# Perturbation Theory for the Information Bottleneck

**Vudtiwat Ngampruetikorn,*** **David J. Schwab**
Initiative for the Theoretical Sciences, CUNY Graduate Center
*vngampruetikorn@gc.cuny.edu

## Abstract

The information bottleneck (IB) method of Tishby, Pereira and Bialek formalizes the notion of extracting relevant information from data. While the IB method offers a precise and appealing framework for understanding learning phenomena, it is analytically intractable in general. Here we derive a perturbation theory for the IB method and analyze the learning onset – the limit of maximum relevant information per each bit, extracted from data. We test our results on a synthetic probability distribution, finding good agreement with the exact solution near the onset of learning. Our work also provides a fresh perspective on the intimate relation between the IB method and the strong data processing inequality.

## 1 Information Bottleneck

Extracting relevant information from data is crucial for all forms of learning. Animals are very adept at isolating biologically useful information from complicated real-world sensory stimuli: for example, we instinctively ignore pixel-level noise when looking for a face in a photo. A failure to disregard irrelevant bits could lead to suboptimal generalization performance especially when the data contains spurious correlations. For instance, an image classifier that relies on background texture to identify objects is likely to fail when presented with a new image showing an object in an 'unusual' background (see, e.g., Refs. [7, 20]). Understanding the principles behind the identification and extraction of relevant bits is therefore of fundamental and practical importance.

Formalizing this aspect of learning, the information bottleneck (IB) method provides a precise notion of relevance with respect to a prediction target: the relevant information in a source ($X$) is the bits that carry information about the target ($Y$) [17]. The relevant bits in $X$ are summarized in a representation ($Z$) via a stochastic map defined by an encoder $q(z|x)$, obeying the Markov constraint $Z \leftrightarrow X \leftrightarrow Y$. In general a trade-off exists between the amount of discarded information (compression) and the remaining relevant information in $Z$ (prediction), thus motivating the IB cost function,

$$L[q(z|x)] = I(Z;X) - \beta I(Z;Y), \tag{1}$$

where $\beta > 0$ denotes the trade-off parameter and $I(A;B)$ the mutual information. The first term favors succinct representations whereas the second encourages predictive ones. The IB loss is minimized by the representations that are most predictive of $Y$ at fixed compression, parametrized by the Lagrange multiplier $\beta$ (see, Fig. 1a).

The IB method offers a highly versatile framework with wide-ranging applications, including neural coding [10], evolutionary population dynamics [13], clustering [16], deep learning [1–3] and reinforcement learning [8]. However the nonlinearity of the IB problem makes it computationally expensive and difficult to analyze, barring a few special cases [5]. This necessitates an investigation of tractable methods for solving the IB problem. The use of variational approximations to reduce the computational cost has paved the way for a massive scale-up of the IB method [3]. Complementing this approach, we report a new analytical result for the IB problem in the tractable limits.

Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), Vancouver, Canada.

