2nd Symposium on Advances in Approximate Bayesian Inference, 2019 1–6

Variational Predictive Information Bottleneck

Alexander A. Alemi
Google Research
ALEMI@GOOGLE.COM

Abstract

In classic papers, Zellner (1988, 2002) demonstrated that Bayesian inference could be derived as the solution to an information theoretic functional. Below we *derive* a generalized form of this functional as a variational lower bound of a predictive information bottleneck objective. This generalized functional encompasses most modern inference procedures and suggests novel ones.

1. Introduction

Consider a data generating process $\phi \sim p(\phi)$ from which we have some N draws that constitute our training set, $\boldsymbol{x_P} = \{x_1, x_2, \dots, x_N\} \sim p(x|\phi)$. We can also imagine (potentially infinitely many) future draws from this same process $\boldsymbol{x_F} = \{x_{N+1}, \dots\} \sim p(x|\phi)$. The *predictive information* $I(\boldsymbol{x_P}; \boldsymbol{x_F})$[1] gives a unique measure of the complexity of a data generating process (Bialek et al., 2001).

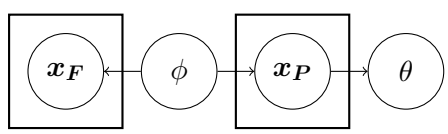

Figure 1: Graphical model under consideration.

The goal of *learning* is to capture this complexity. To perform learning, we form a global representation of the dataset $p(\theta|\boldsymbol{x_P})$. This can be thought of as a learning algorithm, that, given a set of observations, produces a summary statistic of the dataset that we hope is useful for predicting future draws from the same process. This algorithm could be deterministic or more generally, stochastic.

For example, imagine training a neural network on some data with stochastic gradient descent. Here the training data would be $\boldsymbol{x_P}$, the test data $\boldsymbol{x_F}$ and the neural network parameters would be θ. Our training procedure implicitly samples from the distribution $p(\theta|\boldsymbol{x_P})$.

How do we judge the utility of this learned global representation? The mutual information $I(\theta; \boldsymbol{x_F})$ quantifies the amount of information our representation captures about future draws.[2] To maximize learning we therefore aim to maximize this quantity.

1. We use $I(x; y)$ for the *mutual information* between two random variables: $I(x; y) \equiv \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right]$
2. It is interesting to note that in the limit of an infinite number of future draws, $I(\theta; \boldsymbol{x_F})$ approaches $I(\theta; \phi)$. Therefore, the amount of information we have about an infinite number of future draws from

© A.A. Alemi.

This is, of course, only interesting if we constrain how expressive our global representation is, for otherwise we could simply retain the full dataset. The amount of information retained about the observed data: $I(\theta; \boldsymbol{x_P})$ is a direct measure of our representation's complexity. The bits a learner extracts from data provides upper bounds on generalization (Bassily et al., 2017).

2. Predictive Information Bottleneck

Combined, these motivate the *predictive information bottleneck objective*, a generalized *information bottleneck* (Bialek et al., 2001; Tishby et al., 2000):

$$\max_{p(\theta|\boldsymbol{x_P})} I(\theta; \boldsymbol{x_F}) \quad \text{s.t.} \quad I(\theta; \boldsymbol{x_P}) = I_0. \tag{1}$$

We can turn this into an unconstrained optimization problem with the use of a Lagrange multiplier β:

$$\max_{p(\theta|\boldsymbol{x_P})} I(\theta; \boldsymbol{x_F}) - (1-\beta) I(\theta; \boldsymbol{x_P}). \tag{2}$$

While this objective seems wholly out of reach, we can make progress by noting that our random variables satisfy the Markov chain: $\boldsymbol{x_F} \leftarrow \phi \rightarrow \boldsymbol{x_P} \rightarrow \theta$, in which θ and $\boldsymbol{x_F}$ are conditionally independent given $\boldsymbol{x_P}$:

$$I(\theta; \boldsymbol{x_F}, \boldsymbol{x_P}) = I(\theta; \boldsymbol{x_F}) + I(\theta; \boldsymbol{x_P}|\boldsymbol{x_F}) = I(\theta; \boldsymbol{x_P}) + \cancel{I(\theta; \boldsymbol{x_F}|\boldsymbol{x_P})}. \tag{3}$$

