Article

Not peer-reviewed version

β-Optimization in the Information Bottleneck Framework: A Theoretical Analysis

Faruk Alpay *

Posted Date: 12 May 2025

doi: 10.20944/preprints202505.0746.v1

Keywords: information bottleneck; mutual information; variational inference; lagrange multiplier; β-selection; convex optimization; representation learning; neural networks; theoretical analysis

 doi:10.20944/preprints202505.0746.v1

Article

β-Optimization in the Information Bottleneck Framework: A Theoretical Analysis

Faruk Alpay

Independent Researcher; alpay@lightcap.ai

Abstract: The Information Bottleneck (IB) framework formalizes the trade-off between compression and prediction in representation learning. A crucial parameter is the Lagrange multiplier β, which controls the balance between preserving information relevant to a target variable Y and compressing the representation Z of an input X. Selecting an optimal β (denoted β^*) is challenging and typically done via empirical tuning. In this paper, I present a rigorous theoretical analysis of β^*-optimization in both the Variational IB (VIB) and Neural IB (NIB) settings. I define β^* as the critical value of β that marks the boundary between non-trivial (informative) and trivial (uninformative) representations, ensuring maximal compression before the representation collapses. I derive formal conditions for its existence and uniqueness. I prove several key results: (1) the IB trade-off curve (relevance–compression frontier) is concave under mild conditions, implying that β, as the slope of this curve, uniquely characterizes optimal operating points in regular cases; (2) there exists a critical β threshold, $\beta^* = F'(0^+)$ (the slope of the IB curve at zero compression), beyond which the IB solution collapses to a trivial representation; (3) for practical IB implementations (VIB and NIB), I discuss how β^* can be computed algorithmically, including complexity analysis of naive β-sweeping versus adaptive methods like binary search, for which pseudo-code is provided. I provide formal theorems and proofs for concavity properties of the IB Lagrangian, continuity of the IB curve, and boundedness of mutual information quantities. Furthermore, I compare standard IB, VIB, and NIB formulations in terms of the optimal β, showing that while standard IB provides a theoretical target for β^*, variational and neural approximations may deviate from this optimum. My analysis is complemented by a discussion on the implications for deep neural network representations. The results establish a principled foundation for β selection in IB, guiding practitioners to achieve maximal meaningful compression without exhaustive trial-and-error.

Keywords: information bottleneck; mutual information; variational inference; lagrange multiplier; β-selection; convex optimization; representation learning; neural networks; theoretical analysis

---

1. Introduction

In modern representation learning, the Information Bottleneck (IB) principle provides a theoretical framework for extracting a concise yet informative representation of data [1]. Given an input random variable X and a target variable Y, the IB method seeks a "bottleneck" variable Z (the representation) that compresses X while preserving information useful for predicting Y. Formally, Z is chosen to maximize the mutual information $I(Z;Y)$ under a constraint on $I(Z;X)$ [1]. Equivalently, one can solve the Lagrangian formulation introduced by Tishby et al. [1], which defines the IB objective as:

$$\mathcal{L}(p(z|x);\beta) = I(Z;Y) - \beta I(Z;X), \qquad \text{with } \beta \geq 0, \tag{1}$$

where β is a non-negative Lagrange multiplier controlling the trade-off between the prediction term $I(Z;Y)$ and the compression term $I(Z;X)$. By adjusting β, one traces out the Pareto-optimal trade-offs between retaining information about Y versus compressing X.

Selecting an optimal β: In practice, choosing the "right" β is non-trivial and has traditionally relied on brute-force search or cross-validation. Tishby et al. [1] initially suggested "sweeping" over

 doi:10.20944/preprints202505.0746.v1

β to plot the IB curve and then picking an operating point. This trial-and-error approach can be computationally expensive and may fail to capture the subtle trade-offs in complex data. The need for a principled criterion to determine an optimal β (denoted β^*) — one yielding a representation Z^* that achieves a specific kind of optimality in the compression-prediction balance — motivates my work. I ask: How can one characterize and compute β^* theoretically, without resorting solely to empirical tuning?

Recent studies have begun to address this question. Wu et al. (2019) [9] introduced the concept of IB-learnability to provide guidance on choosing β. They define conditions under which a given β will avoid trivial solutions (i.e., Z independent of X) and derive a threshold for β (specific to (X, Y)) related to the existence of meaningful representations. Independently, Rodríguez Gálvez et al. (2020) [8] analyzed the mapping between β and the compression rate, showing that under certain convexity assumptions each point on the IB curve corresponds to a unique β. These works suggest that an optimal β could be identified as a critical value related to the IB curve's geometry. Moreover, a recent multi-objective perspective treats IB as a bi-objective optimization (maximizing $I(Z;Y)$ and minimizing $I(Z;X)$ jointly) to adaptively find trade-offs without a fixed β [12]. Such methods confirm that finding the best balance between compression and prediction is challenging and important.