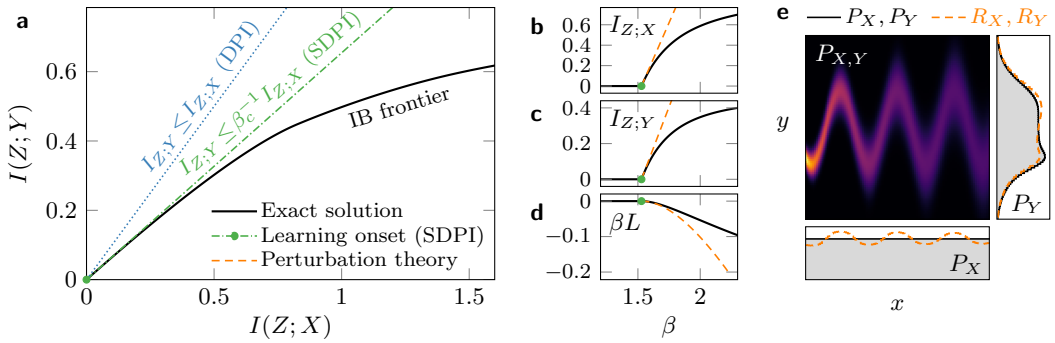


Figure 1: **Information bottleneck & Learning onset.** (a) The IB frontier (solid) is parametrized by the trade-off parameter $\beta$ whose inverse is the slope of this curve. The relevant information is bounded from above by the data processing inequality (DPI) and its tight version, the strong data processing inequality (SDPI) [Eq. (2)] which touches the IB curve at the origin. The slope at the origin is equal to the inverse critical trade-off parameter $\beta_c^{-1}$ which marks the learning onset (circles in (b-d)). (b-d) Our controlled expansions *vs* the exact solution for $P_{X,Y}$ shown in (e) [see legend in (a)]. We obtain the SDPI and the perturbation theory results from Eqs. (13-15) & (21). All information is in bits.

## 2 Learning Onset

Although the IB loss in Eq. (1) encourages a representation to encode every relevant bit in $X$ when $\beta \to \infty$,[1] the optimal representation needs not contain any relevant information at finite $\beta$. To see this, we note that the loss vanishes for any uninformative representation $I(Z;X) = I(Z;Y) = 0$, and thus an informative representation yields a lower loss only when the relevant information in $Z$ is adequately large $I(Z;Y) > \beta^{-1} I(Z;X)$ (which makes the loss negative). But the relevant information is also bounded from above by the data processing inequality (DPI), $I(Z;Y) \leq I(Z;X)$, resulting from the the Markov constraint $Z \leftrightarrow X \leftrightarrow Y$ [6] (see, Fig. 1a). Combining these inequalities yields $\beta^{-1} I(Z;X) < I(Z;Y) \leq I(Z;X)$ which cannot be met when $\beta^{-1} > 1$. Hence the existence of an informative IB minimizer requires $\beta^{-1} \leq 1$. Indeed for any $P_{X,Y}$ with $I(X;Y) > 0$, there exists a critical trade-off parameter $\beta_c(X \to Y) \geq 1$ that marks the learning onset, separating two qualitatively distinct regimes: uninformative regime at $\beta < \beta_c$ and informative regime at $\beta > \beta_c$.

The learning onset is not only a special limit in the IB problem but also physically and practically relevant. It corresponds to the region where the relevant information per encoded bit is greatest and thus places a tight bound on the thermodynamic efficiency of predictive systems [14, 15]. The (inverse) critical trade-off parameter is also a useful measure of correlation between two random variables [9]. Finally estimating the upper bound of $\beta_c$ might help weed out non-viable values of hyperparameters in deep learning techniques such as the variational information bottleneck [18, 19].

**The strong data processing inequality.** We can improve the bound on $\beta_c$ with the tight version of the DPI, the strong data processing inequality (SDPI) [4, 11, 12] (see, Fig. 1a)

$$I(Z;Y) \leq \eta_{\mathrm{KL}}(X \to Y) I(Z;X) \quad \text{with} \quad \eta_{\mathrm{KL}}(X \to Y) \equiv \sup_{R_X \neq P_X} \frac{\mathrm{D}_{\mathrm{KL}}(R_Y || P_Y)}{\mathrm{D}_{\mathrm{KL}}(R_X || P_X)}, \tag{2}$$

where $\eta_{\mathrm{KL}}(X \to Y)$ is the contraction coefficient for the Kullback-Leibler divergence. Here $P_X$ and $P_Y$ denote the probability distributions of $X$ and $Y$. The supremum is over all allowed distributions given the space of $X$, and $R_Y$ is related to $R_X$ via the channel $P_{Y|X}$. Replacing the DPI with the SDPI in the first paragraph of this section, we obtain

$$\beta_c(X \to Y) \geq \eta_{\mathrm{KL}}(X \to Y)^{-1}. \tag{3}$$

In the following section we show that the equality holds, as expected (since the SDPI is tight).[2]

[1] The compression term, while infinitesimally small in this limit, still penalises irrelevant information and prefers a representation $Z$ that is the minimal sufficient statistics of $X$ for $Y$.