This implies:

$$I(\theta; \boldsymbol{x_F}) = I(\theta; \boldsymbol{x_P}) - I(\theta; \boldsymbol{x_P}|\boldsymbol{x_F}). \tag{4}$$

and the equivalent unconstrained optimization problem:[3]

$$\min_{p(\theta|\boldsymbol{x_P})} I(\theta; \boldsymbol{x_P}|\boldsymbol{x_F}) - \beta I(\theta; \boldsymbol{x_P}). \tag{5}$$

The first term here: $I(\theta; \boldsymbol{x_P}|\boldsymbol{x_F})$ is the residual information between our global representation and the dataset after we condition on full knowledge of the data generating procedure. This is a direct measure of the inefficiency of our proposed representation.

3. Variational Predictive Information Bottleneck

Simple variational bounds (Poole et al., 2019) can be derived for this objective, just as was done for the (local) information bottleneck objective in Alemi et al. (2016). First, we demonstrate a variational upper bound on $I(\theta; \boldsymbol{x_P}|\boldsymbol{x_F})$: [4]

$$I(\theta; \boldsymbol{x_P}|\boldsymbol{x_F}) = \left\langle \log \frac{p(\theta|\boldsymbol{x_P})}{p(\theta|\boldsymbol{x_F})} \right\rangle \leq \left\langle \log \frac{p(\theta|\boldsymbol{x_P})}{q(\theta)} \right\rangle. \tag{6}$$

the process is the same as the amount of information we have about the nature and identity of the data generating process itself.

3. A similar transformation for the (local) variational information bottleneck appeared in Fischer (2019).
4. $\langle \cdot \rangle$ is used to denote expectations, and unless denoted otherwise with respect to the full joint density $p(\theta|\boldsymbol{x_P})p(\boldsymbol{x_P}|\phi)p(\phi)p(\boldsymbol{x_F}|\phi)$

Here we upper bound the residual information by using a variational approximation to $p(\theta|\boldsymbol{x_F})$, the marginal of our global representation over all datasets drawn from the same data generating procedure. Any distribution $q(\theta)$ independent of $\boldsymbol{x_F}$ suffices.

Next we variationally lower bound $I(\theta;\boldsymbol{x_P})$ with:

$$I(\theta;\boldsymbol{x_P}) = \left\langle \log \frac{p(\boldsymbol{x_P}|\theta)}{p(\boldsymbol{x_P})} \right\rangle \geq H(\boldsymbol{x_P}) + \sum_i \langle \log q(x_i|\theta) \rangle \,. \tag{7}$$

The entropy of the training data $H(\boldsymbol{x_P})$ is a constant outside of our control that can be ignored. Here we variationally approximate the "posterior" of our global representation with a factorized "likelihood": $\prod_i q(x_i|\theta) = q(\boldsymbol{x_P}|\theta) \sim p(\boldsymbol{x_P}|\theta)$. Notice that while $p(\boldsymbol{x_P}|\theta)$ will not factorize in general, we can certainly consider a family of variational approximations that do.

Combining these variational bounds, we generate the objective:

$$\min_{p(\theta|\boldsymbol{x_P})} \left\langle \log \frac{p(\theta|\boldsymbol{x_P})}{q(\theta)} - \beta \sum_i \log q(x_i|\theta) \right\rangle . \tag{8}$$

We have thus *derived*, as a variational lower bound on the predictive information bottleneck, the objective Zellner (1988) postulates (with $\beta = 1$) is satisfied for inference procedures that optimally process information. As Knoblauch et al. (2019) demonstrates, this encompasses a wide array of modern inference procedures, including Generalized Bayesian Inference (Bissiri et al., 2016) and a generalized Variational Inference, dubbed Gibbs VI (Alquier et al., 2016; Futami et al., 2017).[5] Below we highlight some of these and other connections.

