The Convex Information Bottleneck Lagrangian

Anonymous authors
Paper under double-blind review

Abstract

The information bottleneck (IB) problem tackles the issue of obtaining relevant compressed representations T of some random variable X for the task of predicting Y. It is defined as a constrained optimization problem which maximizes the information the representation has about the task, $I(T;Y)$, while ensuring that a minimum level of compression r is achieved (i.e., $I(X;T) \leq r$). For practical reasons the problem is usually solved by maximizing the IB Lagrangian (i.e., $\mathcal{L}_{\text{IB}}^{\beta}(T) = I(T;Y) - \beta I(X;T)$) for many values of $\beta \in [0,1]$, therefore drawing the IB curve (i.e., the curve of maximal $I(T;Y)$ for a given $I(X;Y)$) and selecting the representation of desired predictability and compression. It is known when Y is a deterministic function of X, the IB curve cannot be explored and other Lagrangians have been proposed to tackle this problem (e.g., the squared IB Lagrangian: $\mathcal{L}_{\text{sq-IB}}^{\beta_{\text{sq}}}(T) = I(T;Y) - \beta_{\text{sq}} I(X;T)^2$)). In this paper we (i) present a general family of Lagrangians which allow for the exploration of the IB curve in all scenarios; (ii) prove that if these Lagrangians are used, there is a one-to-one mapping between the Lagrange multiplier and the desired compression rate r for known IB curve shapes, hence, freeing from the burden of solving the optimization problem for many values of the Lagrange multiplier.

1 Introduction

Let X and Y be two statistically dependent random variables with joint distribution $p(x,y)$. The information bottleneck (IB) (Tishby et al., 2000) investigates the problem of extracting the relevant information from X for the task of predicting Y.

For this purpose, the IB defines a bottleneck variable T obeying the Markov chain $Y \leftrightarrow X \leftrightarrow T$ so that T acts as a representation of X. Tishby et al. (2000) define the relevant information as the information the representation keeps from Y after the compression of X (i.e., $I(T;Y)$), provided a minimum level of compression (i.e, $I(X;T) \leq r$). Therefore, we select the representation which yields the value of the IB curve that best fits our requirements.

Definition 1 (IB functional). *Let X and Y be statistically dependent variables. Let Δ be the set of random variables T obeying the Markov condition $Y \leftrightarrow X \leftrightarrow T$. Then the IB functional is*

$$F_{\text{IB,max}}(r) = \max_{T \in \Delta} \{I(T;Y)\} \textit{ s.t. } I(X;T) \leq r, \; \forall r \in [0,\infty). \tag{1}$$

Definition 2 (IB curve). *The IB curve is the set of points defined by the solutions of $F_{\text{IB,max}}(r)$ for varying values of $r \in [0,\infty)$.*

Definition 3 (Information plane). *The plane is defined by the axes $I(T;Y)$ and $I(X;T)$.*

In practice, solving a constrained optimization problem such as the IB functional is difficult. Thus, in order to avoid the non-linear constraints from the IB functional the IB Lagrangian is defined.

Definition 4 (IB Lagrangian). *Let X and Y be statistically dependent variables. Let Δ be the set of random variables T obeying the Markov condition $Y \leftrightarrow X \leftrightarrow T$. Then we define the IB Lagrangian as*

$$\mathcal{L}_{\textit{IB}}^{\beta}(T) = I(T;Y) - \beta I(X;T). \tag{2}$$

Here $\beta \in [0,1]$ is the Lagrange multiplier which controls the trade-off between the information of Y retained and the compression of X. Note we consider $\beta \in [0,1]$ because (i) for $\beta \leq 0$ many uncompressed solutions such as $T = X$ maximizes $\mathcal{L}_{\text{IB}}^{\beta}$, and (ii) for $\beta \geq 1$ the IB Lagrangian is non-positive due to the data processing inequality (DPI) (Theorem 2.8.1 from Cover & Thomas (2012)) and trivial solutions like $T = \text{const}$ are maximizers with $\mathcal{L}_{\text{IB}}^{\beta} = 0$ (Kolchinsky et al., 2019).

We know the solutions of the IB Lagrangian optimization (if existent) are solutions of the IB functional by the Lagrange's sufficiency theorem (Theorem 5 in Appendix A of Courcoubetis (2003)). Moreover, since the IB functional is concave (Lemma 5 of Gilad-Bachrach et al. (2003)) we know they exist (Theorem 6 in Appendix A of Courcoubetis (2003)).

Therefore, the problem is usually solved by maximizing the IB Lagrangian with adaptations of the Blahut-Arimoto algorithm (Tishby et al., 2000), deterministic annealing approaches (Tishby & Slonim, 2001) or a bottom-up greedy agglomerative clustering (Slonim & Tishby, 2000) or its improved sequential counterpart (Slonim et al., 2002). However, when provided with high-dimensional random variables X such as images, these algorithms do not scale well and deep learning based techniques, where the IB Lagrangian is used as the objective function, prevailed (Alemi et al., 2017; Chalk et al., 2016; Kolchinsky et al., 2017).

Note the IB Lagrangian optimization yields a representation T with a given performance $(I(X;T), I(T;Y))$ for a given β. However there is no one-to-one mapping between β and $I(X;T)$. Hence, we cannot directly optimize for a desired compression level r but we need to perform several optimizations for different values of β and select the representation with the desired performance (e.g., Alemi et al. (2017)). The Lagrange multiplier selection is important since (i) sometimes even choices of $\beta < 1$ lead to trivial representations such that $p_{T|X}(t|x) = p_T(t)$, and (ii) there exist some discontinuities on the performance level w.r.t. the values of β (Wu et al., 2019).

Moreover, recently Kolchinsky et al. (2019) showed how in deterministic scenarios (such as many classification problems where an input x_i belongs to a single particluar class y_i) the IB Lagrangian could not explore the IB curve. Particularly, they showed that multiple β yielded the same performance level and that a single value of β could result in different performance levels. To solve this issue, they introduced the squared IB Lagrangian, $\mathcal{L}_{\text{sq-IB}}^{\beta_{\text{sq}}} = I(T;Y) - \beta_{sq} I(X;T)^2$, which is able to explore the IB curve in any scenario by optimizing for different values of β_{sq}. However, even though they realized a one-to-one mapping between $\beta_s q$ existed, they did not find such mapping. Hence, multiple optimizations of the Lagrangian were still required to fing the best traded-off solution.

The main contributions of this article are:

1. We introduce a general family of Lagrangians (the convex IB Lagrangians) which are able to explore the IB curve in any scenario for which the squared IB Lagrangian (Kolchinsky et al., 2019) is a particular case of. More importantly, the analysis made for deriving this family of Lagrangians can serve as inspiration for obtaining new Lagrangian families which solve other objective functions with intrinsic trade-off such as the IB Lagrangian.
2. We show that in deterministic scenarios (and other scenarios where the IB curve shape is known) **one can use the convex IB Lagrangian to obtain a desired level of performance with a single optimization**. That is, there is a one-to-one mapping between the Lagrange multiplier used for the optmization and the level of compression and informativeness obtained, and we know such mapping. Therefore, eliminating the need of multiple optimizations to select a suitable representation.

Furthermore, we provide some insight for explaining why there are discontinuities in the performance levels w.r.t. the values of the Lagrange multipliers. In a classification setting, we connect those discontinuities with the intrinsic clusterization of the representations when optimizing the IB bottleneck objective.

The structure of the article is the following: in Section 2 we motivate the usage of the IB in supervised learning settings. Then, in Section 3 we outline the important results used about the IB curve in deterministic scenarios. Later, in Section 4 we introduce the convex IB Lagrangian and explain some of its properties. After that, we support our (proved) claims with some empirical evidence on the MNIST dataset (LeCun et al., 1998) in Section 5. The reader can download the PyTorch (Paszke et al., 2017) implementation at `https://gofile.io/?c=G9Dl1L`.

2 The IB in supervised learning

In this section we will first give an overview of supervised learning in order to later motivate the usage of the information bottleneck in this setting.

2.1 Supervised learning overview

In supervised learning we are given a dataset $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$ of n pairs of input features and task outputs. In this case, X and Y are the random variables of the input features and the task outputs. We assume x_i and y_i are sampled i.i.d. from the true distribution $p_{XY}(x,y) = p_{Y|X}(y|x)p_X(x)$. The usual aim of supervised learning is to use the dataset $\mathcal{D}_n$ to learn a particular conditional distribution $q_{\hat{Y}|X,\theta}(\hat{y}|x)$ of the task outputs given the input features, parametrized by θ, which is a good approximation of $p_{Y|X}(y|x)$. We use $\hat{Y}$ and $\hat{y}$ to indicate the predicted task output random variable and its outcome. We call a supervised learning task *regression* when Y is continuous-valued and *classification* when it is discrete.

