BOOTSTRAPPED INFORMATION BOTTLENECK (BIB)

Anonymous authors
Paper under double-blind review

ABSTRACT

The Information Bottleneck (IB) balances prediction sufficiency and input compression. IB methods rely on data-dependent estimates of these two quantities and typically select a trade-off parameter from single-point estimates of these quantities without taking into account estimation uncertainty. We introduce the Bootstrapped Information Bottleneck (BIB), which provides *a distribution over prediction sufficiency vs. compression* summarized with confidence ellipses. We provide the theoretical foundation for BIB by proving *asymptotic consistency of the bootstrap for mutual information estimators*. BIB ensures that sufficiency constraints are met with high probability on unseen data and *improves held-out test accuracy by* $> 7\%$ *compared to point estimate baselines on ImageNet*. Finally, we show experimentally that the BIB data-dependent confidence regions are *three orders of magnitude tighter than existing high-probability bounds on mutual information*.

1 INTRODUCTION

The notion of *information relevance* has been studied in statistics, machine learning, and information theory under the umbrella of sufficient statistics (Fisher, 1922), representation/feature learning, and information bottleneck (IB) (Tishby et al., 2000) over several decades. Under a parametric model of the data, a sufficient statistic contains all the information about unknown parameters that the data provide. The Pitman–Koopman–Darmois theorem establishes that classical sufficient statistics of a bounded dimension, independent of sample size, exist exclusively for exponential families of distributions (Koopman, 1936). Building on the connection between statistical sufficiency and mutual information (Cover, 1999), the IB framework has been proposed as an approximation of minimal sufficient statistics to explain how relevant information is extracted for output prediction (Tishby et al., 2000). IB has proven to be a useful tool in deep learning, tackling prediction tasks while promoting robustness to adversarial attacks (Alemi et al., 2017) and the disentanglement of latent representations (Achille and Soatto, 2018; Burgess et al., 2018; Wang et al., 2024a).

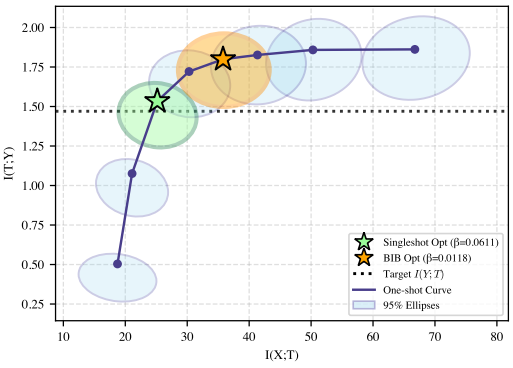


Figure 1: Empirical information curve for VIB on MNIST. Solid line shows point estimates; ellipses capture bootstrap uncertainty using 95% contours. The yellow star denotes the BIB-optimized selection, while the green star denotes the optimal point-estimate selection.

However, a significant challenge in standard IB is the estimation of mutual information (MI). In deterministic networks with strictly monotone nonlinearities, MI-measures are technically vacuous, as they are either infinite or a constant that does not depend on the network's parameters (Goldfeld et al., 2018; Goldfeld and Polyanskiy, 2020). To overcome this challenge, frameworks such as the deep *variational IB* (VIB) framework (Alemi et al., 2017) introduce stochasticity into the network representation where mutual information measures become meaningful, parameter-dependent quantities. A variety of other neural estimators have been deployed to address the IB objective. These range from traditional KDE-based approaches (Kolchinsky et al., 2019) and the Nguyen–Wainwright–Jordan (NWJ) lower bound (Nguyen et al., 2010) to more contemporary neural architectures such as MINE

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(Belghazi et al., 2018), Nonlinear IB (NIB) (Kolchinsky et al., 2019), and InfoNCE (Oord et al., 2018).

Despite the IB framework's origin as an approximation to minimal sufficient statistics, current implementations often fail to satisfy this theoretical promise. Specifically, standard IB methods lack statistical guarantees that the learned representation achieves predictive sufficiency, and the mutual information estimates used to calibrate the prediction-sufficiency trade-off are inaccurate. Estimating mutual information in high dimensions is notoriously difficult, yielding estimators that suffer from substantial bias and variance. Since the direction of the bias in established variational IB frameworks is known (lower bound on compression and upper bound on sufficiency (Alemi et al., 2017; Fischer, 2020; Kolchinsky et al., 2019)), the primary challenge in empirical IB optimization with variational sufficiency and compression estimates is the high variance in finite-sample regimes. Current IB methods solve the IB problem without accounting for the estimator variance. This motivates us to utilize the nonparametric bootstrap (Efron, 1979; Efron and Tibshirani, 1994) to quantify this finite-sample variance and use it to solve the IB problem with rigorous statistical guarantees. While recent work has empirically applied the bootstrap to derive confidence intervals for mutual information estimation (Gritsai et al., 2025), it focuses primarily on practical evaluation and lacks formal theoretical justification for the bootstrap's validity in this setting.

1.1 Summary of contributions

- Algorithmically, we propose BIB, an IB framework that replaces unstable point estimates of mutual information with confidence ellipses in the information plane. BIB enables a statistically rigorous optimization of the IB objective, ensuring the selected model achieves the desired trade-off with *high confidence*.
- Theoretically, we establish that BIB asymptotically satisfies the IB objective with at least 95% confidence. This is shown for both discrete plug-estimators of MI and a broad class of variational neural estimators, including VIB (Alemi et al., 2017), CEB (Fischer, 2020), and NIB (Kolchinsky et al., 2019).
- Empirically, we demonstrate on ImageNet-scale experiments that implementing the IB objective using point estimates of mutual information can be unreliable, as they frequently overestimate the amount of task-relevant information captured by the model (see Figure 1). In contrast, BIB provides a calibrated compression-sufficiency trade-off that successfully meets user-defined sufficiency constraints on unseen data and *improves held-out test accuracy by* $> 7\%$ compared to point-estimate optimization.

2 Information bottleneck method

Let $(X, Y) \sim \mathbb{P}$ and let T be a representation of X satisfying the Markov condition $T - X - Y$, i.e., $p(t \mid x, y) = p(t \mid x)$. The IB objective is to find a representation that is sufficient for predicting Y while being minimal in terms of superfluous information measured by mutual information with X. This is formally expressed using the Shannon mutual information function, $I(\cdot, \cdot)$, as

$$\min_{p(t|x)} I(X;T) \;\text{ s.t. }\; I(Y;T) \geq \alpha \tag{IB Obj}$$

for some user-specified target sufficiency level α that satisfies $0 \leq \alpha \leq I(X;Y)$. Different levels of α yield different trade-offs between compression and sufficiency, and the collection of solutions for all $0 < \alpha \leq I(X;Y)$ trace out the "information curve", i.e., the achievable pairs $(I(X;T), I(Y;T))$, where moving along the curve increases $I(Y;T)$ at the cost of increasing $I(X;T)$. Since analytical solutions to Equation (IB Obj) are intractable, Tishby et al. (2000) proposed the Lagrangian learning objective

$$\min_{p(t|x)} \mathcal{L}_{IB}(\beta) := \min_{p(t|x)} -I(Y;T) \;+\; \beta I(X;T) \tag{IB Lag}$$

where the Lagrangian multiplier $0 < \beta \leq 1$ balances sufficiency for prediction and compression level: if β is close to zero, the objective emphasizes sufficiency, whereas larger values of β encourage compression. This allows for a full exploration of all possible IB trade-offs by varying the β-parameter in Equation (IB Lag). Notably, Wang et al. (2024b) show that when the IB curve is strictly concave, there always exists a bijective mapping from β to the target sufficiency α. However,

it is in general not possible to determine a priori which value of β will yield a given target $I(Y;T)$ level α, and in realistic settings, we select β by training models over a grid of candidate values and choosing the β that achieves the desired compression-sufficiency trade-off, see Figure 1.

2.1 Optimizing the IB Lagrangian

For any fixed β, a practical procedure to optimize Equation (IB Lag) is needed. In their original formulation of the IB principle, Tishby et al. (2000) derived an analytical solution in terms of self-consistent equations for an encoder and decoder distribution. However, these equations have to be solved using an expensive iterative algorithm, which becomes intractable for high-dimensional data. In the current literature, most IB algorithms rely on estimating an upper bound on $I(X;T)$ and a lower bound on $I(Y;T)$, leading to minimization of an upper bound on Equation (IB Lag).

2.1.1 Estimating $I(Y;T)$

In classification, $I(Y;T)$ is commonly estimated by minimizing cross-entropy loss, as is done in most variational IB implementations Alemi et al. (2017); Fischer (2020); Kolchinsky et al. (2019). Introducing a parametric variational decoder $q_\theta(y|t)$ and the Shannon entropy function H, we have

$$I(Y;T) := H(Y) - H(Y \mid T),$$
$$H(Y \mid T) := \mathbb{E}_{p(y|t)}[-\log p(y \mid t)] \leq \mathbb{E}_{p(y|t)}[-\log q_\theta(y \mid t)].$$

Since $H(Y)$ is constant with respect to the model parameters θ, minimizing the cross-entropy loss over θ corresponds to maximizing a variational lower bound on $I(Y;T)$.[1] Given data $\{(y_n, t_n)\}_{n=1}^N$, we can thus estimate

$$\widehat{I}_\theta(Y;T) = \widehat{H}(Y) + \frac{1}{N}\sum_{n=1}^{N} \log q_\theta(y_n \mid t_n)$$

where $\widehat{H}(Y)$ is the empirical entropy of the labels.

2.1.2 Estimating $I(X;T)$

To solve Equation (IB Lag), we minimize an upper bound on $I(X;T)$. Alemi et al. (2017) introduced the *Variational Information Bottleneck* (VIB), which, given the encoder $p_\phi(t|x)$, estimates $I(X;T)$ using the variational bound $\widehat{I}_\phi(X;T) := D_{\mathrm{KL}}(p_\phi(t|x); r(t))$, where r is a chosen prior distribution over the latent representation space, e.g., a Gaussian. Since the true marginal $p(t) = \int p_\phi(t|x)p(x)dx$ is inaccessible, the prior $r(t)$ serves as a variational substitute. In the nonlinear information bottleneck (NIB) (Kolchinsky et al., 2019), $I(X;T)$ is upper bounded by $I(X;T) \leq -\frac{1}{N}\sum_{i=1}^{N} \log \left[\frac{1}{N}\sum_{j=1}^{N} \exp(-D_{\mathrm{KL}}(p(t|x_i); p(t|x_j)))\right]$, given $x_1, \ldots, x_N$ drawn i.i.d. from $p(x)$. This formulation estimates the marginal entropy of the latent space, bypassing the need for a fixed prior. A third approach, the Conditional Entropy Bottleneck (CEB) (Fischer, 2020), establishes a related upper bound by conditioning the bottleneck on the target variable ($I(X;T|Y)$), replacing the marginal prior with a target-conditioned variational marginal $e(t|y)$ to tightly control the compressed representation. Other variational estimators such as MINE Belghazi et al. (2018) and InfoNCE Oord et al. (2018) construct *lower bounds* on $I(X;T)$, making them unsuitable for the IB objective. Therefore, we restrict our focus to VIB, NIB, and CEB that rigorously bound $I(X;T)$ from above.

2.1.3 Empirical optimization and the information curve

To evaluate the Information Bottleneck trade-off, we partition a total dataset of N samples into two disjoint sets: a training set $\mathcal{S}_{\mathrm{tr}}$ of size N_{tr} and a validation set $\mathcal{S}_{\mathrm{val}}$ of size N_{val}, such that $N = N_{\mathrm{tr}} + N_{\mathrm{val}}$. We define the IB loss function over the training data $\mathcal{S}_{\mathrm{tr}} = \{(x_n, y_n)\}_{n=1}^{N_{\mathrm{tr}}}$ as:

$$\mathcal{L}_{\mathrm{IB}}(\theta, \phi; \beta) = \frac{1}{N_{\mathrm{tr}}}\sum_{n=1}^{N_{\mathrm{tr}}} -\log q_\theta(y_n|t_n) + \beta\, \widehat{I}_\psi(X;T) \tag{1}$$

[1] The inequality uses $\mathrm{KL}(p(y \mid t) \| q(y \mid t)) \geq 0$, implying $\mathbb{E}_{p(y,t)}[-\log p(y \mid t)] \leq \mathbb{E}_{p(y,t)}[-\log q(y \mid t)]$.

where $t_n \sim p_\phi(t|x_n)$, and $\widehat{I}_\psi(X;T)$ is a variational upper bound on $I(X;T)$ based on a neural critic f_ψ. To find the optimal β, we optimize Equation (1) K times using the chosen grid $\beta_1, ..., \beta_K$, we compute $(\widehat{I}_\phi(X;T_{\beta_1}), \widehat{I}_\theta(Y;T_{\beta_1})), \ldots, (\widehat{I}_\phi(X;T_{\beta_K}), \widehat{I}_\theta(Y;T_{\beta_K}))$ using the $\mathcal{S}_{\text{val}}$, and we select the β which satisfies our desired compression-sufficiency trade-off. Note that the variational bounds on the true MI hold in expectation: $\mathbb{E}[\hat{I}_\psi(X;T)] \geq I(X;T)$, $\mathbb{E}[\hat{I}_\theta(Y;T)] \leq I(Y;T)$. However, for finite-sample estimates, there is no guarantee for the IB objective, because $\hat{I}_\psi(X;T) \geq I(X;T)$ and $\hat{I}_\theta(Y;T) \leq I(Y;T)$ are not guaranteed to hold. To guarantee that the true mutual information satisfies the (IB Obj), it is sufficient to ensure that our biased, conservative estimate of sufficiency meets the target threshold, provided we accurately account for its estimation variance. This motivates a framework for estimating the *variance* in the variational MI estimates.

3 BIB AND UNCERTAINTY QUANTIFICATION

In this section, we present the Bootstrapped Information Bottleneck (BIB). We first overview the theoretical foundations of the bootstrap for empirical processes and establish BIB asymptotically recovers the true mutual information with plug-in estimates.

3.1 BIB THEORY

3.1.1 BOOTSTRAP CONSISTENCY FOR EMPIRICAL PROCESSES

Let $\{(X_n, Y_n)\}_{n=1}^N$ be sampled i.i.d. from an unknown joint distribution $\mathbb{P}$ on a measurable space $\mathcal{X} \times \mathcal{Y}$, and let $\{(X_n^*, Y_n^*)\}_{n=1}^N$ be drawn uniformly with replacement from the original sample. The empirical measure and its bootstrap analogue are defined as

$$\mathbb{P}_N \;=\; \frac{1}{N}\sum_{n=1}^{N} \delta_{(X_n,Y_n)}, \;\; \mathbb{P}_N^* \;=\; \frac{1}{N}\sum_{n=1}^{N} \delta_{(X_n^*,Y_n^*)}.$$

Define the empirical process and its bootstrap analogue by

$$\mathbb{G}_N := \sqrt{N}\,(\mathbb{P}_N - \mathbb{P}), \;\; \mathbb{G}_N^* := \sqrt{N}\,(\mathbb{P}_N^* - \mathbb{P}_N).$$

$\mathbb{G}_N$ and $\mathbb{G}_N^*$ can be considered random elements in the space $\ell^\infty(\mathcal{F})$ of bounded functions indexed by a class $\mathcal{F}$ of measurable functions. A function class $\mathcal{F}$ is said to be *Donsker* if the empirical process $\{\mathbb{G}_N(f) : f \in \mathcal{F}\}$ converges in probability to a Brownian bridge $\mathbb{G}$ in the space $\ell^\infty(\mathcal{F})$ (van der Vaart and Wellner, 1996). Essentially, this condition ensures that the collection of functions is not "too large", allowing for a functional version of the Central Limit Theorem to hold uniformly over $\mathcal{F}$ (Dudley, 1999). When $\mathcal{F}$ is Donsker, the bootstrap empirical process $\mathbb{G}_N^*$ also converges in probability to the same limit $\mathbb{G}$ in probability (van der Vaart and Wellner, 1996, Theorem 3.6.1). Specifically, this implies that

$$\sup_{h \in \text{BL}_1} \left| \mathbb{E}[h(\mathbb{G}_N^*)] - \mathbb{E}[h(\mathbb{G})] \right| \xrightarrow{\Pr} 0 \tag{2}$$

where BL_1 is the set of all real-valued functions on $\ell^\infty(\mathcal{F})$ with a Lipschitz constant and supremum norm both bounded by 1. In statistical terms, this means that $\sqrt{N}(\mathbb{P}_N^* - \mathbb{P}_N)$ provides an asymptotically consistent approximation to the sampling distribution of $\sqrt{N}(\mathbb{P}_N - \mathbb{P})$. This fact becomes particularly useful when combined with the (functional) delta method, which transfers bootstrap consistency from $\mathbb{P}_N$ to a broad class of smooth plug-in statistics $T(\mathbb{P}_N)$.

3.1.2 FUNCTIONAL DELTA METHOD AND THE BOOTSTRAP

Let $T(\mathbb{P}_N)$ denote an estimator expressed as a functional of the empirical measure. If T is sufficiently smooth in the sense of Hadamard differentiability, we can apply the Delta method to show that the bootstrap gives an asymptotically consistent approximation of the distribution of $T(\mathbb{P}_N)$.

Definition 3.1 (Hadamard Differentiability)**.** Let $\mathbb{D}$ and $\mathbb{E}$ be normed spaces. A map $T : \mathbb{D}_T \subset \mathbb{D} \to \mathbb{E}$ is called *Hadamard differentiable* at $\theta \in \mathbb{D}_T$ tangentially to a subspace $\mathbb{D}_0 \subset \mathbb{D}$ if there exists a

continuous linear map $T'_\theta : \mathbb{D}_0 \to \mathbb{E}$ such that

$$\frac{T(\theta + t_n h_n) - T(\theta)}{t_n} \to T'_\theta(h) \quad \text{as } n \to \infty \tag{3}$$

for all sequences $t_n \to 0$ and $h_n \to h$ such that $h \in \mathbb{D}_0$ and $\theta + t_n h_n \in \mathbb{D}_T$ for all n.

Theorem 3.2 (The Delta-Method for the Bootstrap (van der Vaart and Wellner, 1996, Theorem 3.9.11))**.** *Let $\mathbb{D}$ and $\mathbb{E}$ be normed spaces. Let $T : \mathbb{D}_T \subset \mathbb{D} \mapsto \mathbb{E}$ be Hadamard-differentiable at $\mathbb{P}$ tangentially to a subspace $\mathbb{D}_0$. Let $\mathbb{P}_N$ and $\mathbb{P}^*_N$ be the empirical measure and its bootstrap analogue, respectively, taking values in $\mathbb{D}_T$ such that $\mathbb{G}_N$ and $\mathbb{G}^*_N$ converge in distribution to the same Brownian bridge $\mathbb{G}$ which takes its values in $\mathbb{D}_0$. Then,*

$$\sup_{h \in BL_1} \left| \mathbb{E}\, h\left(\sqrt{N}(T(\mathbb{P}^*_N) - T(\mathbb{P}_N))\right) - \mathbb{E}\, h\left(\sqrt{N}(T(\mathbb{P}_N) - T(\mathbb{P}))\right) \right| \xrightarrow{\Pr} 0.$$

Given the smoothness conditions of T in Definition 3.1, Theorem 3.2 tells us that $\sqrt{N}(T(\mathbb{P}^*_N) - T(\mathbb{P}_N))$ can be used to form asymptotically valid standard errors and confidence intervals for $T(\mathbb{P})$. We now use this theoretical framework for the estimation of mutual information.

3.1.3 BIB THEORY: DISCRETE PLUG-IN MI ESTIMATORS

We first specialize to the discrete case where $\{(X_n, Y_n)\}_{n=1}^N$ takes values in $[r] \times [s]$ and $\mathbb{P} = \{p_{ij}\}_{i=1,\dots,r;\, j=1,\dots,s}$. The empirical measure can be written as $\mathbb{P}_N = \{\hat{p}_{ij}\}_{i,j}$ with $\hat{p}_{ij} = N_{ij}/N$ and $N_{ij} = \sum_{k=1}^N \mathbf{1}\{(X_k, Y_k) = (i,j)\}$, and similarly $\mathbb{P}^*_N = \{\hat{p}^*_{ij}\}_{i,j}$ with $\hat{p}^*_{ij} = N^*_{ij}/N$. Define mutual information as the functional $I(\mathbb{P}) := \sum_{i=1}^r \sum_{j=1}^s p_{ij} \log\left(\frac{p_{ij}}{p_{i\cdot} p_{\cdot j}}\right)$, $p_{i\cdot} = \sum_{j=1}^s p_{ij}$, $p_{\cdot j} = \sum_{i=1}^r p_{ij}$. The discrete plug-in estimator is $I(\mathbb{P}_N)$ and its bootstrap version is $I(\mathbb{P}^*_N)$.

Theorem 3.3 (Bootstrap Validity for Discrete Plug-in Mutual Information)**.** *Let $\mathcal{S}_N = \{(X_n, Y_n)\}_{n=1}^N$ be i.i.d. observations from $\mathbb{P} = \{p_{ij}\}_{i=1,\dots,r;\, j=1,\dots,s}$, and assume $p_{ij} > 0$ for all i, j. Let $\mathbb{P}_N$ denote the empirical measure and $\mathbb{P}^*_N$ denote its bootstrap version. Then $I(\mathbb{P})$ is Hadamard differentiable at $\mathbb{P}$. Consequently,*

$$\Pr\Big(\sqrt{N}\big(I(\mathbb{P}^*_N) - I(\mathbb{P}_N)\big) \le t \,\Big|\, \mathcal{S}_N\Big) \;-\; \Pr\Big(\sqrt{N}\big(I(\mathbb{P}_N) - I(\mathbb{P})\big) \le t\Big) \;\xrightarrow{Pr}\; 0.$$

In particular, bootstrap quantiles of $I(\mathbb{P}^*_N) - I(\mathbb{P}_N)$ yield asymptotically valid confidence intervals for $I(\mathbb{P})$.

Proof outline. The validity of the bootstrap for the discrete mutual information estimator is established in two steps: first, by verifying that the finite-dimensional empirical process is Donsker, and second, by demonstrating the Hadamard differentiability of the plug-in mutual information estimator with respect to the joint probability distribution. The result then follows from the functional delta method for the bootstrap (Theorem 3.2). The full proof is in Appendix A.1.

3.1.4 BIB THEORY: NEURAL ESTIMATORS FOR MI

We now consider a generic and variational IB setup with a parametric and learnable encoder $p_\phi(t|x)$ and decoder $q_\theta(y|t)$ as described in Section 2. This setup covers methods like VIB, NIB, and CEB, which utilize neural network components to evaluate a variational upper bound on $I(X;T)$. *The following theorem holds for a generalized class of such neural estimators.*

Theorem 3.4 (Bootstrap validity for neural IB objectives)**.** *Let $\mathcal{S}_N = \{(X_n, Y_n)\}_{n=1}^N$ be i.i.d. from an unknown joint distribution $\mathbb{P}$ on the compact sample space $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \{1, \dots, M\}$ and $P(Y = y) > 0$ for all y.*

First, consider the compression bound $\hat{I}_\psi(X;T) = \mathcal{T}(\mathbb{P}_N f_\psi(x,t))$, where $\mathcal{T}$ is Hadamard differentiable. Define the function class $\mathcal{F}_\psi = \{(x,t) \mapsto f_\psi(x,t) : \psi \in \Psi\}$. Assume $\mathcal{F}_\psi$ is a class of neural networks indexed by parameters $\psi \in \Psi$, where Ψ is a compact parameter space with bounded weights $\|\psi\|_2 \le W$ and L-Lipschitz activations. Then, for a fixed parameter ψ, $\mathcal{F}_\psi$ is Donsker. Consequently, the distribution of the bootstrapped estimate $\hat{I}^*_\psi(X;T) = \mathcal{T}(\mathbb{P}^*_N f_\psi(x,t))$ converges in distribution to the sampling distribution:

$$\sup_{t \in \mathbb{R}} \Big| \Pr\Big(\sqrt{N}(\hat{I}^*_\psi(X;T) - \hat{I}_\psi(X;T)) \le t \,\Big|\, \mathcal{S}_N\Big) - \Pr\Big(\sqrt{N}(\hat{I}_\psi(X;T) - \mathbb{E}[\hat{I}_\psi(X;T)]) \le t\Big) \Big| \xrightarrow{\Pr} 0.$$

Similarly, consider the sufficiency bound $\hat{I}_\theta(Y;T) = \mathbb{P}_N \frac{q_\theta(y|t)}{\hat{p}(y)}$*, and define the function class* $\mathcal{F}_\theta = \{(y,t) \mapsto \log \frac{q_\theta(y|t)}{p(y)} : \theta \in \Theta\}$*. Assuming* Θ *is a compact parameter space, for a fixed parameter* θ*,* $\mathcal{F}_\theta$ *is Donsker. The distribution of the bootstrapped estimate* $\hat{I}^*_\theta(Y;T) = \mathbb{P}^*_N \frac{q_\theta(y|t)}{\hat{p}(y)}$ *satisfies:*

$$\sup_{t\in\mathbb{R}} \Big| \Pr\Big(\sqrt{N}(\hat{I}^*_\theta(Y;T) - \hat{I}_\theta(Y;T)) \le t \,\Big|\, \mathcal{S}_N\Big) - \Pr\Big(\sqrt{N}(\hat{I}_\theta(Y;T) - \mathbb{E}[\hat{I}_\theta(Y;T)]) \le t\Big) \Big| \xrightarrow{\Pr} 0.$$

In particular, the bootstrap quantiles yield asymptotically valid confidence intervals for the true expected variational bounds $\mathbb{E}[\hat{I}_\psi(X;T)]$ *and* $\mathbb{E}[\hat{I}_\theta(Y;T)]$*.*

Proof outline. To establish the asymptotic validity of the bootstrap for the VIB objective, we verify that the variational function classes are Donsker by satisfying the requirements of Theorem 2.5.2 in van der Vaart and Wellner (1996). Specifically, we prove that $\mathcal{F}_\psi$ and $\mathcal{F}_\theta$ have square-integrable envelope functions and finite uniform entropy integrals, followed by an application of the Delta Theorem. The full proof is in Appendix A.2.

Corollary 3.5 (Confidence limits)**.** *Let* $\mathcal{E}_\beta$ *be the* $1-\delta$ *bootstrap confidence region for the estimated variational bounds. Define the empirical limits* $\hat{I}_{\min}(Y;T_\beta) = \min_{z\in\mathcal{E}_\beta} z_Y$ *and* $\hat{I}_{\max}(X;T_\beta) = \max_{z\in\mathcal{E}_\beta} z_X$*. Under the conditions of Theorem 3.4, the following inequalities hold asymptotically with probability at least* $1-\delta$*:*

$$\hat{I}_{\min}(Y;T_\beta) \le \mathbb{E}[\hat{I}(Y;T_\beta)] \quad \textit{and} \quad \hat{I}_{\max}(X;T_\beta) \ge \mathbb{E}[\hat{I}(X;T_\beta)].$$

The proof of Corollary 3.5 is found in Appendix A.3.

3.2 BIB METHODOLOGY AND ALGORITHMIC OUTLINE

We propose BIB, a framework that solves the IB problem with statistical guarantees. By incorporating the bootstrap into the standard IB pipeline, we guarantee that Equation (IB Obj) is solved using a finite dataset with statistical confidence. Let $\mathcal{E}_\beta$ denote the 95% bootstrap confidence region in the $(I(X;T), I(Y;T))$-plane for a model trained with trade-off parameter β. From this region, we extract the lower bound on sufficiency $\hat{I}_{\min}(Y;T_\beta) = \min_{z\in\mathcal{E}_\beta} z_Y$, and the upper bound on compression, $\hat{I}_{\max}(X;T_\beta) = \max_{z\in\mathcal{E}_\beta} z_X$. As highlighted in Section 2.1.3, the variational MI estimators are designed so that $\mathbb{E}[\hat{I}(Y;T_\beta)] \le I(Y;T_\beta)$ and $\mathbb{E}[\hat{I}(X;T_\beta)] \ge I(X;T_\beta)$. Using Theorem 3.4 and Corollary 3.5, the following chains of inequalities hold with 95% confidence:

$$\begin{aligned} \hat{I}_{\min}(Y;T_\beta) \le \mathbb{E}[\hat{I}(Y;T_\beta)] \le I(Y;T_\beta), \\ \hat{I}_{\max}(X;T_\beta) \ge \mathbb{E}[\hat{I}(X;T_\beta)] \ge I(X;T_\beta). \end{aligned} \tag{4}$$

These bounds are visualized in Figure 3. Since both $\hat{I}_{\min}(Y;T_\beta)$ and $\hat{I}_{\max}(X;T_\beta)$ are tractable, we can solve Equation (IB Obj) by selecting β to minimize $\hat{I}_{\max}(X;T_\beta)$ among all candidate models that strictly satisfy the sufficiency constraint $\hat{I}_{\min}(Y;T_\beta) \ge \alpha$. As summarized in Algorithm 1 (Appendix B), the full BIB procedure first trains K candidate models on a grid $\beta_1, ..., \beta_K$ using the training set. We then estimate sufficiency and compression on B bootstrapped datasets, drawn i.i.d. with replacement from a validation set, to construct the confidence ellipse $\mathcal{E}_\beta$ for each model. Crucially, the computational overhead of BIB is minimal as the procedure only requires resampling of the validation set and performing forward passes through the pre-trained models, avoiding the prohibitive cost of re-training for each bootstrap iteration.

Remark 3.6. BIB focuses exclusively on *finite-sample variance* and not training variance. By treating the trained candidate models as fixed, we isolate the dominant source of unreliability during practical model selection: the finite-sample estimation noise of the MI bounds on the validation set. This ensures robust candidate selection with minimal computational overhead.

4 EXPERIMENTS

In this section, we empirically evaluate the utility and statistical validity of the BIB. First, we validate the coverage properties of the BIB confidence regions on high-dimensional data against ground-truth

targets. Next, we demonstrate the practical value of BIB on MNIST and ImageNet by showing that it leads to a robust IB calibration, satisfying the desired trade-off of sufficiency and compression on unseen data. Finally, we provide uncertainty analysis for popular IB objectives (VIB, NIB, and CEB). A comparison between BIB and previously proposed uncertainty quantification bounds for IB proposed by Shamir et al. (2010) is provided in Appendix D. For completeness, we also show that the BIB machinery also guarantees prediction sufficiency for IB models using MINE Belghazi et al. (2018) and InfoNCE Oord et al. (2018) although similar guarantees for compression $I(X;T)$ cannot be established for these methods. An analysis of the computational overhead and timing results is provided in Appendix G and Appendix H, respectively.

4.1 Empirical coverage of variational bounds

Setup. To validate that BIB confidence regions accurately capture the true variational bounds, we evaluate empirical coverage on a continuous, high-dimensional dataset ($d = 1000$) where the true generative process is known. To emulate the complexity of real-world data, we construct a multi-class Gaussian Mixture Model (GMM) with spatial decay covariance structures ($\Sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.85$). The synthetic setup allows us to compute the ground-truth mutual information quantities $(I(X;T), I(Y;T))$ with high precision via massive-sample Monte Carlo integration ($N_{\text{MC}} = 100{,}000$).

We evaluate VIB, NIB, and CEB. For each method and value of β, we train the model and construct 95% BIB confidence ellipses using $B = 200$ resamples of a finite validation set. We then evaluate the joint coverage, i.e., the frequency with which the ground-truth variational bounds fall within the estimated empirical ellipse, over 100 independent training trials. Crucially, we are measuring the coverage of the *true bounds* rather than the exact true mutual information. However, because the IB objective optimizes a conservative upper bound on compression and a lower bound on sufficiency, achieving nominal coverage of these bounds guarantees that the true mutual information satisfies the user's constraints with at least this stated probability. The coverage analysis of discrete plug-in estimators and comparison with high-probability bounds in Shamir et al. (2010) are done in Appendix D.

Table 1: Empirical joint coverage and area of 95% BIB confidence ellipses across variational estimators. The table reports the empirical joint coverage and mean area of the bootstrap confidence ellipses for VIB, NIB, and CEB across varying trade-off parameters β. Experiments were conducted on a synthetic, high-dimensional Gaussian Mixture Model. Ground-truth targets were approximated via Monte Carlo integration on a massive held-out population set ($N = 100,000$). The BIB confidence regions were constructed using $B = 200$ resamples on a standard validation set ($N = 500$). Joint coverage denotes the proportion of 100 independent trials where the Monte Carlo ground-truth point successfully fell within the estimated 95% confidence region.

	VIB		**NIB**		**CEB**	
β	Joint Coverage	Mean Area	Joint Coverage	Mean Area	Joint Coverage	Mean Area
0.60	0.95	$2.5 \cdot 10^{-3}$	0.98	$5.0 \cdot 10^{-4}$	0.95	$2.8 \cdot 10^{-3}$
0.90	0.96	$2.7 \cdot 10^{-3}$	0.97	$5.0 \cdot 10^{-4}$	0.94	$2.9 \cdot 10^{-3}$
0.99	0.98	$2.7 \cdot 10^{-3}$	0.96	$5.0 \cdot 10^{-4}$	0.96	$3.0 \cdot 10^{-3}$

Results. As shown in Table 1, the empirical joint coverage of the BIB confidence ellipses tightly aligns with the nominal 95% target level across all evaluated variational estimators and compression scales (β). Furthermore, the mean area of the confidence regions remains remarkably small. Unlike worst-case, distribution-free bounds that often yield vacuous intervals in high dimensions, BIB dynamically adapts to the empirical geometry and variance of the specific neural representation being evaluated. These results demonstrate that BIB provides tight, statistically sound uncertainty quantification for modern variational IB objectives, ensuring that the selected compression-sufficiency trade-off will generalize to unseen data.

4.2 Robust IB calibration and improved held-out test accuracy on MNIST and ImageNet

Setup. We evaluate robust calibration on MNIST and ImageNet by training VIB models over a candidate grid of β values. For the ImageNet experiments, we utilize features extracted from a pre-trained ResNet-50; to our knowledge, this work represents the first demonstration of information plane results for these popular variational estimates using a ResNet-50 architecture on ImageNet. We analyze how selection via deterministic point estimates versus BIB ellipses affects the ability to meet sufficiency constraints ($I(Y;T) \geq \alpha$) on 20 held-out test sets. The choice of the threshold α is grounded in the "elbow" phenomenon common in information curves: the point where the representation maintains near-maximal sufficiency while achieving significant compression by discarding redundant features. For MNIST, $\alpha \approx 1.55$; for ImageNet, $\alpha \approx 3.12$, both values are 70% of the empirical label entropy. See Appendix I for dataset-specific hyperparameters and grid details. We use $N_{\text{tr}} = 3600$, $N_{\text{val}} = 400$, and 20 independent experiments each with $N_{\text{test}} = 400$. We choose a grid of β values between 10^{-4} and 10^{-1} to fully explore the information curve. See Appendix I for further implementation details.

Table 2: Robustness comparison: Point-estimate selection vs. BIB. Robustness (**Rob.**) denotes the percent of 20 held-out test sets meeting the sufficiency target α. Accuracies (**Acc.**) are averaged over test sets.

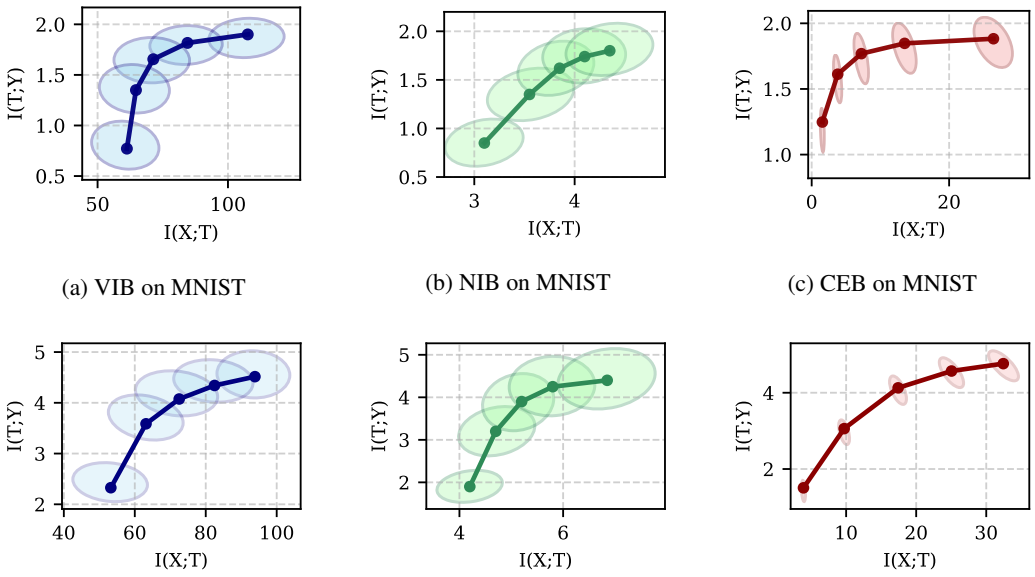
Dataset	**Selection**	$\boldsymbol{\beta}$	**Rob.** (%)	**Acc.** (%)
MNIST	Point	0.0720	75%	83.05%
	BIB	0.0148	**95%**	89.70%
ImageNet	Point	0.0794	70%	57.50%
	BIB	0.0200	**100%**	**65.20%**

Results and discussion. As shown in Table 2, point-estimate selection is consistently over-optimistic, frequently failing the sufficiency constraint on unseen data due to estimator variance. In contrast, BIB guards against over-compression by selecting an accurate β based on the 95% confidence lower bound. On ImageNet, achieves 100% robustness, meaning all 20 held-out test sets successfully satisfy the sufficiency threshold α, while the point-estimate baseline only achieves 70%. This also leads to 7.7% increased accuracy on a held-out test set. These results show that accounting for estimator uncertainty is fundamental to realizing the users' desired sufficiency level and maintaining predictive performance in both small-scale and large-scale regimes. This is thoroughly illustrated in Figure 1: although a point on the IB curve can sit above the target sufficiency, the corresponding 95% bootstrap ellipse may overlap with the region below the threshold, indicating a possibility that the model fails to satisfy the constraint. Qualitatively, the increased sufficiency gained from robustness translates to superior class disentanglement in the latent space of the encoder. We provide a visual comparison of these representations in Appendix F. We also note that BIB introduces a minimal computational overhead of only $5-11\%$ (see Appendix H) and scales efficiently to large datasets such as ImageNet, as its post-hoc design avoids costly re-training. We run similar robustness experiments and demonstrate similar findings for NIB and CEB in Appendix E.

4.3 Uncertainty quantification of popular IB methods

Setup and results. We evaluate the stability of various IB objectives, specifically VIB, NIB, and CEB, on MNIST and the large-scale ImageNet dataset. For each baseline, we sweep the Lagrangian multiplier β to trace both the single-estimate information curve and the 95% confidence ellipse, providing a clear comparison of how different IB objectives behave under identical data conditions.

We showcase the uncertainty and estimator stability for each IB approach in Figure 2. It is clear that all approaches suffer from estimator variance; thus, BIB is a useful tool for both simple and complex learning tasks, regardless of the IB objective used. The choice of objective affects both the variance of $I(X;T)$ and $I(Y;T)$. Because each method constructs its variational upper bound on $I(X;T)$ using distinct mathematical formulations, the compression estimates operate on different absolute scales along the x-axis. Yet, there is substantial finite-sample variance relative to the respective bounds



(a) VIB on MNIST (b) NIB on MNIST (c) CEB on MNIST

(d) VIB on ImageNet (e) NIB on ImageNet (f) CEB on ImageNet

Figure 2: Information plane analysis and estimator sensitivity on MNIST and ImageNet. Each panel displays the information curve (solid lines) across varying β values alongside 95% confidence bootstrap ellipses (shaded regions). The ellipses illustrate the empirical variance of the mutual information estimates generated from $B = 200$ resamples of the validation set. Panels compare three distinct Information Bottleneck frameworks: VIB, NIB, and CEB.

across all methods and both datasets. This observation is consistent with the findings in Poole et al. (2019), which notice that variational MI estimators inherently suffer from high variance.

Summary of sensitivity and ablation results. Finally, we investigate how estimator uncertainty is influenced by dataset size (N) and representation dimensionality ($\dim(T)$). As shown in Figure 6, the bootstrap ellipses generally shrink as N increases, reflecting the expected reduction in estimator variance and the growing reliability of point estimates. Furthermore, we find that a large architecture tends to increase estimator uncertainty. A comprehensive analysis of these variables across different IB objectives, including detailed tables and visualizations, is provided in Appendix G.

5 Limitations

The BIB framework provides a robust method for selecting IB trade-offs. As the sample size N increases, the bootstrap variance shrinks toward zero, rendering point estimates more relevant for model selection in this case. Hence, BIB is of particular practical value for low to medium data regimes and large architectures. It would be interesting to go beyond the non-parametric bootstrap for medium-sized data at the cost of additional compute, which we leave for future work.

6 Conclusion

We have introduced BIB, a simple and efficient procedure that augments the standard IB pipeline with uncertainty quantification, which is applicable to most popular variational IB frameworks. By replacing unreliable point estimates with confidence ellipses in the information plane, BIB provides a robust IB calibration that accounts for estimator variance. Theoretically, we guarantee that the BIB satisfies the IB objective asymptotically with 95% confidence. Empirically, we demonstrated that the BIB consistently achieves any desired compression-sufficiency trade-off on unseen data, guaranteeing better generalization and predictive accuracy.

AI USE STATEMENT

All ideas and the proposed method, theoretical claims and proofs, experimental designs, and final text are developed and verified by the authors without AI tools.

The authors utilized AI tools as a writing aid to help improve the clarity of the manuscript and to assist with code generation for the experiments.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our theoretical and empirical contributions, we provide the following supplementary material in the appendices: complete proofs for all theoretical claims, detailing the assumptions made for each claim (Appendix A), a complete step-by-step algorithm for the BIB (Appendix B), and full implementation details for all numerical experiments, including information on the data, hyperparameter configurations, and model architectures (Appendix I). The code for reproducing all experiments will be made available upon acceptance.

REFERENCES

Ronald A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, 1922.

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

Bernard Osgood Koopman. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical society*, 39(3):399–409, 1936.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017.

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Chenyu Wang, Sharut Gupta, Xinyi Zhang, Sana Tonekaboni, Stefanie Jegelka, Tommi Jaakkola, and Caroline Uhler. An information criterion for controlled disentanglement of multimodal data. *arXiv preprint arXiv:2410.23996*, 2024a.

Ziv Goldfeld, Ewout van den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks. *arXiv preprint arXiv:1810.05728*, 2018.

Ziv Goldfeld and Yury Polyanskiy. The information bottleneck problem and its applications in machine learning. *IEEE Journal on Selected Areas in Information Theory*, 1(1):19–38, 2020.

Artemy Kolchinsky, Brendan D Tracey, and David H Wolpert. Nonlinear information bottleneck. *Entropy*, 21(12):1181, 2019.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R. Devon Hjelm. Mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ian Fischer. The conditional entropy bottleneck. *Entropy*, 22(9):999, 2020.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1994.

German Gritsai, Megan Richards, Maxime Méloux, Kyunghyun Cho, and Maxime Peyrard. MIST: Mutual information via supervised training. *arXiv preprint arXiv:2511.18945*, 2025.

Chenyu Wang, Sharut Gupta, Xinyi Zhang, Sana Tonekaboni, Stefanie Jegelka, Tommi Jaakkola, and Caroline Uhler. An information criterion for controlled disentanglement of multimodal data. *arXiv preprint arXiv:2410.23996*, 2024b.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer, New York, 1996.

Richard M. Dudley. *Uniform Central Limit Theorems*, volume 63. Cambridge University Press, Cambridge, UK, 1999. ISBN 978-0-521-46102-3.

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

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning (ICML)*, 2019.

TABLE OF CONTENTS

1 Introduction **1**

1.1 Summary of contributions 2

2 Information bottleneck method **2**

2.1 Optimizing the IB Lagrangian 3

3 BIB and uncertainty quantification **4**

3.1 BIB theory 4

3.2 BIB methodology and algorithmic outline 6

4 Experiments **6**

4.1 Empirical coverage of variational bounds 7

4.2 Robust IB calibration and improved held-out test accuracy on MNIST and ImageNet 8

4.3 Uncertainty quantification of popular IB methods 8

5 Limitations **9**

6 Conclusion **9**

A Proofs **14**

A.1 Proof of Theorem 3.3 14

A.2 Proof of Theorem 3.4 15

A.3 Proof of Corollary 3.5 16

B BIB Algorithm **16**

C Plug-in estimator **16**

C.1 Comparison to high probability bounds 16

D Additional plots **18**

E Extended Robustness Results across Estimators **19**

F Representation characteristics **19**

G Ablation Study **20**

G.1 Setup 20

G.2 Results 20

H Computational overhead **21**

I Implementation Details **22**

I.1 Comparison to High-Probability Bounds 22
I.2 Using BIB for IB Calibration 23
I.3 Behaviors of Different IB Methods on MNIST and ImageNet 23
I.4 Ablation Study 23

A PROOFS

A.1 PROOF OF THEOREM 3.3

Proof. To establish the asymptotic validity of the bootstrap for the discrete plug-in mutual information estimator, we must verify the two conditions required by Theorem 3.2: the Donsker property of the underlying empirical process and the Hadamard differentiability of the functional I. First, note that the empirical measure $\mathbb{P}_N$ and its bootstrap version $\mathbb{P}_N^*$ are indexed by the finite class of indicator functions

$$\mathcal{F} = \{\mathbf{1}\{(x, y) = (i, j)\} : 1 \le i \le r, 1 \le j \le s\}.$$

Since any finite class of measurable functions with a finite envelope is a Donsker class, the empirical process $\mathbb{G}_N = \sqrt{N}(\mathbb{P}_N - \mathbb{P})$ converges weakly to a Brownian bridge $\mathbb{G}$ in $\ell^\infty(\mathcal{F})$. By Theorem 3.6.1 in van der Vaart and Wellner (1996), $\mathbb{G}_N^* = \sqrt{N}(\mathbb{P}_N^* - \mathbb{P}_N)$ also converges to $\mathbb{G}$ in $\ell^\infty(\mathcal{F})$. Next, we will investigate the conditions of the Delta method for the bootstrap, i.e., Hadamard differentiability for the plug-in MI estimator (Theorem 3.2). Let

$$\mathcal{P}_{r,s} := \Big\{\mathbb{P} = \{p_{ij}\}_{i,j} : p_{ij} > 0,\ \sum_{i=1}^{r}\sum_{j=1}^{s} p_{ij} = 1\Big\} \subset \mathbb{R}^{rs},$$

and let a small perturbation $h = \{h_{ij}\}_{i,j} \in \mathbb{R}^{rs}$ with defined marginals $h_{i\cdot} = \sum_{j=1}^{s} h_{ij},\ h_{\cdot j} = \sum_{i=1}^{r} h_{ij}$ satisfy

$$\sum_{i=1}^{r}\sum_{j=1}^{s} h_{ij} = 0,$$

so that $\mathbb{P} + th \in \mathcal{P}_{r,s}$ for all sufficiently small t. The mutual information functional can be written as

$$I(\mathbb{P}) = \sum_{i=1}^{r}\sum_{j=1}^{s} p_{ij} \log p_{ij} - \sum_{i=1}^{r} p_{i\cdot} \log p_{i\cdot} - \sum_{j=1}^{s} p_{\cdot j} \log p_{\cdot j}.$$

Let $\phi(x) = x \log x$, which is continuously differentiable on $(0, \infty)$ with derivative $\phi'(x) = 1 + \log x$. Since $p_{ij} > 0$ for all i, j, all arguments of ϕ lie in the interior of its domain. Let $\mathbb{P}_t = \mathbb{P} + th$. A first-order Taylor expansion of ϕ yields

$$\phi(p_{ij} + t\, h_{ij}) = \phi(p_{ij}) + t\, h_{ij}(1 + \log p_{ij}) + o(t),$$

uniformly over i, j. Summing over indices, we obtain

$$\begin{aligned} I(\mathbb{P}_t) - I(\mathbb{P}) = t \sum_{i=1}^{r}\sum_{j=1}^{s} h_{ij}(1 + \log p_{ij}) \\ - t \sum_{i=1}^{r} h_{i\cdot}(1 + \log p_{i\cdot}) - t \sum_{j=1}^{s} h_{\cdot j}(1 + \log p_{\cdot j}) + o(t). \end{aligned}$$

Using $\sum_{i,j} h_{ij} = 0$, the constant terms cancel, and we conclude that

$$\begin{aligned} \dot{I}_{\mathbb{P}}(h) &:= \lim_{t \to 0} \frac{I(\mathbb{P} + th) - I(\mathbb{P})}{t} \\ &= \sum_{i=1}^{r}\sum_{j=1}^{s} h_{ij} \log\left(\frac{p_{ij}}{p_{i\cdot} p_{\cdot j}}\right). \end{aligned}$$

This derivative is linear and continuous in h, and hence $I(\cdot)$ is Hadamard differentiable at $\mathbb{P}$. Hence, by (Theorem 3.2),

$$\sup_{t \in \mathbb{R}} \left| \Pr\Big(\sqrt{N}\big(I(\mathbb{P}_N^*) - I(\mathbb{P}_N)\big) \le t \,\Big|\, \mathcal{S}_N\Big) - \Pr\Big(\sqrt{N}\big(I(\mathbb{P}_N) - I(\mathbb{P})\big) \le t\Big) \right| \xrightarrow{Pr} 0.$$

□

A.2 Proof of Theorem 3.4

Proof. We establish that the classes $\mathcal{F}_\theta$ and $\mathcal{F}_\psi$ are Donsker by verifying the two conditions of Theorem 2.5.2 in van der Vaart and Wellner (1996): the existence of a square-integrable envelope and a finite uniform entropy integral.

Existence of square-integrable envelopes. By assumption, the sample spaces $\mathcal{X}$ is compact, the parameter spaces Θ and Ψ are compact (with bounded weights $\|\psi\|_2 \leq W$), and the activation functions are L-Lipschitz. Since $\mathcal{T}$ is the space of mapped variables from $\mathcal{X}$, where the mapping is decomposed into bounded mappings by the Lipschitz property and bounded weights, $\mathcal{T}$ is compact as well. Since neural networks are continuous functions of both their inputs and their parameters, every $f_\psi \in \mathcal{F}_\psi$ is bounded on the compact domain $\mathcal{X} \times \mathcal{T}$. Specifically, there exists a constant C_1 such that $|f_\psi(x,t)| \leq C$ for all $(x,t) \in \mathcal{X} \times \mathcal{T}$ and $\psi \in \Psi$. We define the envelope function as C_1, which is square-integrable.

A similar argument applies to $\mathcal{F}_\theta$. The decoder $q_\theta(y|t)$ uses a softmax output layer which bounds

$$q_\theta(y \mid t) = \frac{\exp(a_{\theta,y}(t))}{\sum_{j=1}^{M} \exp(a_{\theta,j}(t))},$$

where $y \in \{1, ..., M\}$. By the compactness of $\mathcal{T}$ and the assumption of Lipschitz activations and boundedness of Θ, $|a_{\theta,j}(t)| \leq a$ for some $a < \infty$ for all t, θ. This implies that $b_0 \leq q_\theta(y|t) \leq b_1$ for some $0 < b_0 \leq b_1$, which again implies that $\left|\log \frac{q_\theta(y|t)}{p(y)}\right| \leq C_2$ for some $C_2 < \infty$ for all y, t, θ. Thus, C_2 serves as a square-integrable envelope for $\mathcal{F}_\theta$.

Uniform Entropy Condition. Next, we show that $\mathcal{F}_\psi$ and $\mathcal{F}_\theta$ satisfy the uniform entropy condition

$$\int_0^\infty \sup_Q \sqrt{\log \mathcal{N}\big(\varepsilon, \mathcal{F}, L_2(Q)\big)} d\varepsilon < \infty,$$

where the supremum is taken over all finitely discrete probability measures Q. The covering number, denoted $\mathcal{N}(\varepsilon, \mathcal{F}, L_2(Q))$, represents the number of balls of radius ε needed to cover the class $\mathcal{F}$ under the $L_2(Q)$ metric. The $L_2(Q)$ metric measures the distance between two functions $f, g \in \mathcal{F}$ as $\sqrt{\int |f-g|^2 dQ}$. We consider the covering number of $\mathcal{F}_\psi$. Since $\Psi \subset \mathbb{R}^w$ is a compact parameter space with bounded weights $\|\psi\|_2 \leq W$, $\mathcal{N}(\varepsilon, \Psi, \|\cdot\|_2) \leq (C/\varepsilon)^w$ for some constant C. Under the assumptions of L-Lipschitz activations and bounded weights acting on the compact domains $\mathcal{X}$ and $\mathcal{T}$, the mapping from the parameter space to the function class $\psi \mapsto f_\psi$ is Lipschitz continuous. