks*, 15(4):800–810, 2004.

[5] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[6] Tom Richardson and Ruediger Urbanke. *Modern coding theory*. Cambridge University Press, 2008.

[7] Jorma Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of statistics*, pages 416–431, 1983.

[8] Chris S Wallace and David L Dowe. Minimum message length and kolmogorov complexity. *The Computer Journal*, 42(4):270–283, 1999.