Usually supervised learning methods employ intermediate representations of the inputs before making predictions about the outputs; e.g., hidden layers in neural networks (Chapter 5 from Bishop (2006)) or transformations in a feature space through the kernel trick in kernel machines like SVMs or RVMs (Sections 7.1 and 7.2 from Bishop (2006)). Let T be a possibly stochastic function of the input features X with a parametrized conditional distribution $q_{T|X,\theta}(t|x)$, then, T obeys the Markov condition $Y \leftrightarrow X \leftrightarrow T$. The mapping from the representation to the predicted task outputs is defined by the parametrized conditional distribution $q_{\hat{Y}|T,\theta}(\hat{y}|t)$. Therefore, in representation-based machine learning methods the full Markov Chain is $Y \leftrightarrow X \leftrightarrow T \leftrightarrow \hat{Y}$. Hence, the overall estimation of the conditional probability $p_{Y|X}(y|x)$ is given by the marginalization of the representations,

$$q_{\hat{Y}|X,\theta}(\hat{y}|x) = \int_{\forall t} q_{\hat{Y}|T,\theta}(\hat{y}|t) q_{T|X,\theta}(t|x) dt. \tag{3}$$

In order to achieve the goal of having a good estimation of the conditional probability distribution $p_{Y|X}(y|x)$, we usually define an *instantaneous cost function* $\jmath_\theta(x,y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. This serves as a heuristic to measure the loss our algorithm (parametrized by θ) obtains when trying to predict the realization of the task output y with the input realization x.

Clearly, we are interested in minimizing the expectation of the instantaneous cost function over all the possible input features and task outputs, which we call the *cost function*. However, since we only have a finite dataset $\mathcal{D}_n$ we have instead to minimize the *empirical cost function*.

Definition 5 (Cost function and empirical cost function). *Let X and Y be the input features and task output random variables and $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ their realizations. Let also $\jmath_\theta(x,y)$ be the instantaneous cost function, θ the parametrization of our learning algorithm, and $\mathcal{D}_n = \{x_i, y_i\}_{i=1}^n$ the given dataset. Then we define:*

1. The cost function:
$$J(\theta) = \mathbb{E}_{p_{XY}}[\jmath_\theta(x,y)] \tag{4}$$

2. The emprical cost function:
$$\hat{J}(\theta, \mathcal{D}_n) = \frac{1}{n}\sum_{i=1}^n \jmath_\theta(x_i, y_i) \tag{5}$$

The discrepancy between the normal and empirical cost functions is called the *generalization gap* or *generalization error* (see Section 1 of Xu & Raginsky (2017), for instance) and intuitevely, the smaller this gap is, the better our model generalizes (i.e., the better it will perform to new, unseen samples in terms of our cost function).

Definition 6 (Generalization gap). *Let $J(\theta)$ and $\hat{J}(\theta, \mathcal{D}_n)$ be the cost and the empirical cost functions as defined in Definition 5. Then, the generalization gap is defined as*

$$\text{gen}(\theta, \mathcal{D}_n) = J(\theta) - \hat{J}(\theta, \mathcal{D}_n), \tag{6}$$

and it represents the error incurred when the selected distribution is the one parametrized by θ when the rule $\hat{J}(\theta, \mathcal{D}_n)$ is used instead of $J(\theta)$ as the function to minimize.

Ideally, we would want to minimize the cost function. Hence, we usually try to minimize the empirical cost function and the generalization gap simultaneously. The modifications to our learning algorithm which intend to reduce the generalization gap but not hurt the performance on the empirical cost function are known as *regularization*.

2.2 WHY DO WE USE THE IB?

Definition 7 (Representation cross-entropy cost function). *Let X and Y be two statistically dependent variables with joint distribution $p_{XY}(x,y) = p_{Y|X}(y|x)p_X(x)$. Let also T be a random variable obeying the Markov condition $Y \leftrightarrow X \leftrightarrow T$ and $q_{T|X,\theta}(t|x)$ and $q_{\hat{Y}|T,\theta}(\hat{y}|t)$ be the encoding and decoding distributions of our model, parametrized by θ. Finally, let $\mathbb{C}(p(z)||q(z)) = -\mathbb{E}_{p(Z)}[\log(q(z))]$ be the cross entropy between two probability distributions p and q. Then, the cross-entropy cost function is*

$$J_{\text{CE}}(\theta) = \mathbb{E}_{q_{T|X,\theta}p_X}\left[\mathbb{C}(q_{Y|T,\theta}(y|t)||q_{\hat{Y}|T,\theta}(\hat{y}|t))\right] = \mathbb{E}_{p_{XY}}\left[\jmath_{\text{CE},\theta}(x,y)\right], \tag{7}$$

where $\jmath_{\text{CE},\theta}(x,y) = \mathbb{C}(q_{T|X,\theta}(t|x)||q_{\hat{Y}|T,\theta}(\hat{y}|t))$ is the instantaneous representation cross-entropy cost function and $q_{Y|T,\theta}(y|t) = \int_{\forall x} p_{Y|X}(y|x)q_{T|X,\theta}(t|x)p_X(x)/q_{T,\theta}(t)dx$.

The cross-entropy is a widely used cost function in classification tasks (e.g., Krizhevsky et al. (2012); Shore & Gray (1982); Teahan (2000)) which has many interesting properties (Shore & Johnson, 1981). Moreover, it is known that minimizing the $J_{\text{CE}}(\theta)$ maximizes the mutual information $I(T;Y)$ (see Section 2 of Kolchinsky et al. (2019) or Section II A. of Vera et al. (2018)).

Definition 8 (Nuisance). *A nuisance is any random variable which affects the observed data X but is not informative to the task we are trying to solve. That is, Ξ is a nuisance for Y if $Y \perp \Xi$ or $I(\Xi, Y) = 0$.*

Similarly, we know that minimizing $I(X;T)$ minimizes the generalization gap for restricted classes when using the cross-entropy cost function (Theorem 1 of Vera et al. (2018)), and when using $I(T;Y)$ directly as an objective to maximize (Theorem 4 of Shamir et al. (2010)). Furthermore, Achille & Soatto (2018) in Proposition 3.1 upper bound the information of the input representations, T, with nuisances that affect the observed data, Ξ, with $I(X;T)$. Therefore minimizing $I(X;T)$ helps generalization by not keeping useless information of Ξ in our representations.

Thus, jointly maximizing $I(T;Y)$ and minimizing $I(X;T)$ is a good choice both in terms of performance in the available dataset and in new, unseen data, which motivates studies on the IB.

3 THE INFORMATION BOTTLENECK IN DETERMINISTIC SCENARIOS

Kolchinsky et al. (2019) showed that when Y is a deterministic function of X (i.e., $Y = f(X)$), the IB curve is piecewise linear. More precisely, it is shaped as stated in Proposition 1.

Proposition 1 (The IB curve is piecewise linear in deterministic scenarios). *Let X be a random variable and $Y = f(X)$ be a deterministic function of X. Let also T be the bottleneck variable that solves the IB functional. Then the IB curve in the information plane is defined by the following equation:*

$$\begin{cases} I(T;Y) = I(X;T) & \text{if} \quad I(X;T) \in [0, I(X;Y)] \\ I(T;Y) = H(Y) & \text{if} \quad I(X;T) > I(X;Y) \end{cases} \tag{8}$$

Furthermore, they showed that the IB curve could not be explored by optimizing the IB Lagrangian for multiple β because the curve was not strictly concave. That is, there was not a one-to-one relationship between β and the performance level.

Theorem 1 (In deterministic scenarios, the IB curve cannot be explored using the IB Lagrangian). *Let X be a random variable and $Y = f(X)$ be a deterministic function of X. Let also T be the bottleneck variable that solves $\arg\max_{T\in\Delta}\{\mathcal{L}_{\text{IB}}^{\beta}\}$ with Δ the set of r.v. obeying the Markov condition $Y \leftrightarrow X \leftrightarrow T$. Then:*

1. *Any solution $T \in \Delta$ s.t. $I(X;T) \in [0, I(X;Y))$ and $I(T;Y) = I(X;T)$ solves $\arg\max_{T\in\Delta}\{\mathcal{L}_{\text{IB}}^{\beta}\}$ for $\beta = 1$.*
2. *Any solution $T \in \Delta$ s.t. $I(X;T) > I(X;Y)$ and $I(T;Y) = I(X;Y)$ solves $\arg\max_{T\in\Delta}\{\mathcal{L}_{\text{IB}}^{\beta}\}$ for $\beta = 0$.*
3. *The solution of $I(X;T) = I(T;Y) = I(X;Y)$ is achieved $\forall\beta \in (0,1)$. Furthermore, this is the only solution $\beta \in (0,1)$ yields.*

4 The Convex IB Lagrangian

4.1 Exploring the IB curve

Clearly, a situation like the one depicted in Theorem 1 is not desirable, since we cannot aim for different levels of compression or performance. For this reason, we generalize the effort from Kolchinsky et al. (2019) and look for families of Lagrangians which are able to explore the IB curve. Inspired by the squared IB Lagrangian, $\mathcal{L}_{\text{sq-IB}}^{\beta_{\text{sq}}}(T) = I(T;Y) - \beta_{\text{sq}} I(X;T)^2$, we look at the conditions a function of $I(X;T)$ requires in order to be able to explore the IB curve. In this way, we realize that any monotonically increasing and strictly convex function will be able to do so, and we call the family of Lagrangians with these characteristics the *convex IB Lagrangians*, due to the nature of the introduced function.