[2] In general $\eta_{\mathrm{KL}}(X \to Y)$ and $\beta_c(X \to Y)$ are asymmetric in $X$ and $Y$.

## 3 Perturbation theory

Our theory is based on a controlled expansion around the critical trade-off parameter $\beta_c$ and some uninformative encoder $q_0(z|x) = q_0(z)$,

$$q(z|x) = q_0(z|x) + \varepsilon q_1(z|x) + \varepsilon^2 q_2(z|x) + \dots \tag{4}$$

$$I(Z;X) = \varepsilon I^{(1)}_{Z;X}[q_1] + \varepsilon^2 I^{(2)}_{Z;X}[q_1, q_2] + \dots, \tag{5}$$

where $\varepsilon \equiv \beta - \beta_c \to 0^+$ and $\sum_z q_n(z|x) = \delta_{n,0}$ to ensure normalization. Note that $I^{(0)}_{Z;X}$ vanishes for uninformative $q_0$. The first and second-order informations read

$$I^{(1)}_{Z;X}[q_1] = \sum\nolimits_x p(x) \sum\nolimits_{z\in\mathcal{Z}_1} q_1(z|x) \ln \frac{q_1(z|x)}{q_1(z)} \tag{6}$$

$$I^{(2)}_{Z;X}[q_1, q_2] = \sum\nolimits_x p(x) \Bigg( \sum\nolimits_{z\in\mathcal{Z}_0} \frac{q_1(z|x)^2 - q_1(z)^2}{2q_0(z)} + \sum\nolimits_{z\in\mathcal{Z}_1} q_2(z|x) \ln \frac{q_1(z|x)}{q_1(z)} + \sum\nolimits_{z\in\mathcal{Z}_2} q_2(z|x) \ln \frac{q_2(z|x)}{q_2(z)} \Bigg), \tag{7}$$

where $\mathcal{Z}_0 = \mathrm{supp}(q_0)$ and $\mathcal{Z}_n = \mathrm{supp}(q_n) \setminus \bigcup_{i=0}^{n-1} \mathcal{Z}_i$ (i.e., $\mathcal{Z}_n$ contains representation classes or space that first appear in the support of the $n$th-order encoder).[3] The expansions for $q(z)$ and $q(z|y)$ take the same form as Eq. (4), and the expressions for $I(Z;Y)$ are given by Eqs. (5-7) but with $Y$ replacing $X$ everywhere. Finally we write down the loss function as a power series in $\varepsilon$,

$$L[q(z|x)] = \varepsilon L^{(1)}[q_1] + \varepsilon^2 L^{(2)}[q_1, q_2] + \dots, \tag{8}$$

where

$$L^{(1)}[q_1] = I^{(1)}_{Z;X}[q_1] - \beta_c I^{(1)}_{Z;Y}[q_1] \tag{9}$$

$$L^{(2)}[q_1, q_2] = I^{(2)}_{Z;X}[q_1, q_2] - \beta_c I^{(2)}_{Z;Y}[q_1, q_2] - I^{(1)}_{Z;Y}[q_1]. \tag{10}$$

**First-order theory.** Minimizing the first-order loss yields[4]

$$\min L^{(1)} = L^{(1)}[q_1^*] = 0 \quad \text{with} \quad \frac{q_1^*(z|x)}{q_1^*(z)} = \exp\left(\beta_c \sum\nolimits_y p(y|x) \ln \frac{q_1^*(z|y)}{q_1^*(z)}\right) \text{ for } z \in \mathcal{Z}_1. \tag{11}$$

As the ratio $q_1(z|x)/q_1(z)$ does not depend on $z$, we eliminate the superfluous dependence on $z$ by defining[5]

$$r(x) \equiv \frac{q_1(z|x)p(x)}{q_1(z)} \quad \text{for } z \in \mathcal{Z}_1, \text{ and } \quad r(y) = \sum\nolimits_x p(y|x) r(x). \tag{12}$$

