Some comments on Deterministic Information Bottleneck

Sarah E. Marzen[1,*]

[1] *W. M. Keck Science Department, Claremont CA 91711*

(Dated: May 8, 2024)

We make the case that although Deterministic Information Bottleneck may be a contribution to clustering, it should not be used to aid lossy compression without the addition of blocklength. We therefore suggest a new objective function that does so and leave its testing to future work.

PACS numbers: 02.50.-r 89.70.+c 05.45.Tp
Keywords: –

In 2017, the "Deterministic Information Bottleneck" (DIB) method was published and is now highly cited. The basic idea is that according to the information bottleneck method [1], an extension of rate-distortion theory [2] to an informational distortion measure, one should minimize expected distortion and also minimize the mutual information between the compressed signal and the original variable at the same time. The information bottleneck method is great for clustering, e.g. as in Ref. [3], and it also is a contribution to information theory [4], but if one is not careful, one might misunderstand where the objective function comes from. In this short arXiv paper, we simply aim to describe where these rate-distortion objectives come from and why DIB is a contribution to clustering but not to information theory as it stands. We then modify the objective so that it might be a contribution to information theory and leave the testing for later work.

From one of my own papers [5]: "In the classic rate-distortion setup, one sends a sequence of n letters $x_{0:n}$ to an encoder that chooses one of M words for those n letters and then sends that word to a decoder which produces a guess as to what those letters were, $\hat{x}_{0:n}$. The material constraint is actually $\log M/n$, not a mutual information. This corresponds to a more intuitive notion of resource constraints in the biological sense– number of molecules or number of neurons, normalized by "blocklength" n. Some distortion measure is defined, $d(x, \hat{x})$, which in some extensions can be a distortion of the entire block $x_{0:n}$ relative to $\hat{x}_{0:n}$ rather than letter-by-letter. There are some rates $\log M/n$ and distortions $\sum d(x_i, \hat{x}_i)/n$ that are achievable and some that are unachievable given any combination of encoder and decoder. A theorem shows that the curve separating achievable from unachievable is given by replacing the rate $\log M/n$ with a mutual information $I[X; \hat{X}]$ and the average distortion with an expected distortion if all is memoryless. This curve is accurate in the limit that blocklength n goes to infinity. Otherwise, the rate-distortion curve that separates achievable from unachievable is given by $R_n(D)$ rather than $R(D)$, and $R_n(D)$ is horribly difficult to calculate [2]."

In the information bottleneck method, one replaces $d(x, \hat{x})$ with an informational distortion measure (a Kullback-Leibler divergence), which does not correspond to the classic rate-distortion setup, in which $d(\hat{x}, x)$ does not depend on $P(Y|\hat{X} = \hat{x})$. As a result, the objective in the information bottleneck method becomes $I[R; Y] - \beta I[X; R]$, where R is the compressed variable, X is the variable that we compress, and Y is the variable that we try to retain information about. There is a rate-distortion theorem justifying this objective [4].

The wonderful thing about the rate-distortion theorem is that it allows one to test if a sensor is near-optimal. One calculates the rate and expected distortion and places it relative to the rate-distortion curve, as in Ref. [6]. It is also known that the actual soft clusters obtained in order to get the rate-distortion curve are terrible for coding, as their coding rate is given by $H[R]$ rather than by $I[R; X]$ [2], although they may be useful for machine learning as stated earlier. As such, they should not be used as a null model to decide if an information-theoretic objective is successful at producing good codes.

DIB seems to be a fix to the need for clusters with a good coding rate. In DIB, we replace $I[X; R]$ with $H[R] + \lambda H[R|X]$ for some other λ and we take the limit of determinism, where λ becomes very large. However, there is no rate-distortion theorem justifying this objective as revealing sensors that are closer to optimal, and in fact, there should not be, as the objective that reveals if a sensor is close to the boundary between achievable and unachievable is the original information bottleneck objective function.

We might ask– does DIB lead to good information theory codes? The objective indeed leads to deterministic codes. Deterministic codes are what one would desire for lossy compression, but it is well-known that the best codes often require large blocklengths n, though there

* smarzen@cmc.edu

are exceptions. We therefore simply propose that perhaps the objective function of $\langle d(\hat{x}_{0:n}, x_{0:n})\rangle + \beta H[R] + \alpha H[R|X_{0:n}] - \lambda H[\hat{X}_{0:n}|R]$ should be investigated for information theorists, despite the computational overhead that is incurred by including blocklength.

All of this does not mean that DIB is not great at getting interesting clusters for machine learning and data science purposes.

---

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

[2] Thomas Berger. *Rate distortion theory: A mathematical basis for data compression*. Prentice-Hall, Inc., 1971.

[3] Susanne Still, William Bialek, and Léon Bottou. Geometric clustering using the information bottleneck method. *Advances in neural information processing systems*, 16, 2003.

[4] Michael Gastpar, Bixio Rimoldi, and Martin Vetterli. To code, or not to code: Lossy source-channel communication revisited. *IEEE Transactions on Information Theory*, 49(5):1147–1158, 2003.

[5] Sarah Marzen. Resource-rational reinforcement learning and sensorimotor causal states. *arXiv preprint arXiv:2404.18775*, 2024.

[6] Stephanie E Palmer, Olivier Marre, Michael J Berry, and William Bialek. Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, 112(22):6908–6913, 2015.