Theorem 2 (Convex IB Lagrangians). *Let Δ be the set of r.v. T obeying the Markov condition $Y \leftrightarrow X \leftrightarrow T$. Then, if h is a* ***monotonically increasing and strictly convex function****, the IB curve can always be recovered by the solutions of $\arg\max_{T\in\Delta}\{\mathcal{L}_{\text{IB},h}^{\beta_h}(T)\}$, with*

$$\mathcal{L}_{\text{IB,h}}^{\beta_h}(T) = I(T;Y) - \beta_h h(I(X;T)). \tag{9}$$

That is, for each point $(I(X;T), I(T;Y))$ s.t. $dI(T;Y)/dI(X;T) > 0$ there is a unique β_h for which maximizing $\mathcal{L}_{\text{IB,h}}^{\beta_h}(T)$ achieves this solution. Furthermore, β_h is strictly decreasing w.r.t. $I(X;T)$. We call $\mathcal{L}_{\text{IB},h}^{\beta_h}(T)$ the convex IB Lagrangian.

The proof of this theorem can be found on Appendix A. Furthermore, by exploiting the IB curve duality (Lemma 10 of Gilad-Bachrach et al. (2003)) we were able to derive other families of Lagrangians which allow for the exploration of the IB curve (Appendix E).

Remark 1. *Clearly, we can see how if h is the identity function (i.e., $h(I(X;T)) = I(X;T)$) then we end up with the normal IB Lagrangian. However, since the identity function is not strictly convex, it cannot ensure the exploration of the IB curve.*

4.2 Aiming for a specific compression level

Let B_h denote the domain of Lagrange multipliers β_h for which we can find solutions in the IB curve with the convex IB Lagrangian. Then the convex IB Lagrangians do not only allow us to explore the IB curve with different β_h. They also allow us to identify the specific β_h that obtains a given point $(I(X;T), I(T;Y))$, provided we know the IB curve in the information plane. Conversely, the convex IB Lagrangian allows to find the specific point $(I(X;T), I(T;Y))$ that is obtained by a given β_h.

Proposition 2 (Bijective mapping between IB curve point and convex IB Lagrange multiplier). *Let the IB curve in the information plane be known; i.e., $I(T;Y) = f_{\text{IB}}(I(X;T))$ is known. Then there is a bijective mapping from Lagrange multipliers $\beta_h \in B_h \setminus \{0\}$ from the convex IB Lagrangian to points in the IB curve $(I(X;T), f_{\text{IB}}(I(X;T))$. Furthermore, these mappings are:*

$$\beta_h = \frac{df_{\text{IB}}(I(X;T))}{dI(X;T)} \frac{1}{h'(I(X;T))} \quad \textit{and} \quad I(X;T) = (h')^{-1}\left(\frac{df_{\text{IB}}(I(X;T))}{dI(X;T)} \frac{1}{\beta_h}\right), \tag{10}$$

where h' is the derivative of h and $(h')^{-1}$ is the inverse of h'.

It is interesting since in deterministic scenarios we know the shape of the IB curve (Theorem 1) and since the convex IB Lagrangians allow for the exploration of the IB curve (Theorem 2). A proof for Proposition 2 can be found in Appendix B.

Remark 2. *The inclusion of the function h is what allows us to find the bijection between β_h and $I(X;T)$. The previous definition from Tishby et al. (2000) of β as $d(I(T;Y))/dI(X;T)$ did not.*

A direct result derived from this proposition is that we know the domain of Lagrange multipliers, B_h, which allow for the exploration of the IB curve if the shape of the IB curve is known. Furthermore, if the shape is not known we can at least bound that range.

Corollary 1 (Domain of convex IB Lagrange multiplier with known IB curve shape). *Let the IB curve in the information plane be $I(T;Y) = f_{\text{IB}}(I(X;T))$ and let $I_{\max} = I(X;Y)$. Let also $I(X;T) = r_{\max}$ be the minimum mutual information s.t. $f_{\text{IB}}(r_{\max}) = I_{\max}$ (i.e., $r_{\max} = \min_r\{f_{\text{IB}}(r) = I_{\max}\}$). Then, the range of Lagrange multipliers that allow the exploration of the IB curve with the convex IB Lagrangian is $B_h = [\beta_{h,\min}, \beta_{h,\max}]$, with*

$$\beta_{h,\min} = \lim_{r \to r_{\max}^-} \left\{ \frac{f'_{\text{IB}}(r)}{h'(r)} \right\} \geq 0 \quad \textit{and} \quad \beta_{h,\max} = \lim_{r \to 0^+} \left\{ \frac{f'_{\text{IB}}(r)}{h'(r)} \right\}, \tag{11}$$

where $f'_{\text{IB}}(r)$ and $h'(r)$ are the derivatives of $f_{\text{IB}}(I(X;T))$ and $h(I(X;T))$ w.r.t. $I(X;T)$ evaluated at r respectively.

Corollary 2 (Domain of convex IB Lagrange multiplier bound). *The range of the Lagrange multipliers that allow the exploration of the IB curve is contained by $[0, \beta_{\text{h,top}}]$ which is also contained by $[0, \beta^+_{h,\text{top}}]$, where*

$$\beta_{h,\text{top}} = \frac{(\inf_{\Omega_x \subset \mathcal{X}}\{\beta_0(\Omega_x)\})^{-1}}{\lim_{r \to 0^+}\{h'(r)\}}, \textit{ and } \beta^+_{h,\text{top}} = \frac{1}{\lim_{r \to 0^+}\{h'(r)\}}, \tag{12}$$

$h'(r)$ is the derivative of $h(I(X;T))$ w.r.t. $I(X;T)$ evaluated at r, $\mathcal{X}$ is the set of possible realizations of X and β_0[1] and Ω_x are defined as in (Wu et al., 2019). That is, $B_h \subseteq [0, \beta_{\text{h,top}}] \subseteq [0, \beta^+_{\text{h,top}}]$.

Corollaries 1 and 2 allow us to reduce the range search for β when we want to explore the IB curve. Practically, $\inf_{\Omega_x \subset \mathcal{X}}\{\beta_0(\Omega_x)\}$ might be difficult to calculate so Wu et al. (2019) derived an algorithm to approximate it. However, we still recommend 1 for simplicity. The proofs for both corollaries are found in Appendices C and D.

5 Experimental support

In order to showcase our claims we use the MNIST dataset (LeCun et al., 1998). We simply modify the nonlinear-IB method (Kolchinsky et al., 2017), which is a neural network that minimizes the cross-entropy while also minimizing a differentiable kernel-based estimate of $I(X;T)$ (Kolchinsky & Tracey, 2017). Then we use this technique to maximize a lower bound on the convex IB Lagrangians by applying the functions h to the $I(X;T)$ estimate.

For a fair comparison, we use the same network architecture as that in (Kolchinsky et al., 2017): First, a stochastic encoder [2] $T = f_{\theta,\text{enc}}(X) + W$ with $W \sim \mathcal{N}(0, I_2)$ such that $T \in \mathbb{R}^2$. Here $f_{\theta,\text{enc}}$ is a three fully-conected layer encoder with 800 ReLU units on the first two layers and 2 linear units on the last layer. Second, a deterministic decoder $q_{\hat{Y}|T,\theta}(\hat{y}|t) = f_{\theta,\text{dec}}(t)$. Here, $f_{\theta,\text{dec}}$ is a fully-conected 800 ReLU unit layers followed by an output layer with 10 softmax units. For further details about the experiment setup and additional results for different values of α and η please refer to Appendix F.

In Figure 1 we show our results for two particularizations of the convex IB Lagrangians:

[1] Note in (Wu et al., 2019) they consider the dual problem (see Appendix E) so when they refer to β^{-1} it translates to β in this article.

[2] The encoder needs to be stochastic to (i) ensure a finite and well-defined mutual information (Kolchinsky et al., 2019; Amjad & Geiger, 2019) and (ii) make gradient-based optimization methods over the IB Lagrangian useful (Amjad & Geiger, 2019).

[3] The clusters were obtained using the DBSCAN algorithm (Ester et al., 1996; Schubert et al., 2017).