Substituting Eqs. (12) in (6) & (11), we obtain

$$I^{(1)}_{Z;X} = \mathrm{D_{KL}}[r(x)||p(x)] \sum\nolimits_{z\in\mathcal{Z}_1} q_1(z), \qquad I^{(1)}_{Z;Y} = \mathrm{D_{KL}}[r(y)||p(y)] \sum\nolimits_{z\in\mathcal{Z}_1} q_1(z) \tag{13}$$

$$r(x) = p(x) \exp\left(-\beta_c (\mathrm{D_{KL}}[p(y|x)||r(y)] - \mathrm{D_{KL}}[p(y|x)||p(y)])\right). \tag{14}$$

Since the first-order loss vanishes [Eq. (11)], we have $I^{(1)}_{Z;X}[q_1^*] - \beta_c I^{(1)}_{Z;Y}[q_1^*] = 0$ and thus

$$\beta_c = \frac{\mathrm{D_{KL}}[r(x)||p(x)]}{\mathrm{D_{KL}}[r(y)||p(y)]}. \tag{15}$$

Note that an uninformative solution $r(x) = p(x)$ always satisfies Eq. (14) and we must seek a nontrivial solution $r(x) \neq p(x)$.

[3] Our theory generalizes the expansions in Refs. [18, 19] which considered the case $\mathcal{Z}_1 = \mathcal{Z}_2 = \emptyset$.

[4] Unlike in the original IB problem, here the optimization is unconstrained since the normalization $\sum_z q_1(z|x) = 0$ sums over both $\mathcal{Z}_0$ and $\mathcal{Z}_1$, and only the latter enters our first-order theory.

[5] Both $r(x)$ and $r(y)$ are non-negative and normalized: $\sum_x r(x) = \sum_y r(y) = 1$.

We now show that the critical trade-off parameter is the inverse contraction coefficient. First we note that $r(x)$ in Eq. (14) is a solution to a different optimization, described by a loss function $\mathcal{L}[f] = \mathrm{D_{KL}}[f(x)||p(x)] - \beta_c \, \mathrm{D_{KL}}[f(y)||p(y)]$. That is, $\delta\mathcal{L}/\delta f|_{f\to r} = 0$ and $\min \mathcal{L} = \mathcal{L}[r] = 0$. It follows immediately that $\delta(\frac{\mathrm{D_{KL}}[f(y)||p(y)]}{\mathrm{D_{KL}}[f(x)||p(x)]})/\delta f|_{f\to r} = 0$ for $\mathrm{D_{KL}}[r(x)||p(x)] > 0$, therefore

$$\beta_c^{-1} = \frac{\mathrm{D_{KL}}[r(y)||p(y)]}{\mathrm{D_{KL}}[r(x)||p(x)]} = \sup_{f\neq p} \frac{\mathrm{D_{KL}}[f(y)||p(y)]}{\mathrm{D_{KL}}[f(x)||p(x)]} = \eta_{\mathrm{KL}}(X \to Y). \tag{16}$$

While our first-order theory provides a method for identifying the critical trade-off parameter by solving Eqs. (14) & (15), it is incomplete. The optimal encoder in Eq. (11) is determined up to a multiplicative factor. Consequently the informations in Eq. (13) still depend on $q_1(z)$ which can take any positive value. This scale invariance is unphysical and is broken in the second-order theory.

**Second-order theory.** From Eqs. (7) & (10), we write down the second-order loss

$$L^{(2)}[q_1, q_2] = \sum\nolimits_{z\in\mathcal{Z}_0} \frac{\sum_{x,x'} q_1(z|x)K(x,x')q_1(z|x')}{2q_0(z)} - I^{(1)}_{Z;Y}[q_1] \tag{17a}$$

$$+ \sum\nolimits_x p(x) \sum\nolimits_{z\in\mathcal{Z}_1} q_2(z|x) \left( \ln \frac{q_1(z|x)}{q_1(z)} - \beta_c \sum\nolimits_y p(y|x) \ln \frac{q_1(z|y)}{q_1(z)} \right) \tag{17b}$$

$$+ \sum\nolimits_x p(x) \sum\nolimits_{z\in\mathcal{Z}_2} q_2(z|x) \left( \ln \frac{q_2(z|x)}{q_2(z)} - \beta_c \sum\nolimits_y p(y|x) \ln \frac{q_2(z|y)}{q_2(z)} \right) \tag{17c}$$

where $K(x,x') \equiv \delta(x,x')p(x) + (\beta_c - 1)p(x)p(x') - \beta_c \sum_y p(y)p(x|y)p(x'|y)$. Optimizing $L^{(2)}$ with respect to $q_2$ (for $\mathcal{Z}_1$ and $\mathcal{Z}_2$ separately) results in stationary conditions, which equate the terms in the parentheses of Eqs. (17b) & (17c) to zero.[6] Eliminating $I^{(1)}_{Z;Y}$ with Eq. (13), we have

$$L^{(2)}[q_1] = -\,\mathrm{D_{KL}}[r(y)||p(y)] \sum\nolimits_{z\in\mathcal{Z}_1} q_1^*(z) + \sum\nolimits_{z\in\mathcal{Z}_0} \frac{\sum_{x,x'} q_1(z|x)K(x,x')q_1(z|x')}{2q_0(z)}. \tag{18}$$

Minimizing this loss with respect to $q_1$ and subject to the normalization $\sum_z q_1(z|x) = 0$ gives

$$\sum\nolimits_{x'} K(x,x') \frac{q_1^*(z|x')}{q_0(z)} = -\left(\sum\nolimits_{z'\in\mathcal{Z}_1} q_1(z')\right) \sum\nolimits_{x'} K(x,x') \frac{r(x')}{p(x')} \quad \text{for } z \in \mathcal{Z}_0. \tag{19}$$

Substituting the above in Eq. (18) yields

$$L^{(2)}[q_1^*] = -\,\mathrm{D_{KL}}[r(y)||p(y)] \sum_{z\in\mathcal{Z}_1} q_1(z) + \frac{1}{2} \left( \sum_{x,x'} \frac{r(x)K(x,x')r(x')}{p(x)p(x')} \right) \left( \sum_{z\in\mathcal{Z}_1} q_1(z) \right)^2. \tag{20}$$

The final minimization with respect to $\sum_{z\in\mathcal{Z}_1} q_1(z)$ results in

$$\min L^{(2)} = -\frac{1}{2\kappa}\,\mathrm{D_{KL}}[r(y)||p(y)]^2 \quad \text{and} \quad \sum\nolimits_{z\in\mathcal{Z}_1} q_1^*(z) = \frac{1}{\kappa}\,\mathrm{D_{KL}}[r(y)||p(y)], \tag{21}$$

where $\kappa = \sum_{x,x'} \frac{r(x)K(x,x')r(x')}{p(x)p(x')} > 0$. These results break the scale invariance in our first-order theory [Eq. (11)] and fix the leading corrections to mutual information in Eq. (13).

In Fig. 1 we demonstrate that our theory [Eqs. (13-15) & (21)] correctly predicts the critical trade-off parameter and captures the behaviors of the mutual information and IB loss in the vicinity of the learning onset for a synthetic joint distribution (shown in Fig. 1e).

## 4 Outlook

We derive a perturbation theory for the IB problem and offer a glimpse of the intimate connections between the learning onset and the strong data processing inequality. In future works we aim to build on our results to develop an algorithm for estimating the contraction coefficient from samples and explore novel methods for solving the IB problem in this limit. It would be interesting to further leverage the wealth of rigorous results from the literature on hypercontractivity and strong data processing inequalities to better understand the learning onset in the IB problem. In addition, various numerical techniques developed for the IB problem could significantly extend the range of applicability of contraction coefficients.

[6] This optimization is unconstrained since the second-order loss does not depend on $q_2$ with $z \in \mathcal{Z}_0$ (see, footnote 4). The resulting stationary conditions are identical to Eq. (11) for $q_1$ with $z \in \mathcal{Z}_1$ and $q_2$ with $z \in \mathcal{Z}_2$.

## Broader Impact

Our work expects to benefit researchers working on the information bottleneck, hypercontractivity, strong data processing inequality, and related problems. We do not anticipate that this work would advantage or disadvantage any group.

## Acknowledgments and Disclosure of Funding

We thank Shervin Parsi and Sarang Gopalakrishnan for useful discussions. This work was supported in part by the National Institutes of Health under award number R01EB026943 and the National Science Foundation, through the Center for the Physics of Biological Function (PHY-1734030), and through the Simons Foundation.

## References

[1] A Achille and S Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2897, 2018. doi:10.1109/TPAMI.2017.2784440.

[2] A Achille and S Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1, 2018. `http://jmlr.org/papers/v19/17-646.html`.

[3] AA Alemi, I Fischer, JV Dillon, and K Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. `https://openreview.net/forum?id=HyxQzBceg`.

[4] V Anantharam, AA Gohari, S Kamath, and C Nair. On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover. `http://arxiv.org/abs/1304.6133`.

[5] G Chechik, A Globerson, N Tishby, and Y Weiss. Information bottleneck for Gaussian variables. *Journal of Machine Learning Research*, 6:165, 2005. `https://www.jmlr.org/papers/v6/chechik05a.html`.

[6] TM Cover and JA Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd ed, 2006.

[7] Y Dubois, D Kiela, DJ Schwab, and R Vedantam. Learning optimal representations with the decodable information bottleneck. 2020. `https://arxiv.org/abs/2009.12789`.

[8] A Goyal, R Islam, DJ Strouse, Z Ahmed, H Larochelle, M Botvinick, S Levine, and Y Bengio. Transfer and exploration via the information bottleneck. In *International Conference on Learning Representations*, 2019. `https://openreview.net/forum?id=rJg8yhAqKm`.

[9] H Kim, W Gao, S Kannan, S Oh, and P Viswanath. Discovering potential correlations via hypercontractivity. In *Advances in Neural Information Processing Systems 30*, 2017. `http://papers.nips.cc/paper/7044-discovering-potential-correlations-via-hypercontractivity.pdf`.

[10] SE Palmer, O Marre, MJ Berry II, and W Bialek. Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, 112(22):6908, 2015. doi:10.1073/pnas.1506855112.

[11] Y Polyanskiy and Y Wu. Strong data-processing inequalities for channels and Bayesian networks. In *Convexity and Concentration*, Springer, New York, 2017. doi:10.1007/978-1-4939-7005-6_7.

[12] M Raginsky. Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355, 2016. doi:10.1109/TIT.2016.2549542.

[13] V Sachdeva, T Mora, AM Walczak, and S Palmer. Optimal prediction with resource constraints using the information bottleneck. *bioRxiv*, 2020. doi:10.1101/2020.04.29.069179.

[14] S Still. Thermodynamic cost and benefit of memory. *Physical Review Letters*, 124:050601, 2020. doi:10.1103/PhysRevLett.124.050601.

[15] S Still, DA Sivak, AJ Bell, and GE Crooks. Thermodynamics of prediction. *Physical Review Letters*, 109: 120604, 2012. doi:10.1103/PhysRevLett.109.120604.

[16] DJ Strouse and DJ Schwab. The information bottleneck and geometric clustering. *Neural Computation*, 31 (3):596, 2019. doi:10.1162/neco_a_01136.

[17] N Tishby, FCN Pereira, and W Bialek. The information bottleneck method. In *37th Allerton Conference on Communication, Control and Computing*, 1999. `http://arxiv.org/abs/physics/0004057`.

[18] T Wu and I Fischer. Phase transitions for the information bottleneck in representation learning. In *International Conference on Learning Representations*, 2020. `https://openreview.net/forum?id=HJloElBYvB`.

[19] T Wu, I Fischer, IL Chuang, and M Tegmark. Learnability for the information bottleneck. *Entropy*, 21(10): 924, 2019. doi:10.3390/e21100924.

[20] K Xiao, L Engstrom, A Ilyas, and A Madry. Noise or signal: The role of image backgrounds in object recognition. 2020. `https://arxiv.org/abs/2006.09994`.