4. Connections

If, in Equation (8), we identity $q(\theta)$ with a *fixed prior* and $q(x|\theta)$ with a *fixed likelihood* of a generative model, optimizing this objective for $p(\theta|\boldsymbol{x_P})$ in the space of all probability densities gives the generalized Boltzmann distribution (Jaynes, 1957):

$$p(\theta|\boldsymbol{x_P}) = q(\theta) \left[\prod_i q(x_i|\theta) \right]^\beta / \mathcal{Z}, \tag{9}$$

where $\mathcal{Z}$ is the *partition function.* [6] This is a generalized form of Bayesian Inference called the *power likelihood* (Holmes and Walker, 2017; Royall and Tsou, 2003). Here the inverse temperature β acts as a Lagrange multiplier controlling the trade-off between the amount of information we retain about our observed data ($I(\theta;\boldsymbol{x_P})$) and how much predictive information we capture ($I(\theta;\boldsymbol{x_F})$). As $\beta \to \infty$ (temperature goes to zero), we recover the maximum likelihood solution. At $\beta = 1$ (temperature $= 1$) we recover ordinary Bayesian inference. As $\beta \to 0$ (temperature goes to infinity), we recover just prior predictive inference that ignores the data entirely. These limits are summarized in Table 1.

5. To incorporate the Generalized VI (Knoblauch et al., 2019) with divergence measures other than KL, we need only replace our mutual informations (which are KL based) with their corresponding generalizations.
6. $\mathcal{Z} \equiv \int d\theta \, q(\theta) \left[\prod_i q(x_i|\theta) \right]^\beta$

Limit	Inference	Equivalent Objective	$p(\theta\|\boldsymbol{x_P})$
β	Generalized Bayes	$\max I(\theta;\boldsymbol{x_F}) - (1-\beta)I(\theta;\boldsymbol{x_P})$	$\propto q(\theta)q(\boldsymbol{x_P}\|\theta)^\beta$
$\beta \to 0$	Prior Predictive	$\min I(\theta;\boldsymbol{x_P}\|\boldsymbol{x_F})$	$q(\theta)$
$\beta \to 1$	Bayesian	$\max I(\theta;\boldsymbol{x_F})$	$\propto q(\theta)q(\boldsymbol{x_P}\|\theta)$
$\beta \to \infty$	Maximum Likelihood	$\max I(\theta;\boldsymbol{x_P})$	$\arg\min_\theta q(\boldsymbol{x_P}\|\theta)$

Table 1: Power Bayes can be recovered as a variational lower bound on the predictive information bottleneck objective (Equation (5)).

More generally, notice that in Equation (8) the densities $q(x|\theta)$ and $q(\theta)$ are not literally the likelihood and prior of a generative model, they are variational approximations that we have complete freedom to specify. This allows us to describe other more generalized forms of Bayesian inference such as Divergence Bayes or the full Generalized Bayes (Knoblauch et al., 2019; Bissiri et al., 2016) provided we can interpret the chosen loss function as a conditional distribution.

If we limit the domain of $p(\theta|\boldsymbol{x_P})$ to a restricted family of parametric distributions, we immediately recover not only standard variational inference, but a broad generalization known as Gibbs Variational Inference (Knoblauch et al., 2019; Alquier et al., 2016; Futami et al., 2017).

Furthermore, nothing prevents us from making $q(x|\theta)$ or $q(\theta)$ themselves parametric and simultaneously optimizing those. Optimizing the prior with a fixed likelihood, unconstrained $p(\theta|\boldsymbol{x_P})$, and $\beta = 1$ the objective mirrors Empirical Bayesian (Maritz and Lwin, 2018) approaches, including the notion of *reference priors* (Mattingly et al., 2018; Berger et al., 2009). Alternatively, optimizing a parametric likelihood with a parametric representation $p(\theta|\boldsymbol{x_P})$, fixed prior, and $\beta = 1$ equates to a Neural Process (Garnelo et al., 2018).