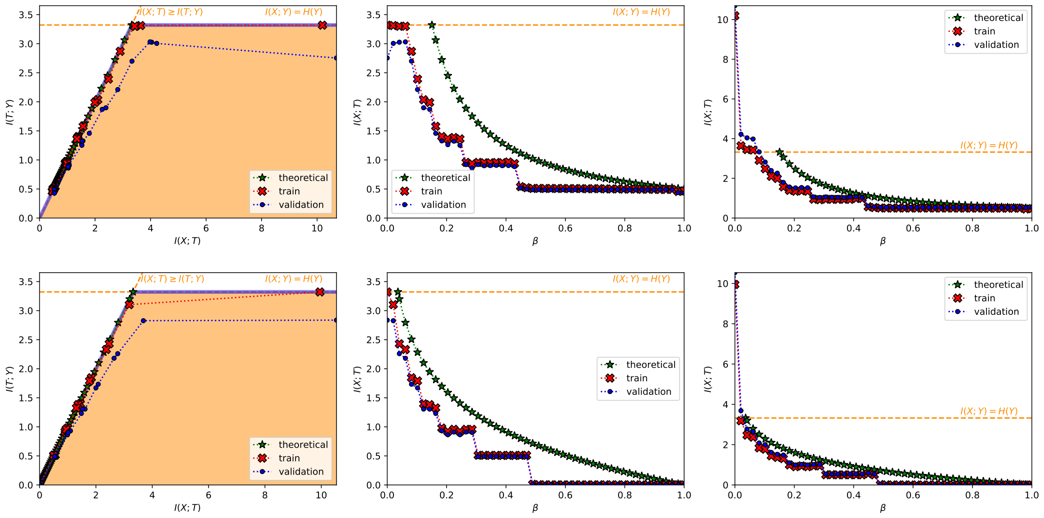
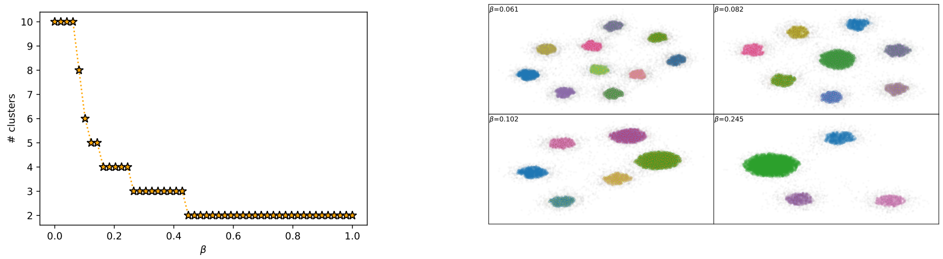


Figure 1: The top row shows the results for the power IB Lagrangian with $\alpha = 1$, and the bottom row for the exponential IB Lagrangian with $\eta = 1$. In each row, from left to right it is shown (i) the information plane, where the region of possible solutions of the IB problem is shadowed in light orange and the information-theoretic limits are the dashed orange line; (ii) $I(T;Y)$ as a function of β_h; and (iii) the compression $I(X;T)$ as a function of β_h. In all plots the red crosses joined by a dotted line represent the values computed with the training set, the blue dots the values computed with the validation set and the green stars the theoretical values computed as dictated by Proposition2. Moreover, in all plots it is indicated $I(X;Y) = H(Y) = \log_2(10)$ in a dashed, orange line. All values are shown in bits.



(a) Number of clusters for different β_{pow}.

(b) Example of clusters for different β_{pow}.

Figure 2: Depiction of the clusterization behavior[3]of the bottleneck variable for the power IB Lagrangian with $\alpha = 1$.

1. the power IB Lagrangians[4]: $\mathcal{L}_{\text{IB,pow}}^{\beta_{\text{pow}}}(T,\alpha) = I(T;Y) - \beta_{\text{pow}} I(X;T)^{(1+\alpha)}$, $\alpha > 0$.
2. the exponential IB Lagrangians: $\mathcal{L}_{\text{IB,exp}}^{\beta_{\text{exp}}}(T,\eta) = I(T;Y) - \beta_{\text{exp}} \exp(\eta I(X;T))$, $\eta > 0$.

We can clearly see how both Lagrangians are able to explore the IB curve (first column from Figure 1) and how the theoretical performance trend of the Lagrangians matches the experimental results (second and third columns from Figure 1). There are small mismatches between the theoretical and experimental performance. This is because using the nonlinear-IB, as stated by Kolchinsky et al. (2019), does not guarantee that we find optimal representations due to factors like: (i) innacurate estimation of $I(X;T)$, (ii) restrictions on the structure of T, (iii) use of an estimation of the decoder instead of the real one and (iv) the typical non-convex optimization issues that arise with gradient-based methods. The main difference comes from the discontinuities in performance for in-

[4] Note when $\alpha = 1$ we have the squared IB functional from Kolchinsky et al. (2019).

creasing β, which cause is still unknown (cf. Wu et al. (2019)). It has been observed, however, that the bottleneck variable performs an intrinsic clusterization in classification tasks (see, for instance (Kolchinsky et al., 2017; 2019; Alemi et al., 2018) or Figure 2b). We realized how this clusterization matches with the quantized performance levels observed (e.g., compare Figure 2a with the top center graph in Figure 1); with maximum performance when the number of clusters is equal to the cardinality of Y and reducing performance with a reduction of the number of clusters. We do not have a mathematical proof for the exact relationship between these two phenomena; however, we agree with Wu et al. (2019) that it is an interesting matter and hope this realization serves as motivation to derive new theory.

To sum up, in order to achieve a desired level of performance with the convex IB Lagrangian as an objective one should:

1. In a deterministic or close to deterministic setting (see ϵ-deterministic definition in Kolchinsky et al. (2019)): Use the adequate β_h for that performance using Proposition 2. Then if the perfomance is lower than desired (i.e., we are placed in the wrong performance plateau), gradually reduce the value of β_h until reaching the previous performance plateau.
2. In a stochastic setting: Draw the IB curve with multiple values of β_h on the range defined by Corollary 2 and select the representations that best fit their interests.

In practice, there are different criterions for choosing the function h. For instance, the exponential IB Lagrangian could be more desirable than the power IB Lagrangian when we want to draw the IB curve since it has a finite range of β_h. This is $B_h = [(\eta \exp(\eta H_{\text{max}}))^{-1}, \eta^{-1}]$ for the exponential IB Lagrangian vs. $B_h = [((1+\alpha) H_{\text{max}}^{\alpha})^{-1}, \infty)$ for the power IB Lagrangian. Furthermore, there is a trade-off between (i) how much the selected h function ressembles the identity (e.g., with α or η close to zero), since it will suffer from similar problems as the original IB Lagrangian; and (ii) how fast it grows (e.g., higher values of α or η), since it will suffer from value convergence; i.e., optimizing for separate values of β_h will achieve similar levels of performance (Figure 3). Please, refer to Appendix G for a more thorough explanation of this phenomenon.

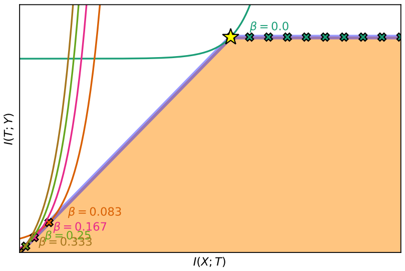


Figure 3: Example of value convergence with the exponential IB Lagrangian with $\eta = 3$. We show the intersection of the isolines of $\mathcal{L}_{\text{IB,exp}}^{\beta_{\text{exp}}}$ for different $\beta_{\text{exp}} \in B_{\text{exp}}$[5] with the IB curve.

6 Conclusion

The information bottleneck is a widely used and studied technique. However, it is known that the IB Lagrangian cannot be used to achieve varying levels of performance in deterministic scenarios. Moreover, in order to achieve a particular level of performance multiple optimizations with different Lagrange multipliers must be done to draw the IB curve and select the best traded-off representation.

In this article we introduced a general family of Lagrangians which allow to (i) achieve varying levels of performance in any scenario, and (ii) **pinpoint a specific Lagrange multiplier β_h to optimize for a specific performance level in known IB curve scenarios** (e.g., deterministic). Furthermore, we showed the β_h domain when the IB curve is known and a β_h domain bound for exploring the IB curve when it is unkown. This way we can reduce and/or avoid multiple optimizations and, hence, reduce the computational effort for finding well traded-off representations. Finally, (iii) we provided some insight to the discontinuities on the performance levels w.r.t. the Lagange multipliers by connecting those with the intrinsic clusterization of the bottleneck variable.

[5] $B_h \approx [1.56 \cdot 10^{-5}, 3^{-1}]$ using Corollary 1.

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. 2017.

Alexander A Alemi, Ian Fischer, and Joshua V Dillon. Uncertainty in the variational information bottleneck. *arXiv preprint arXiv:1807.00906*, 2018.

Rana Ali Amjad and Bernhard Claus Geiger. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Matthew Chalk, Olivier Marre, and Gasper Tkacik. Relevant sparse codes with variational information bottleneck. In *Advances in Neural Information Processing Systems*, pp. 1957–1965, 2016.

Costas Courcoubetis. *Pricing Communication Networks Economics, Technology and Modelling*. Wiley Online Library, 2003.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pp. 226–231, 1996.

Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby. An information theoretic tradeoff between complexity and accuracy. In *Learning Theory and Kernel Machines*, pp. 595–609. Springer, 2003.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Artemy Kolchinsky and Brendan Tracey. Estimating mixture entropy with pairwise distances. *Entropy*, 19(7):361, 2017.

Artemy Kolchinsky, Brendan D Tracey, and David H Wolpert. Nonlinear information bottleneck. *arXiv preprint arXiv:1705.02436*, 2017.

Artemy Kolchinsky, Brendan D Tracey, and Steven Van Kuyk. Caveats for information bottleneck in deterministic scenarios. In *ICLR*, 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):19, 2017.

Ohad Shamir, Sivan Sabato, and Naftali Tishby. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29-30):2696–2711, 2010.

John Shore and Rodney Johnson. Properties of cross-entropy minimization. *IEEE Transactions on Information Theory*, 27(4):472–482, 1981.

John E Shore and Robert M Gray. Minimum cross-entropy pattern classification and cluster analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (1):11–17, 1982.

Noam Slonim and Naftali Tishby. Agglomerative information bottleneck. In *Advances in neural information processing systems*, pp. 617–623, 2000.

Noam Slonim, Nir Friedman, and Naftali Tishby. Unsupervised document classification using sequential information maximization. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 129–136. ACM, 2002.

William John Teahan. Text classification and segmentation using minimum cross-entropy. In *Content-Based Multimedia Information Access-Volume 2*, pp. 943–961. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2000.

Naftali Tishby and Noam Slonim. Data clustering by markovian relaxation and the information bottleneck method. In *Advances in neural information processing systems*, pp. 640–646, 2001.

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

Matias Vera, Pablo Piantanida, and Leonardo Rey Vega. The role of the information bottleneck in representation learning. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 1580–1584. IEEE, 2018.

Tailin Wu, Ian Fischer, Isaac Chuang, and Max Tegmark. Learnability for the information bottleneck. In *ICLR*, 2019.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pp. 2524–2533, 2017.

A Proof of Theorem 2

Proof. We start the proof by remembering the optimization problem at hand (Definition 1):

$$F_{\text{IB,max}}(r) = \max_{T \in \Delta}\{I(T;Y)\} \text{ s.t. } I(X;T) \leq r \tag{13}$$

We can modify the optimization problem by

$$\max_{T \in \Delta}\{I(T;Y)\} \text{ s.t. } h(I(X;T)) \leq h(r) \tag{14}$$

iff h is a monotonically non-decreasing function since otherwise $h(I(X;T)) \leq h(r)$ would not hold necessarily. Now, let us assume $\exists T^* \in \Delta$ and β_h^* s.t. T^* maximizes $\mathcal{L}_{\text{IB,h}}^{\beta_h^*}(T)$ over all $T \in \Delta$, and $I(X;T^*) \leq r$. Then, we can operate as follows:

$$\begin{aligned}
\max_{\substack{T \in \Delta \\ h(I(X;T)) \leq h(r)}}\{I(T;Y)\} &= \max_{\substack{T \in \Delta \\ h(I(X;T)) \leq h(r)}}\{I(T;Y) - \beta_h^*(h(I(X;T)) - h(r) + \xi)\} && (15) \\
&\leq \max_{T \in \Delta}\{I(T;Y) - \beta_h^*(h(I(X;T)) - h(r) + \xi)\} && (16) \\
&= I(T^*;Y) - \beta_h^*(h(I(X;T^*) - h(r) + \xi) = I(T^*;Y). && (17)
\end{aligned}$$

Here, the equality from equation (15) comes from the fact that since $I(X;T) \leq r$, then $\exists \xi \geq 0$ s.t. $h(I(X;T)) - h(r) + \xi = 0$. Then, the inequality from equation (16) holds since we have expanded the optimization search space. Finally, in equation (17) we use that T^* maximizes $\mathcal{L}_{\text{IB,h}}^{\beta_h^*}(T)$ and that $I(X;T^*) \leq r$.

Now, we can exploit that $h(r)$ and ξ do not depend on T and drop them in the maximization in equation (16). We can then realize we are maximizing over $\mathcal{L}_{\text{IB,h}}^{\beta_h^*}(T)$; i.e.,

$$\underset{\substack{T \in \Delta \\ h(I(X;T)) \leq h(r)}}{\arg\max} \{I(T;Y)\} \leq \underset{T \in \Delta}{\arg\max}\{I(T;Y) - \beta_h^*(h(I(X;T)) - h(r) + \xi)\} \tag{18}$$

$$= \underset{T \in \Delta}{\arg\max}\{I(T;Y) - \beta_h^* h(I(X;T))\} = \underset{T \in \Delta}{\arg\max}\{\mathcal{L}_{\text{IB,h}}^{\beta_h^*}(T)\}. \tag{19}$$

Therefore, since $I(T^*;Y)$ satisfies both the maximization with $T^* \in \Delta$ and the constraint $I(X;T^*) \leq r$, maximizing $\mathcal{L}_{\text{IB,h}}^{\beta_h^*}(T)$ obtains $F_{\text{IB,max}}(r)$.

Now, we know if such β_h^* exists, then the solution of the Lagrangian will be a solution for $F_{\text{IB,max}}(r)$. Then, if we consider Theorem 6 from the Appendix of Courcoubetis (2003) and consider the maximization problem instead of the minimization problem, we know if both $I(T;Y)$ and $-h(I(X;T))$ are concave functions, then a set of Lagrange multipliers S_h^* exists with these conditions. We can make this consideration because f is concave if $-f$ is convex and $\max\{f\} = \min\{-f\}$. We know $I(T;Y)$ is a concave function of T for $T \in \Delta$ (Lemma 5 of Gilad-Bachrach et al. (2003)) and $I(X;T)$ is convex w.r.t. T given $p_X(x)$ is fixed (Theorem 2.7.4 of Cover & Thomas (2012)). Thus, if we want $-h(I(X;T))$ to be concave **we need** h **to be a convex function**.

Finally, we will look at the conditions of h so that for every point $(I(X;T), I(T;Y))$ in the IB curve, there exists a unique β_h^* s.t. $\mathcal{L}_{\text{IB,h}}^{\beta_h^*}(T)$ is maximized. That is, the conditions of h s.t. $|S_h^*| = 1$. For this purpose we will look at the solutions of the Lagrangian optimization:

$$\frac{d\mathcal{L}_{\text{IB,h}}^{\beta_h}(T)}{dT} = \frac{d(I(T;Y) - \beta_h h(I(X;T)))}{dT} = \frac{dI(T;Y)}{dT} - \beta_h \frac{dh(I(X;T))}{dI(X;T)} \frac{dI(X;T)}{dT} = 0 \tag{20}$$

Now, if we integrate both sides of equation (20) over all $T \in \Delta$ we obtain

$$\beta_h = \frac{dI(T;Y)}{dI(X;T)} \left(\frac{dh(I(X;T))}{dI(X;T)} \right)^{-1} = \frac{\beta}{h'(I(X;T))}, \tag{21}$$

where β is the Lagrange multiplier from the IB Lagrangian (Tishby et al., 2000) and $h'(I(X;T))$ is $\frac{dh(I(X;T))}{dI(X;T)}$. Also, if we want to avoid indeterminations of β_h we need $h'(I(X;T))$ not to be 0. Since we already imposed h to be monotonically non-decreasing, we can solve this issue by strengthening this condition. That is, we will require h **to be monotonically increasing**.

We would like β_h to be continuous, this way there would be a unique β_h for each value of $I(X;T)$. We know β is a non-increasing function of $I(X;T)$ (Lemma 6 of Gilad-Bachrach et al. (2003)). Hence, if we want β_h **to be a strictly decreasing function of** $I(X;T)$, we will require h' to be an strictly increasing function of $I(X;T)$. Therefore, **we will require** h **to be a strictly convex function**.

Thus, if h is an strictly convex and monotonically increasing function, for each point $(I(X;T), I(T;Y))$ in the IB curve s.t. $dI(T;Y)/dI(X;T) > 0$ there is a unique β_h for which maximizing $\mathcal{L}_{\text{IB,h}}^{\beta_h}(T)$ achieves this solution.

□

B Proof of Proposition 2

Proof. In Theorem 2 we showed how each point of the IB curve $(I(X;T), I(T;Y))$ can be found with a unique β_h maximizing $\mathcal{L}_{\text{IB,h}}^{\beta_h}$. Therefore since we also proved $\mathcal{L}_{\text{IB,h}}^{\beta_h}$ is strictly concave w.r.t. T we can find the values of β_h that maximize the Lagrangian for fixed $I(X;T)$.