In deep learning, two notable IB-based paradigms have emerged: the Variational Information Bottleneck (VIB) [5] and what I term the Neural Information Bottleneck (NIB). VIB refers to the variational approximation introduced by Alemi et al. [5], which employs deep neural networks and variational inference to approximate $I(Z;X)$ and $I(Z;Y)$, making IB applicable to high-dimensional continuous data. NIB, in this paper, denotes IB implementations that use neural estimation or deterministic encoders instead of the analytic variational bound – for example, using neural mutual information estimators (like MINE [10]) or the Deterministic IB (DIB) method [4]. Both VIB and NIB aim to optimize the same IB trade-off, but their behavior and optimal β may differ due to approximation error or different notions of compression (stochastic vs. deterministic). A comprehensive theoretical treatment must encompass both settings and clarify how β^* manifests in each.

This paper provides a full-length theoretical analysis of β-optimization within the IB framework. My contributions are as follows:

- **Rigorous Definition of β^***: I formalize β^* as the critical β value that marks the boundary between non-trivial (informative) and trivial (uninformative) representations. This β^* corresponds to the slope of the IB curve at the origin, $F'(0^+)$, representing the point of maximal compression beyond which the representation collapses. This definition is made precise in Section 3.
- **Theoretical Properties and Existence**: I derive conditions under which β^* exists and is unique. Key properties such as the concavity of the IB curve (as a function $I(Z;Y)$ vs $I(Z;X)$) and the continuity and monotonicity of optimal solutions w.r.t. β are proven. I show, for instance, that $I(Z_\beta;Y)$ and $I(Z_\beta;X)$ are non-increasing in β. I prove that there is a critical $\beta_c = F'(0^+)$ (which I define as β^*) beyond which the only solution is the trivial one (Z carries no information from X).
- **Algorithmic Discussion and Complexity**: I discuss how one can solve for β^* in practice. I provide pseudo-code for a binary search procedure on β. I analyze the complexity of naïvely sweeping β versus more efficient methods that leverage my theoretical insights (e.g., using the properties of the IB Lagrangian to pinpoint β^*). I also compare the computational complexity of VIB and NIB approaches.
- **Comparisons of IB, VIB, and NIB**: I provide a comparative analysis of how β^* should be interpreted in standard IB theory versus in VIB and NIB implementations. I show, for example, that if the VIB approximation is tight, the chosen β in VIB corresponds closely to the true β^*; however, if variational bounds are loose [6], the effective trade-off might differ. In the NIB setting (e.g., DIB [4]), I examine how replacing the mutual information constraint with alternative penalties changes the β^*-criterion. Formal propositions highlight these differences.

The remainder of the paper is organized as follows. In Section 2, I review the IB framework and introduce the formal definition of β^*, along with preliminaries on mutual information and optimization

 doi:10.20944/preprints202505.0746.v1

under constraints. In Section 3, I present my main theorems on the existence and characterization of β^*, with proofs of concavity, continuity, and boundedness conditions that underpin β^*-optimization. In Section 4, I connect my findings to algorithmic strategies and practical IB variants (VIB/NIB), and I outline implications for neural network models. Finally, Section 5 summarizes my contributions and suggests future research directions.

Throughout, I maintain an academic tone and mathematical rigor, aiming to ensure clarity and completeness of all proofs and definitions. All key results are backed by references to foundational works (e.g., information theory [14] and IB literature) for verification and context. I use formal notation consistent with information theory texts. By providing both theorems and intuitive explanations, I hope this work serves as a solid theoretical foundation for choosing β in Information Bottleneck applications.

2. Methodology

2.1. Background: Information Bottleneck Framework

Consider random variables X (input or source) and Y (target or label) with a joint distribution $p(x, y)$. The Information Bottleneck method [1] introduces an auxiliary variable Z (the representation or "bottleneck") such that $Y - X - Z$ forms a Markov chain. This means Z is obtained by some probabilistic encoder $p(z|x)$ and is intended to keep only the information in X that is relevant to predicting Y. The IB principle can be stated as the constrained optimization problem [1]:

$$\max_{p(z|x)} I(Z;Y) \quad \text{s.t.} \quad I(Z;X) \leq R,$$

where $I(Z;Y)$ is the mutual information between Z and Y, and $I(Z;X)$ measures how much information Z retains about X. The constraint $I(Z;X) \leq R$ (for some compression level R) enforces that Z is a compressed version of X.

Using Lagrange duality, one solves this by introducing a Lagrange multiplier $\beta \geq 0$ to form the IB Lagrangian [1]:

$$\mathcal{L}(p(z|x);\beta) = I(Z;Y) - \beta I(Z;X). \tag{2}$$

Here β controls the trade-off: a larger β places more penalty on $I(Z;X)$ (compression), while a smaller β prioritizes $I(Z;Y)$ (prediction). Varying β from 0 to ∞ traces out the IB curve, which is the set of optimal $(I(Z;X), I(Z;Y))$ pairs achievable [1]. Specifically, as β increases from 0, optimal $I(Z;X)$ typically decreases (stricter compression) and optimal $I(Z;Y)$ decreases (some predictive information is sacrificed). For $\beta = 0$, one recovers the unconstrained maximum $I(Z;Y)$ (which is $I(X;Y)$ if $Z = X$ is allowed); for $\beta \to \infty$, one enforces extreme compression ($I(Z;X) \to 0$), usually at the cost of $I(Z;Y) \to 0$ (trivial Z independent of X).

Mutual Information Basics: Recall that mutual information $I(U;V)$ is defined as $I(U;V) = H(U) - H(U|V) = H(V) - H(V|U)$, where $H(\cdot)$ is Shannon entropy. Under the Markov chain $Y - X - Z$, the Data Processing Inequality (DPI) states $I(Z;Y) \leq I(X;Y)$ [14]. The inequality $I(Z;Y) \leq I(Z;X)$ also often holds for meaningful IB solutions, as Z cannot convey more information about Y (which is related to X) than it holds about X itself, especially if Z is a deterministic function of X or if an additional Markov chain $X - Z - Y_{pred}$ is assumed for prediction. These imply that on the IB plane (with $I(Z;X)$ on x-axis and $I(Z;Y)$ on y-axis), all achievable points lie under the line $I(Z;Y) = I(Z;X)$ (typically) and below $I(Z;Y) = I(X;Y)$, and to the left of $I(Z;X) = H(X)$. The feasible region is bounded: $0 \leq I(Z;Y) \leq I(X;Y)$ and $0 \leq I(Z;X) \leq H(X)$.

 doi:10.20944/preprints202505.0746.v1

Optimal Representations: Solving Eq. (2) means finding an encoder $p^*(z|x)$ that maximizes $\mathcal{L}$. For fixed β, standard results give a set of self-consistent equations for the optimum. In the discrete case, these are [1]:

$$p^*(z|x) = \frac{p^*(z)}{N(x,\beta)} \exp(-\beta D_{\mathrm{KL}}(p(y|x)\|p^*(y|z))), \tag{3}$$

$$p^*(z) = \sum_x p(x)p^*(z|x), \tag{4}$$

$$p^*(y|z) = \frac{1}{p^*(z)} \sum_x p(x,y)p^*(z|x), \tag{5}$$

where $D_{\mathrm{KL}}(\cdot\|\cdot)$ is the Kullback-Leibler divergence and $N(x,\beta)$ is a normalization factor for $p^*(z|x)$. These iterative updates (e.g., Blahut-Arimoto style algorithm for IB) converge to a (locally) optimal $p(z|x)$.

2.2. Defining β^* (Optimal Trade-off Parameter)

I now define β^*, the optimal β in the IB sense. Intuitively, β^* should correspond to a point on the IB curve that represents a critical trade-off.

Definition 1 (β^* as Critical Point on the IB Curve)**.** *Let $F(r) = \max\{I(Z;Y) : I(Z;X) \leq r\}$ be the IB frontier function, giving the maximal achievable $I(Z;Y)$ for a given compression level $r = I(Z;X)$. Assume $F(r)$ is concave and differentiable on $0 < r < H(X)$. The parameter β in the IB Lagrangian (2) corresponds to the slope of the IB curve, i.e., $\beta = F'(r)$ at an optimal point $(r, F(r))$. I define β^* as the critical value β_c corresponding to the slope of the IB curve at the origin (maximal compression end):*

$$\beta^* = \beta_c = F'(0^+) = \lim_{r\to 0^+} F'(r). \tag{6}$$

This β^* is the largest value of β for which the IB solution is marginally non-trivial. For any $\beta > \beta^*$, the optimal solution is the trivial encoding ($Z \perp X$, $I(Z;X) = 0, I(Z;Y) = 0$). For $\beta \leq \beta^*$, a non-trivial solution with $I(Z;X) > 0$ and $I(Z;Y) > 0$ can exist.

This definition aligns with the concept of IB-Learnability by Wu et al. [9], who identified a similar threshold. Their threshold $\beta_{c,Wu} = 1/\eta_{XY}$, where $\eta_{XY} = \lim_{I(Z;X)\to 0} I(Z;Y)/I(Z;X) = F'(0^+)$ (assuming $F(0) = 0$). If their β parameter is interpreted as $1/\text{slope}$ (e.g., in a Lagrangian like $I(Z;X) - \lambda I(Z;Y)$), then it is consistent. However, if their β is used in the same Lagrangian form as Eq. (2), then their threshold would be $F'(0^+)$. The precise relationship depends on the specific Lagrangian formulation. In this paper, β is consistently the coefficient of $I(Z;X)$ as in Eq. (2), and thus $\beta^* = F'(0^+)$.

This β^* represents the point of maximal compression pressure under which the system still extracts some meaningful information. Beyond this β^*, the compression penalty is so high that it is optimal to discard all information about X. Thus, β^* identifies the operating point that is most compressed while still being informative.

2.3. IB in VIB and NIB Settings

Variational IB (VIB): VIB approximates the IB objective by introducing a parametric encoder $q_\phi(z|x)$ and a decoder $p_\theta(y|z)$, optimizing:

$$\mathcal{L}_{\mathrm{VIB}}(\phi,\theta) = \mathbb{E}_{p(x)}\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(y|z)] - \beta\mathbb{E}_{p(x)}[D_{\mathrm{KL}}(q_\phi(z|x)\|p(z))],$$

where $p(z)$ is a fixed prior (e.g., standard Gaussian). The first term is a lower bound on $I(Z;Y)$ (related to cross-entropy), and the second term $K = \mathbb{E}_{p(x)}[D_{\mathrm{KL}}(q_\phi(z|x)\|p(z))]$ relates to $I(Z;X)$. Specifically,

 doi:10.20944/preprints202505.0746.v1

$K = I_q(Z;X) + D_{\mathrm{KL}}(q(z)\|p(z))$, where $q(z) = \mathbb{E}_{p(x)}[q_\phi(z|x)]$ is the true marginal. Thus K is an upper bound on $I_q(Z;X)$. The parameter β in VIB plays the same qualitative role.

Neural IB (NIB): This term covers IB implementations using, e.g., neural mutual information estimators like MINE [10] for $I(Z;X)$, or alternative complexity measures. An example is Deterministic IB (DIB) [4], which optimizes $\max I(Z;Y) - \beta H(Z)$. Since $I(Z;X) \leq H(Z)$ (as $I(Z;X) = H(Z) - H(Z|X)$), penalizing $H(Z)$ is different from penalizing $I(Z;X)$. DIB encourages deterministic encoders where $H(Z|X) \approx 0$, so $I(Z;X) \approx H(Z)$. An optimal β^* can be defined similarly for these variants as the threshold for non-trivial solutions.

My theoretical results primarily apply to the standard IB formulation, but their implications for VIB and NIB will be discussed.

3. Theoretical Results

I now present the main theoretical results regarding β^* in the IB framework.

3.1. Properties of the IB Lagrangian and the Trade-off Curve

Lemma 1 (Monotonicity and Bounds)**.** *Let $(I_\beta^X, I_\beta^Y) = (I(Z_\beta;X), I(Z_\beta;Y))$ be the coordinates of an optimal IB solution Z_β for a given $\beta \geq 0$. Then, as functions of β: (i) I_β^X is non-increasing. (ii) I_β^Y is non-increasing. Moreover, $0 \leq I_\beta^Y \leq I(X;Y)$ and $0 \leq I_\beta^X \leq H(X)$. As $\beta \to 0$, $I_\beta^Y \to I(X;Y)$ (if $Z = X$ is achievable) and $I_\beta^X \to I(X;X) = H(X)$ (if $Z = X$). As $\beta \to \infty$, $I_\beta^X \to 0$ and $I_\beta^Y \to 0$.*

Proof. (i) Let $\beta_1 < \beta_2$. Let Z_1 be optimal for β_1 and Z_2 for β_2. From optimality: $I(Z_1;Y) - \beta_1 I(Z_1;X) \geq I(Z_2;Y) - \beta_1 I(Z_2;X)$ $I(Z_2;Y) - \beta_2 I(Z_2;X) \geq I(Z_1;Y) - \beta_2 I(Z_1;X)$ Adding these inequalities: $-\beta_1 I(Z_1;X) - \beta_2 I(Z_2;X) \geq -\beta_1 I(Z_2;X) - \beta_2 I(Z_1;X)$ Rearranging gives: $(\beta_2 - \beta_1)(I(Z_1;X) - I(Z_2;X)) \geq 0$. Since $\beta_2 - \beta_1 > 0$, we must have $I(Z_1;X) - I(Z_2;X) \geq 0$, so $I(Z_1;X) \geq I(Z_2;X)$. Thus I_β^X is non-increasing.

(ii) Since I_β^X is non-increasing, and $\beta = F'(I_\beta^X)$ where F is the concave IB curve (Theorem 1), an increase in β corresponds to moving to a point on the curve with smaller I_β^X. Since $F(r)$ is non-decreasing, smaller I_β^X generally implies smaller or equal I_β^Y. More formally, $I_\beta^Y = F(I_\beta^X)$. Since F is non-decreasing and I_β^X is non-increasing, I_β^Y must be non-increasing.

The bounds $0 \leq I_\beta^Y \leq I(X;Y)$ (by DPI, $Y - X - Z$) and $0 \leq I_\beta^X \leq H(X)$ (since $I(Z;X) \leq H(X)$) are standard. As $\beta \to 0$, the objective becomes $\max I(Z;Y)$. The solution approaches $Z = X$ (if cardinality allows), yielding $I(Z;Y) \approx I(X;Y)$ and $I(Z;X) \approx I(X;X) = H(X)$. As $\beta \to \infty$, the term $-\beta I(Z;X)$ dominates. To maximize the Lagrangian, $I(Z;X)$ must be minimized, so $I_\beta^X \to 0$. Consequently, $I_\beta^Y \to 0$ (since $I(Z;Y) \leq I(Z;X)$ if Z is a deterministic function of X, or more generally, $I(Z;Y) \to 0$ as $I(Z;X) \to 0$). □

Theorem 1 (Concavity of the IB Curve)**.** *The set of achievable pairs $(I(Z;X), I(Z;Y))$, resulting from any encoder $p(z|x)$, forms a convex set in the $(I(Z;X), I(Z;Y))$-plane. The frontier $F(r) = \max\{I(Z;Y) : I(Z;X) \leq r\}$ is a concave, non-decreasing function of r.*

Proof. This is a standard result in information theory, often proven using a time-sharing argument [14,15]. Consider two encoders $p_1(z|x)$ and $p_2(z|x)$ achieving points $(r_1, u_1) = (I_1(Z;X), I_1(Z;Y))$ and (r_2, u_2) respectively. A new encoder can be constructed by choosing $p_1(z|x)$ with probability λ and $p_2(z|x)$ with probability $1 - \lambda$. The resulting representation Z achieves $(r, u) = (\lambda r_1 + (1 - \lambda) r_2, \lambda u_1 + (1 - \lambda) u_2)$. This means any point on the line segment connecting (r_1, u_1) and (r_2, u_2) is achievable. Thus, the set of all achievable points is convex. The function $F(r)$ is the upper boundary of this convex set, and is therefore concave and non-decreasing. □

The concavity of $F(r)$ implies that its derivative $F'(r)$ (where it exists) is non-increasing. Since $\beta = F'(r)$ for an optimal solution, this is consistent with Lemma 1. If $F(r)$ is strictly concave, the

 doi:10.20944/preprints202505.0746.v1

mapping from r to $\beta = F'(r)$ is one-to-one. If $F(r)$ has linear segments, multiple r values can correspond to the same β (if β is the slope of that segment), or one r can correspond to a range of β values (if r is a kink point).

Example 1 (Deterministic Y and IB Curve Plateau)**.** *Suppose $Y = f(X)$ is a deterministic function of X. Then $I(X;Y) = H(Y)$. The IB curve $F(r)$ may exhibit a plateau at $I(Z;Y) = H(Y)$ for $r \geq r_{min}$, where r_{min} is the minimum $I(Z;X)$ required to perfectly predict Y. For $r \geq r_{min}$, $F'(r) = 0$. Thus, any $\beta \in [0, F'(r_{min}^-))$ (if r_{min} is a kink) or $\beta = 0$ (if smooth) would yield a solution on this plateau. This scenario is discussed in [7]. The β^* as defined by $F'(0^+)$ would still be positive and characterize the other end of the curve, unless $F'(0^+) = 0$ (e.g., if Y is independent of X).*

3.2. Existence and Characterization of β^*

Theorem 2 (Existence of Critical β^*)**.** *There exists a unique critical value $\beta^* = F'(0^+) \in [0, \infty)$ such that: (i) For all $\beta > \beta^*$, the optimal IB solution to Eq. (2) is the trivial encoder ($Z \perp X$, yielding $I(Z;X) = 0$ and $I(Z;Y) = 0$). (ii) For $\beta < \beta^*$, if $F'(0^+) > 0$, a non-trivial optimal solution with $I(Z;X) > 0$ and $I(Z;Y) > 0$ exists. (iii) At $\beta = \beta^*$, a non-trivial solution may exist if the slope $F'(0^+)$ is achieved by some $r > 0$.*

Proof Sketch. We want to maximize $\mathcal{L}(r;\beta) = F(r) - \beta r$ over $r \geq 0$. The value of the trivial solution ($r = 0$) is $F(0) - \beta \cdot 0 = 0$ (assuming $F(0) = 0$, i.e., zero compression implies zero information about Y if Z is independent of X). Consider a non-trivial solution $r > 0$. If $F(r)$ is differentiable, the first-order condition for an interior maximum is $F'(r) - \beta = 0$, so $\beta = F'(r)$. Since $F(r)$ is concave, $F'(r)$ is non-increasing. Let $s_0 = F'(0^+) = \lim_{r \to 0^+} F'(r)$. This is the maximum possible slope. (i) If $\beta > s_0$: Since $F'(r) \leq s_0$ for all $r > 0$, there is no $r > 0$ such that $F'(r) = \beta$. Consider the function $g(r) = F(r) - \beta r$. Its derivative is $g'(r) = F'(r) - \beta$. If $\beta > s_0$, then $F'(r) < \beta$ for all $r > 0$ where $F'(r)$ is defined (assuming $F'(r)$ is strictly decreasing or s_0 is not attained for $r > 0$). So $g'(r) < 0$. This means $g(r)$ is decreasing. Thus, its maximum over $r \geq 0$ is at $r = 0$. So the trivial solution is optimal. (ii) If $\beta < s_0$: Then there exists some $r^* > 0$ such that $F'(r^*) = \beta$ (if $F'(r)$ spans the range $(0, s_0]$). For small $r > 0$, $F(r) \approx s_0 r$. Then $F(r) - \beta r \approx (s_0 - \beta)r$. Since $s_0 - \beta > 0$, this value is positive for $r > 0$, hence better than the trivial solution's value of 0. So a non-trivial solution is optimal. (iii) If $\beta = s_0$: The optimum may occur at $r = 0$ or at some $r > 0$ if $F'(r) = s_0$ for some $r > 0$ (e.g., if $F(r)$ starts with a linear segment of slope s_0). The uniqueness of $\beta^* = s_0$ follows from it being a specific value determined by $F(r)$ at $r = 0^+$. A more formal proof can be constructed by analyzing the properties of the dual function $g(\beta) = \max_{p(z|x)} \mathcal{L}(p(z|x);\beta)$, which is convex in β. The transition point defines β^*. Wu et al. [9] provide a detailed analysis of such a threshold (termed β_c in their work, potentially defined as $1/s_0$ depending on their Lagrangian formulation, but conceptually similar). □

This theorem establishes the existence and uniqueness of β^* as defined. It is the point where the IB solution effectively collapses if compression pressure is increased further.

Theorem 3 (Properties of the Representation at β^*)**.** *Let $\beta^* = F'(0^+)$ be the critical threshold from Theorem 2. If an optimal encoder Z_{β^*} exists for $\beta = \beta^*$: (i) Z_{β^*} represents the maximally compressed encoding of X that can still retain non-zero information about Y (if $F'(0^+) > 0$). (ii) The quantity $1/\beta^* = 1/F'(0^+)$ can be interpreted as the maximum possible "predictive efficiency" $I(Z;Y)/I(Z;X)$ achievable in the limit of very high compression ($I(Z;X) \to 0$). (iii) Z_{β^*} is not generally a minimally sufficient statistic for Y in the sense of achieving $I(Z_{\beta^*};Y) = I(X;Y)$ with minimal $I(Z_{\beta^*};X)$. That property corresponds to a point at the low-compression end of the IB curve (typically $\beta \to 0$).*

Proof. (i) By definition of β^*, for any $\beta > \beta^*$, the optimal solution is trivial ($I(Z;X) = 0, I(Z;Y) = 0$). For $\beta \leq \beta^*$ (specifically, for β approaching β^* from below), non-trivial solutions exist. Thus, β^* marks the boundary. A solution Z_{β^*} (if non-trivial) is obtained under the highest compression pressure (β) that still permits $I(Z;Y) > 0$. (ii) Since $\beta^* = F'(0^+) = \lim_{r \to 0} \frac{F(r) - F(0)}{r - 0} = \lim_{I(Z;X) \to 0} \frac{I(Z;Y)}{I(Z;X)}$ (assuming

 doi:10.20944/preprints202505.0746.v1

$F(0) = 0$). Thus, $1/\beta^*$ is the inverse of this limiting slope, representing bits of $I(Z;X)$ per bit of $I(Z;Y)$ at maximal compression. Or, β^* itself is the marginal gain in $I(Z;Y)$ per bit of $I(Z;X)$ at $I(Z;X) \to 0$. (iii) A minimally sufficient statistic Z_{MSS} for Y from X satisfies $I(Z_{MSS};Y) = I(X;Y)$ and $I(Z_{MSS};X)$ is minimized subject to this. This point $(I(Z_{MSS};X), I(X;Y))$ lies on the IB curve, typically where $I(Z;X)$ is relatively large (low compression). The corresponding $\beta_{MSS} = F'(I(Z_{MSS};X))$ is typically small (e.g., $\beta_{MSS} \approx 0$ if this point is on a plateau where $F'(r) = 0$). This is distinct from $\beta^* = F'(0^+)$, which is typically large. □

This theorem clarifies the nature of the representation at $\beta^* = F'(0^+)$. It is not about full sufficiency but about the efficiency at the edge of informativeness.

3.3. Algorithmic Implications and Complexity

Determining β^* in practice involves finding $F'(0^+)$.

- **Sweep and Search**: One can solve the IB optimization for a range of β values and observe where the solution transitions from non-trivial to trivial. A binary search on β is more efficient. The complexity is roughly $O(N_{\text{iter}} \cdot C_{\text{IB\_solve}})$, where $C_{\text{IB\_solve}}$ is the cost of solving the IB objective for a fixed β, and N_{iter} is the number of iterations for the search (e.g., $\log_2(\text{range}/\epsilon)$ for binary search with precision ϵ). For discrete X, Y, $C_{\text{IB\_solve}}$ using Blahut-Arimoto style iterations is polynomial in alphabet sizes $|X|, |Y|, |Z|$ [1].
- **Frontier Geometry Methods**: If $F(r)$ is known analytically (e.g., Gaussian IB [3]), $F'(0^+)$ can be computed directly. Alternatively, methods estimating the (hyper)contraction coefficient of the channel $X \to Y$ can estimate $F'(0^+)$ and thus β^* (or $1/\beta^*$ depending on formulation) [9]. Multi-objective optimization techniques might generate the Pareto front, from which $F'(0^+)$ could be estimated [12].

Algorithm 1 outlines a binary search approach to find β^*.

Algorithm 1 Binary Search for $\beta^* = F'(0^+)$

Require: Joint distribution $p(x,y)$, tolerance ϵ_β, small threshold $\delta_I > 0$.
Initialize $\beta_{low} = 0$, $\beta_{high} =$ large_value (e.g., estimated upper bound for $F'(0^+)$).
function SOLVEIB(β_{test})
  Solve $Z_{test} = \arg\max_{p(z|x)}(I(Z;Y) - \beta_{test} I(Z;X))$.
  return $I(Z_{test};X)$.
end function
while $(\beta_{high} - \beta_{low}) > \epsilon_\beta$ **do**
  $\beta_{mid} = (\beta_{low} + \beta_{high})/2$.
  $I_X\_mid = \text{SolveIB}(\beta_{mid})$. ▷ Variable name changed to be valid
  if $I_X\_mid < \delta_I$ **then** ▷ Solution is trivial or near-trivial
    $\beta_{high} = \beta_{mid}$.
  else ▷ Solution is non-trivial
    $\beta_{low} = \beta_{mid}$.
  end if
end while
$\beta^* \approx \beta_{low}$ (or β_{high}, depending on convention for boundary).
return β^*

This algorithm seeks the largest β for which the solution is non-trivial. δ_I is a small positive constant to numerically check for $I(Z;X) \approx 0$.

4. Discussion

4.1. β^* in Variational IB (VIB)

In VIB [5], β plays a similar role, but several factors can affect the observed β^*:

 doi:10.20944/preprints202505.0746.v1

- **Approximation Error**: VIB uses bounds for $I(Z;Y)$ and $I(Z;X)$. If these bounds are not tight, or if the parametric encoder/decoder families are not expressive enough, the VIB-optimized curve may differ from the true IB curve [6]. This can shift the empirically observed β^* for collapse.
- **Collapse Phenomenon**: VIB models are known to exhibit a collapse phenomenon: for too large β, the encoder $q_\phi(z|x)$ learns to ignore x and $q_\phi(z|x) \approx p(z)$ (the prior), making $I(Z;X) \approx 0$. This empirical collapse threshold in VIB is the analogue of the theoretical β^*.
- **Practical Estimation**: Practitioners often find a suitable β by sweeping values and observing validation performance. The largest β that maintains good performance before a sharp drop could be considered an empirical estimate of a "useful" β, which might be lower than the strict collapse threshold $\beta^* = F'(0^+)$ if some minimal $I(Z;Y)$ is required.

The theory of β^* provides a target: VIB training should ideally operate with $\beta \leq \beta^*$ to avoid complete information loss.

4.2. β^* in Neural IB (NIB)

NIB approaches, like DIB [4] or methods using MINE [10], also involve a trade-off parameter analogous to β.

- **Deterministic IB (DIB)**: DIB optimizes $I(Z;Y) - \beta H(Z)$. Since $I(Z;X) \leq H(Z)$, DIB penalizes an upper bound on $I(Z;X)$. DIB tends to find deterministic encoders where $I(Z;X) \approx H(Z)$. The critical β^*_{DIB} for collapse in DIB will exist but may have a different numerical value than β^*_{IB} due to the different complexity term.
- **MI Estimators**: Using neural MI estimators for $I(Z;X)$ can be noisy. Detecting the exact β^* where $I(Z;X)$ (and thus $I(Z;Y)$) truly vanishes can be hard. However, the principle of a collapse threshold remains.

In all NIB variants, β^* (appropriately defined for the specific objective) marks the boundary of useful compression.

4.3. Generalization and Robustness Considerations

The IB framework is linked to generalization in machine learning [2,11]. Compressing representations (larger β) can discard irrelevant information, potentially improving generalization by preventing overfitting. $\beta^* = F'(0^+)$ represents the most extreme compression. Operating slightly below β^* might yield representations that are highly compressed yet still informative. Choosing β to optimize generalization often involves finding a balance, possibly at a "knee" of the IB curve, which is different from $\beta^* = F'(0^+)$. However, β^* provides a hard upper limit on useful β values. Similar arguments apply to robustness: IB might discard fragile, non-robust features.

4.4. Multi-Target or Multi-Layer Extensions

For multiple targets $Y_1, Y_2, \ldots$, or for IB applied at multiple layers of a deep network [16], the concept of β-optimization becomes more complex. One might have a vector of β parameters or layer-specific β^*_i. The fundamental idea of a critical threshold where information is lost would likely generalize, but its characterization would be more involved.

4.5. Limitations and Assumptions

The analysis relies on certain assumptions:

- Concavity and differentiability of $F(r)$: For some distributions, $F(r)$ might have kinks or linear segments. β^* as $F'(0^+)$ still exists (as a one-sided derivative).
- Existence of optimal encoders: Assumed for theoretical IB. In practice (VIB/NIB), model capacity and optimization are critical. If models are too restricted, apparent collapse might occur earlier due to capacity limits rather than β itself.

 doi:10.20944/preprints202505.0746.v1

5. Conclusion

In this work, I presented a comprehensive theoretical study of β-optimization in the Information Bottleneck framework. I formally defined β^* as the critical Lagrange multiplier $\beta^* = F'(0^+)$, which marks the boundary where the IB solution transitions from being informative (non-trivial) to uninformative (trivial). This β^* represents the maximal compression pressure under which a representation can still convey some information about the target.

My analysis yielded the following key takeaways:

- β^* exists and is unique for a given X–Y distribution, identifiable as $F'(0^+)$ (Theorem 2). It signifies the point beyond which further increase in the compression penalty β leads to a complete loss of information.
- The IB trade-off curve $F(r)$ is concave (Theorem 1), ensuring a well-behaved relationship between β (as the slope $F'(r)$) and the optimal information measures $(I(Z;X), I(Z;Y))$.
- At $\beta^* = F'(0^+)$, the representation Z_{β^*} is maximally compressed while potentially retaining the initial, most "efficient" bits of information about Y (Theorem 3). This is distinct from concepts like minimal sufficiency for Y (i.e., $I(Z;Y) = I(X;Y)$), which occurs at the other end of the IB curve (typically $\beta \approx 0$).
- Algorithmic approaches, such as binary search (Algorithm 1), can be used to estimate β^* in practice.
- The interpretation of β^* extends to VIB and NIB, where analogous collapse phenomena are observed, though the exact value may be affected by approximations or alternative objective formulations.

My findings offer practical guidance by providing a principled understanding of β^*. This can help reduce reliance on ad-hoc tuning. For instance, estimating $F'(0^+)$ or using informed search strategies can guide the selection of β.

Future Work: Several avenues for future research emerge:

- Developing adaptive algorithms that dynamically tune β towards β^* (or a desired point relative to β^*) during training.
- Investigating robust estimation of β^* from finite samples, especially in high-dimensional settings.
- Extending the theory of β^*-optimization to more complex scenarios, such as multi-target IB, sequential IB (e.g., for time-series data or reinforcement learning), or hierarchical IB in deep networks.
- Exploring the relationship between $\beta^* = F'(0^+)$ and other notions of an "optimal" β, such as one corresponding to the "knee" of the IB curve or one optimizing generalization performance on a validation set. While $\beta^* = F'(0^+)$ is a mathematically precise critical point, other definitions might be more relevant for specific practical goals.

In conclusion, this work grounds β-optimization in the IB framework with a formal understanding of β^* as the critical threshold for informativeness. I hope this theoretical analysis contributes to a more principled application of the Information Bottleneck method in designing efficient and effective representation learning systems.

References

1. Tishby, N., Pereira, F.C., Bialek, W. (2000). The information bottleneck method. *Proc. of 37th Allerton Conference on Communication, Control, and Computing*.
2. Shamir, O., Sabato, S., Tishby, N. (2010). Learning and Generalization with the Information Bottleneck. In *Proc. of the 2010 IEEE Information Theory Workshop (ITW)*.
3. Chechik, G., Globerson, A., Tishby, N., Weiss, Y. (2005). Information bottleneck for Gaussian variables. *Journal of Machine Learning Research*, 6:165–188.
4. Strouse, D.J., Schwab, D.J. (2017). The Deterministic Information Bottleneck. *Neural Computation*, 29(6):1611–1630. DOI: 10.1162/NECO_a_00961.

 doi:10.20944/preprints202505.0746.v1

5. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K. (2017). Deep Variational Information Bottleneck. *Proc. of the International Conference on Learning Representations (ICLR)*.
6. Kolchinsky, A., Tracey, B.D., Wolpert, D.H. (2019). Nonlinear Information Bottleneck. *Entropy*, 21(12):1181. DOI: 10.3390/e21121181.
7. Kolchinsky, A., Tracey, B.D., Van Kuyk, S. (2019). Caveats for Information Bottleneck in deterministic scenarios. *Proc. of the International Conference on Learning Representations (ICLR)*.
8. Rodríguez Gálvez, B., Thobaben, R., Skoglund, M. (2020). The Convex Information Bottleneck Lagrangian. *Entropy*, 22(1):98. DOI: 10.3390/e22010098.
9. Wu, T., Fischer, I., Chuang, I.L., Tegmark, M. (2019). Learnability for the Information Bottleneck. *Entropy*, 21(10):924. DOI: 10.3390/e21100924.
10. Belghazi, M.I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, D. (2018). MINE: Mutual Information Neural Estimation. *Proc. of the International Conference on Machine Learning (ICML)*.
11. Achille, A., Soatto, S. (2018). Emergence of Invariance and Disentanglement in Deep Representations. *Journal of Machine Learning Research*, 19(54):1–34.
12. Zhao, Z., Liu, Y., Peng, X., Li, Y., Liu, T., Tao, D. (2024). Exploring Complex Trade-offs in Information Bottleneck through Multi-Objective Optimization. *arXiv preprint arXiv:2310.00789*.
13. Liu, Y., Zhao, Z., Peng, X., Liu, T., Tao, D. (2023). Exploring the Trade-Off in the Variational Information Bottleneck for Regression with a Single Training Run. *Entropy*, 25(7):1043. DOI: 10.3390/e25071043. (Corrected volume/issue for 2023, assuming typical Entropy numbering. Original was 26(12):1043 which is unlikely for 2023).
14. Cover, T.M., Thomas, J.A. (2012). *Elements of Information Theory*. Wiley, 2nd Ed.
15. Witsenhausen, H.S., Wyner, A.D. (1975). A conditional entropy bound for a pair of discrete random variables. *IEEE Transactions on Information Theory*, 21(5):493–501.
16. Shwartz-Ziv, R., Tishby, N. (2017). Opening the Black Box of Deep Neural Networks via Information. *arXiv preprint arXiv:1703.00810*.