Consider next *data augmentation*, where we have some stochastic process that modifies our data with implicit conditional density $t(x'|x)$. If the augmentation procedure is centered about zero so that $\langle x' \rangle_{t(x'|x)} = x$ and our chosen likelihood function is concave, then we have:

$$\log q(x|\theta) = \log q(\langle x' \rangle_{x' \sim t(x'|x)}|\theta) \geq \langle \log q(x'|\theta) \rangle_{x' \sim t(x'|x)}\,, \tag{10}$$

which maintains our bound. For example, for an exponential family likelihood and any centered augmentation procedure (like additive mean zero noise), doing generalized Bayesian inference on an augmented dataset is also a lower bound on the predictive information bottleneck objective.

5. Conclusion and Future Work

We have shown that a wide range of existing inference techniques are variational lower bounds on a single predictive information bottleneck objective.

This connection highlights the drawbacks of these traditional forms of inference. In all cases considered in the previous section, we made two choices that loosened our variational bounds. First, we approximated $p(\boldsymbol{x_P}|\theta)$, with a factorized approximation $q(\boldsymbol{x_P}|\theta) = \prod_i q(x_i|\theta)$. Second, we approximated the future conditional *marginal* $p(\theta|\boldsymbol{x_F}) =$

$\int d\boldsymbol{x_P}\, p(\theta|\boldsymbol{x_P})p(\boldsymbol{x_P}|\boldsymbol{x_F})$ as an unconditional "prior". Neither of these approximations is necessary.

For example, consider the following tighter "prior":

$$q(\theta|\boldsymbol{x_F}) \sim \int d\boldsymbol{x_P}'\, p(\theta|\boldsymbol{x_P}')q(\boldsymbol{x_P}'|\boldsymbol{x_F}). \tag{11}$$

Here we reuse a tractable global representation $p(\theta|\boldsymbol{x_P})$ and instead create a variational approximation to the density of alternative datasets drawn from the same process: $q(\boldsymbol{x_P}'|\boldsymbol{x_F})$.

We believe this information-theoretic, representation-first perspective on learning has the potential to motivate new and better forms of inference. [7]

Acknowledgments

Acknowledgements go here.

References

Alexander A. Alemi and Ian Fischer. TherML: Thermodynamics of machine learning, 2018.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of gibbs posteriors. *Journal of Machine Learning Research*, 17(236):1–41, 2016. URL http://jmlr.org/papers/v17/15-290.html.

Raef Bassily, Shay Moran, Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. Learners that use little information. *arXiv preprint arXiv:1710.05233*, 2017.

James O Berger, José M Bernardo, Dongchu Sun, et al. The formal definition of reference priors. *The Annals of Statistics*, 37(2):905–938, 2009.

William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural computation*, 13(11):2409–2463, 2001.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016. doi: 10.1111/rssb.12158. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12158.

Ian Fischer. The conditional entropy bottleneck, 2019. URL https://openreview.net/forum?id=rkVOXhAqY7.

Futoshi Futami, Issei Sato, and Masashi Sugiyama. Variational inference based on robust divergences. *arXiv preprint arXiv:1710.06595*, 2017.

7. Extensions of this representation-first perspective beyond a single global representation towards supervised or unsupervised learning can be found in Alemi and Fischer (2018).

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018.

CC Holmes and SG Walker. Assigning a value to a power likelihood in a general bayesian model. *Biometrika*, 104(2):497–503, 2017.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Generalized variational inference. *arXiv preprint arXiv:1904.02063*, 2019.

Johannes S Maritz and T Lwin. *Empirical bayes methods*. Routledge, 2018.

Henry H Mattingly, Mark K Transtrum, Michael C Abbott, and Benjamin B Machta. Maximizing the information learned from finite data selects a simple model. *Proceedings of the National Academy of Sciences*, 115(8):1760–1765, 2018.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A Alemi, and George Tucker. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*, 2019.

Richard Royall and Tsung-Shan Tsou. Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):391–404, 2003.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Arnold Zellner. Optimal information processing and bayes's theorem. *The American Statistician*, 42(4):278–280, 1988.

Arnold Zellner. Information processing and bayesian analysis. *Journal of Econometrics*, 107 (1-2):41–50, 2002.