First, we look at the solutions of the Lagrangian maximization:

$$\frac{d\mathcal{L}_{\text{IB,h}}^{\beta_h}(T)}{dT} = \frac{d(f_{\text{IB}}(I(X;T)) - \beta_h h(I(X;T)))}{dT} = \frac{df_{\text{IB}}(I(X;T))}{dT} - \beta_h \frac{dh(I(X;T))}{dI(X;T)} \frac{dI(X;T)}{dT} = 0. \tag{22}$$

Then as before we can integrate at both sides for all $T \in \Delta$ and solve for β_h:

$$\beta_h = \frac{df_{\text{IB}}(I(X;T))}{dI(X;T)} \frac{1}{h'(I(X;T))}. \tag{23}$$

Moreover, since h is a strictly convex function its derivative h' is strictly decreasing. Hence, h' is an invertible function (since a strictly decreasing function is bijective and a function is invertible iff it is bijective by definition). Now, if we consider $\beta_h > 0$ to be known and $I(X;T)$ to be the unknown we can solve for $I(X;T)$ and get:

$$I(X;T) = (h')^{-1}\left(\frac{df_{\text{IB}}(I(X;T))}{dI(X;T)} \frac{1}{\beta_h}\right). \tag{24}$$

Note we require β_h not to be 0 so the mapping is defined. □

C Proof of Corollary 1

Proof.

Lemma 1. *Let $\mathcal{L}_{\text{IB,h}}^{\beta_h}(T)$ be a convex IB Lagrangian, then* $\max_{T\in\Delta}\{\mathcal{L}_{\text{IB,h}}^{0}(T)\} = I(X;Y)$.

Proof. If we write $\mathcal{L}_{\text{IB,h}}^{0}(T) = I(T;Y)$, we see that maximizing this Lagrangian is directly maximizing $I(T;Y)$. We know $I(T;Y)$ is a concave function of T for $T \in \Delta$ (Theorem 2.7.4 from Cover & Thomas (2012)); hence it has a maximum. We also know $I(T;Y) \leq I(X;Y)$. Moreover, we know $I(X;Y)$ can be achieved if, for example, Y is a deterministic function of T (since then the Markov Chain $X \leftrightarrow T \leftrightarrow Y$ is formed). Thus, $\max_{T\in\Delta}\{\mathcal{L}_{\text{IB,h}}^{0}(T)\} = I(X;Y)$. □

For $\beta_h = 0$ we know maximizing $\mathcal{L}_{\text{IB,h}}(T)$ can obtain the point in the IB curve $(r_{\max}, I_{\max})$ (Lemma 1).

Moreover, we know that for every point $(I(X;T), f_{\text{IB}}(I(X;T)))$, $\exists!\beta_h$ s.t. $\max\{\mathcal{L}_{\text{IB,h}}^{\beta_h}(T)\}$ achieves that point (Theorem 2). Thus, $\exists!\beta_{h,\min}$ s.t. $\lim_{r\to r_{\max}^-}(r, f_{\text{IB}}(r))$ is achieved. From Proposition 2 we know this $\beta_{h,\min}$ is given by

$$\beta_{h,\min} = \lim_{r\to r_{\max}^-}\left\{\frac{f'_{\text{IB}}(r)}{h'(r)}\right\}. \tag{25}$$

Since we know $f_{\text{IB}}(I(X;T))$ is a concave non-decreasing function in $(0, r_{\max})$ (Lemma 5 of Gilad-Bachrach et al. (2003)) we know it is continuous in this interval. In addition we know β_h is strictly decreasing w.r.t. $I(X;T)$ (Theorem 2). Furthermore, by definition of $r_{\max}$ and knowing $I(T;Y) \leq I(X;Y)$ we know $f'_{\text{IB}}(r) = 0$, $\forall r > r_{\max}$. Therefore, we cannot ensure the exploration of the IB curve for β'_h s.t. $0 < \beta'_h < \beta_{h,\min}$.

Then, since h is a strictly increasing function in $(0, r_{\text{max}})$, h' is positive in that interval. Hence, taking into account β_h is strictly decreasing we can find a maximum β_h when $I(X;T)$ approaches to 0. That is,

$$\beta_{h,\text{max}} = \lim_{r \to 0^+} \left\{ \frac{f'_{\text{IB}}(r)}{h'(r)} \right\}, \tag{26}$$

□

D Proof of Corollary 2

Proof. If we use Corollary 1, it is straightforward to see that $\beta_h \subseteq [L_-, L_+]$ if $\beta_{\text{h,min}} \geq L_-$ and $\beta_{\text{h,max}} \leq L_+$ for all IB curves f_{IB} and functions h. Therefore, we look at a domain bound dependent on the function choice. That is, if we can find $\beta_{\text{min}} \leq f'_{\text{IB}}(r)$ and $\beta_{\text{max}} \geq f'_{\text{IB}}(r)$ for all IB curves and all values of r, then

$$B_h \subseteq \left[\frac{\beta_{\text{min}}}{\lim_{r \to r^-_{\text{max}}} \{h'(r)\}}, \frac{\beta_{\text{max}}}{\lim_{r \to 0^+} \{h'(r)\}} \right]. \tag{27}$$

The region for all possible IB curves regardless of the relationship between X and Y is depicted in Figure 4. The hard limits are imposed by the DPI (Theorem 2.8.1 from Cover & Thomas (2012)) and the fact that the mutual information is non-negative (Corollary 2.90 for discrete and first Corollary of Theorem 8.6.1 for continuous random variables from Cover & Thomas (2012)). Hence, a minimum and maximum values of f'_{IB} are given by the minimum and maximum values of the slope of the Pareto frontier. Which means

$$B_h \subseteq \left[0, \frac{1}{\lim_{r \to 0^+} \{h'(r)\}} \right]. \tag{28}$$

Note $0/(\lim_{r \to r^-_{\text{max}}} \{h'(r)\}) = 0$ since h is monotonically increasing and, thus, h' will never be 0.

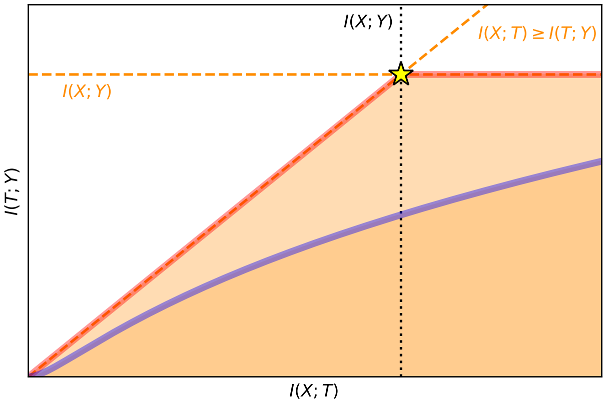


Figure 4: Graphical representation of the IB curve in the information plane. Dashed lines in orange represent tight bounds confining the region (in light orange) of possible IB curves (delimited by the red line, also known as the Pareto frontier). Black dotted lines are informative values. In blue we show an example of a possible IB curve confining a region (in darker orange) of an IB curve which does not achieve the Pareto frontier. Finally, the yellow star represents the point where the representation keeps the same information about the input and the output.

Finally, we can tighten the bound using the results from Wu et al. (2019), where, in Theorem 2, they showed the slope of the Pareto frontier could be bounded in the origin by $f'_{\text{IB}} \leq (\inf_{\Omega_x \subset \mathcal{X}}\{\beta_0(\Omega_x)\})^{-1}$. Finally, we know that in deterministic classification tasks $\inf_{\Omega_x \subset \mathcal{X}}\{\beta_0(\Omega_x)\} = 1$, which aligns with Kolchinsky et al. (2019) and what we can observe from Figure 4. Therefore,

$$B_h \subseteq \left[0, \frac{(\inf_{\Omega_x \subset \mathcal{X}}\{\beta_0(\Omega_x)\})^{-1}}{\lim_{r \to 0^+}\{h'(r)\}}\right] \subseteq \left[0, \frac{1}{\lim_{r \to 0^+}\{h'(r)\}}\right]. \tag{29}$$

□

E Other Lagrangian Families

We can use the same ideas we used for the convex IB Lagrangian to formulate new families of Lagrangians that allow the exploration of the IB curve. For that we will use the duality of the IB curve (Lemma 10 of (Gilad-Bachrach et al., 2003)). That is:

Definition 9 (IB dual functional). *Let X and Y be statistically dependent variables. Let also Δ be the set of random variables T obeying the Markov condition $Y \leftrightarrow X \leftrightarrow T$. Then the IB dual functional is*

$$F_{\text{IB,min}}(i) = \min_{T \in \Delta}\{I(X;T)\} \text{ s.t. } I(T;Y) \geq i, \ \forall i \in [0, I(X;Y)]. \tag{30}$$

Theorem 3 (IB curve duality). *Let the IB curve be defined by the solutions of $F_{\text{IB,max}}(r)$ for varying $r \in [0, \infty)$. Then,*

$$\forall r \exists i \textit{ s.t. } (r, F_{\text{IB,max}}(r)) = (F_{\text{IB,min}}(i), i) \tag{31}$$

and

$$\forall i \exists r \textit{ s.t. } (F_{\text{IB,min}}(i), i) = (r, F_{\text{IB,max}}(r)). \tag{32}$$

From this definition it follows that minimizing the *dual IB Lagrangian*, $\mathcal{L}^{\beta_{\text{dual}}}_{\text{IB,dual}}(T) = I(X;T) - \beta_{\text{dual}} I(T;Y)$, for $\beta_{\text{dual}} = \beta^{-1}$ is equivalent to maximizing the IB Lagrangian. In fact, the original Lagrangian for solving the problem was defined this way (Tishby et al., 2000). We decided to use the maximization version because the domain of useful β is bounded while it is not for β_{dual}.

Following the same reasoning as we did in the proof of Theorem 2, we can ensure the IB curve can be explored if:

1. We minimize $\mathcal{L}^{\beta_g}_{\text{IB,g}}(T) = I(X;T) - \beta_g g(I(T;Y))$.
2. We maximize $\mathcal{L}^{\beta_{g,\text{dual}}}_{\text{IB,g,dual}}(T) = g(I(T;Y)) - \beta_{\text{g,dual}} I(X;T)$.
3. We minimize $\mathcal{L}^{\beta_{h,\text{dual}}}_{\text{IB,h,dual}}(T) = h(I(X;T)) - \beta_{h,\text{dual}} I(T;Y)$.

Here, h is a monotonically increasing strictly convex function, g is a monotonically increasing strictly concave function, and $\beta_g, \beta_{g,\text{dual}}, \beta_{h,\text{dual}}$ are the Lagrange multipliers of the families of Lagrangians defined above.

In a similar manner, one could obtain relationships between the Lagrange multipliers of the IB Lagrangian and the convex IB Lagrangian with these Lagrangian families. Also, one could find a range of values for these Lagrangians to allow for the IB curve exploration and define a bijective mapping between their Lagrange multipliers and the IB curve. However, (i) as mentioned in Section 2.2, $I(T;Y)$ is particularly interesting to maximize without transformations because of its meaning. Moreover, (ii) like β_{dual}, the domain of useful β_g and $\beta_{h,\text{dual}}$ is not upper bounded. These two reasons make these other Lagrangians less preferable. We only include them here for completeness. Nonetheless, we encourage the curiours reader to explore these families of Lagrangians too.

F Experimental setup details

In order to generate the empirical support results from Section 5 we used the nonlinear IB (Kolchinsky et al., 2017) on the MNIST dataset (LeCun et al., 1998). This dataset contains 60,000 training samples and 10,000 testing samples of hand-written digits. The samples are 28x28 pixels and are labeled from 0 to 9; i.e., $\mathcal{X} = \mathbb{R}^{784}$ and $\mathcal{Y} = \{0, 1, ..., 9\}$.

As in (Kolchinsky et al., 2019) we trained the neural network with the Adam optimization algorithm (Kingma & Ba, 2014) with a learning rate of 10^{-4} but we introduced a 0.6 decay rate every 10 iterations. After talking with the authors of the nonlinear IB (Kolchinsky et al., 2017), we decided to estimate the gradients of both $I_\theta(X;T)$ and the cross entropy with the same mini-batch of 128 samples. Moreover, we did not learn the covariance of the mixture of Gaussians used for the kernel density estimation of $I_\theta(X;T)$ and we set it to $(\exp(-1))^2$. We trained for 100 epochs[6]. All the weights were initialized according to the method described by Glorot & Bengio (2010) using a Gaussian distribution. The reader can find the PyTorch (Paszke et al., 2017) implementation at `https://gofile.io/?c=G9Dl1L`.

Then, we used the DBSCAN algorithm (Ester et al., 1996; Schubert et al., 2017) for clustering. Particularly, we used the scikit-learn (Pedregosa et al., 2011) implementation with $\epsilon = 0.3$ and `min_samples` = 50.

In Figure 5 we show how the IB curve can be explored with different values of α for the power IB Lagrangian and in Figure 6 for different values of η and the exponential IB Lagrangian.

Finally, in Figure 7 we show the clusterization for the same values of α and η as in Figures 5 and 6. In this way the connection between the performance discontinuities and the clusterization is more evident. Furthermore, we can also observe how the exponential IB Lagrangian maintains better the theoretical performance than the power IB Lagrangian (see Appendix G for an explanation of why).

G Guidelines for selecting a proper function in the Convex IB Lagrangian

When chossing the right h function, it is important to find the right balance between avoiding value convergence and aiming for strong convexity. Practically, this balance is found by looking at how much faster h grows w.r.t. the identity function.

G.1 Avoiding value convergence

In order to explain this issue we are going to use the example of classification on MNIST (LeCun et al., 1998), where $I(X;Y) = H(Y) = \log_2(10)$, and again the power and exponential IB Lagrangians.

If we use Proposition 2 on both Lagrangians we obtain the bijective mapping between their Lagrange multipliers and a certain level of compression in the classification setting:

1. Power IB Lagrangian: $\beta_{\text{pow}} = ((1+\alpha)I(X;T)^\alpha)^{-1}$ and $I(X;T) = ((1+\alpha)\beta_{\text{pow}})^{-\frac{1}{\alpha}}$.
2. Exponential IB Lagrangian: $\beta_{\text{exp}} = (\eta \exp(\eta I(X;T)))^{-1}$ and $I(X;T) = -\log(\eta\beta_{\text{exp}})/\eta$.

Hence, we can simply plot the curves of $I(X;T)$ vs. β_h for different hyperparameters α and η (see Figure 8). In this way we can observe how increasing the growth of the function (e.g., increasing α or η in this case) too much provokes that many different values of β_h converge to very similar values of $I(X;T)$. This is an issue both for drawing the curve (for obvious reasons) and for aiming for a specific performance level. Due to the nature of the estimation of the IB Lagrangian, the theoretical and practical value of β_h that yield a specific $I(X;T)$ may vary slightly (see Figure 1). Then if we select a function with too high growth, a small change in β_h can result in a big change in the performance obtained.

[6] Note in the last version of the Nonlinear IB article in *arxiv* (v8) they explain many of this issues.

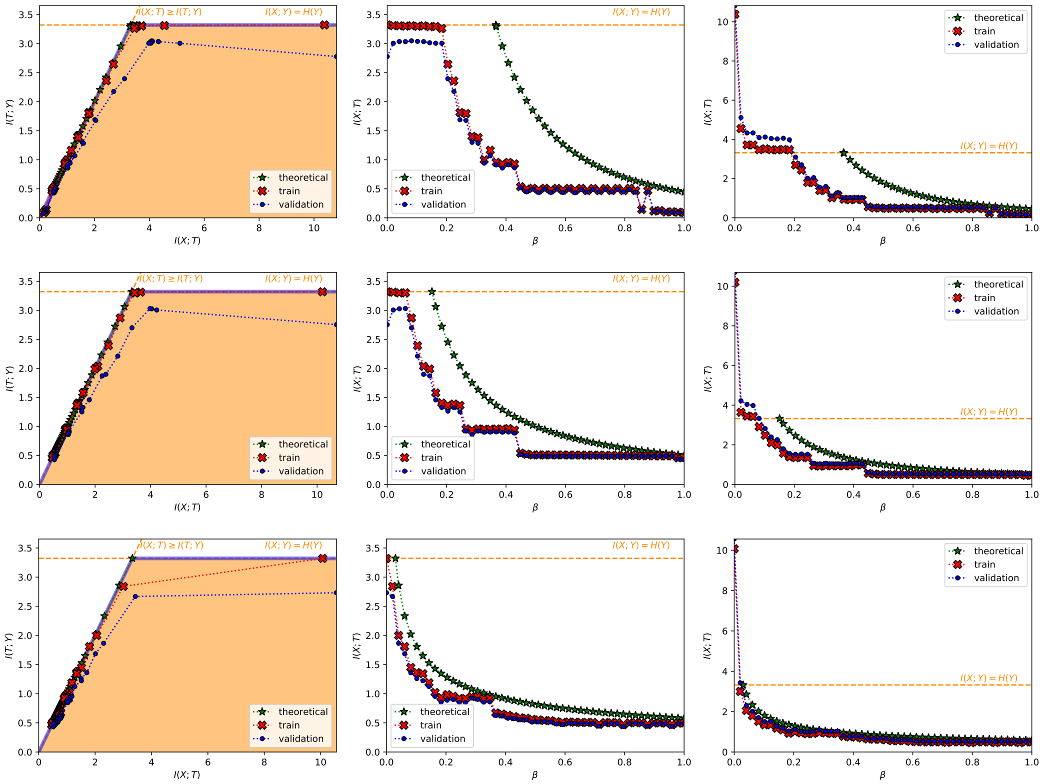


Figure 5: Results for the power IB Lagrangian with $\alpha = \{0.5, 1, 2\}$, from top to bottom. In each row, from left to right it is shown (i) the information plane, where the region of possible solutions of the IB problem is shadowed in light orange and the information-theoretic limits are the dashed orange line; (ii) $I(T;Y)$ as a function of β_h; and (iii) the compression $I(X;T)$ as a function of β_h. In all plots the red crosses joined by a dotted line represent the values computed with the training set, the blue dots the values computed with the validation set and the green stars the theoretical values computed as dictated by Proposition 2. Moreover, in all plots it is indicated $I(X;Y) = H(Y) = \log_2(10)$ in a dashed, orange line. All values are shown in bits.

G.2 Aiming for strong convexity

Definition 10 (**μ-Strong convexity**). *If a function $f(r)$ is twice continuous differentiable and its domain is confined in the real line, then it is μ-strong convex if $f''(r) \geq \mu \geq 0 \; \forall r$.*

Experimentally, we observed when the growth of our function $h(r)$ is small in the domain of interest $r > 0$ the convex IB Lagrangian does not perform well. Later we realized that this was closely related with the strength of the convexity of our function.

In Theorem 2 we imposed the function h to be strictly convex to enforce having a unique β_h for each value of $I(X;T)$. Hence, since in practice we are not exactly computing the Lagrangian but an estimation of it (e.g., with the nonlinear IB (Kolchinsky et al., 2017)) we require strong convexity in order to be able to explore the IB curve.

We now look at the second derivative of the power and exponential function: $h''(r) = (1+\alpha)\alpha r^{\alpha-1}$ and $h''(r) = \eta^2 \exp(\eta r)$ respectively. Here we see how both functions are inherently 0-strong convex for $r > 0$ and $\alpha, \eta > 0$. However, values of $\alpha < 1$ and $\eta < 1$ could lead to low μ-strong convexity in certain domains of r. Particularly, the case of $\alpha < 1$ is dangerous because the function approaches 0-strong convexity as r increases, so the power IB Lagrangian performs poorly when low α are used to find high performances.

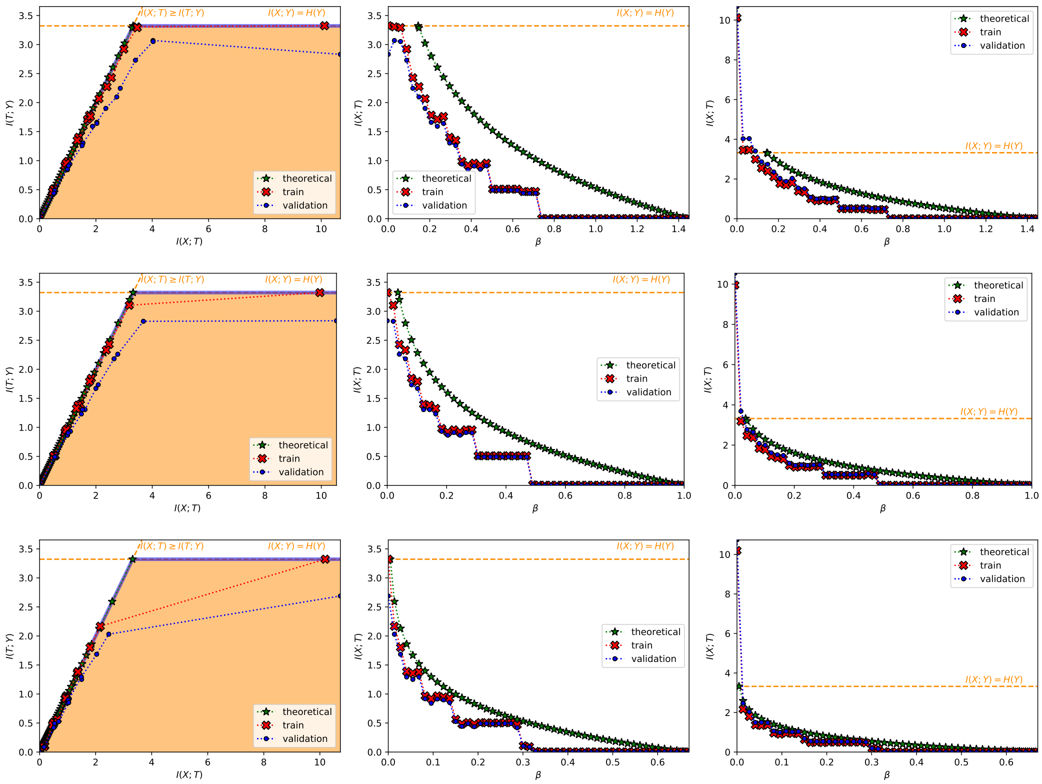


Figure 6: Results for the exponential IB Lagrangian with $\eta = \{\log(2), 1, 1.5\}$, from top to bottom. In each row, from left to right it is shown (i) the information plane, where the region of possible solutions of the IB problem is shadowed in light orange and the information-theoretic limits are the dashed orange line; (ii) $I(T;Y)$ as a function of β_h; and (iii) the compression $I(X;T)$ as a function of β_h. In all plots the red crosses joined by a dotted line represent the values computed with the training set, the blue dots the values computed with the validation set and the gren stars the theoretical values computed as dictated by Proposition 2. Moreover, in all plots it is indicated $I(X;Y) = H(Y) = \log_2(10)$ in a dashed, orange line. All values are shown in bits.

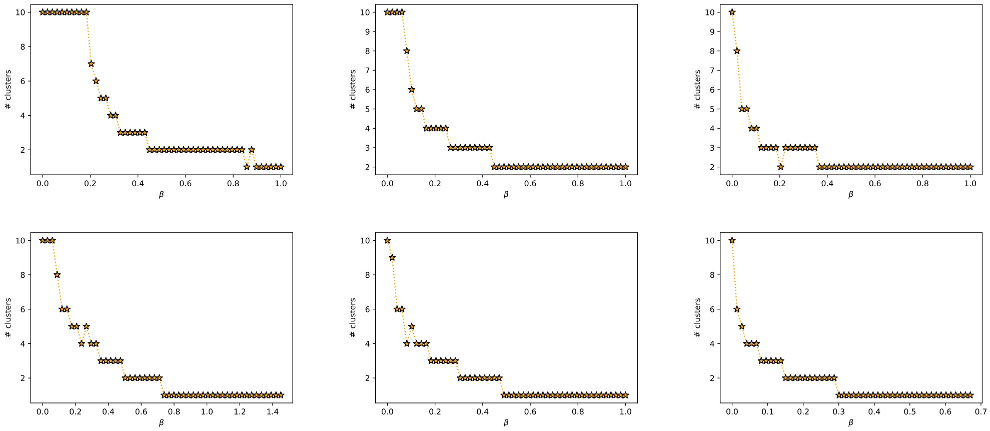


Figure 7: Depiction of the clusterization behavior of the bottleneck variable. In the first row, from left to right, the power IB Lagrangian with different values of $\alpha = \{0.5, 1, 2\}$. In the second row, from left to right, the exponential IB Lagrangian with different values of $\eta = \{\log(2), 1, 1.5\}$.

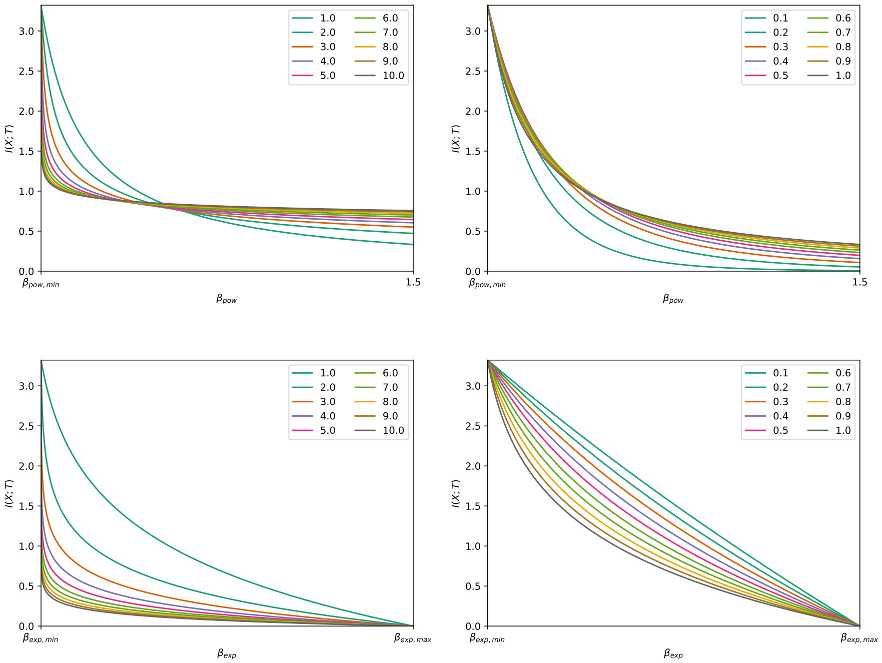


Figure 8: Theoretical bijection between $I(X;T)$ and different α from $\beta_{\text{h,min}}$ to 1.5 in the power IB Lagrangian (top), and different η in the domain B_h in the exponential IB Lagrangian (bottom).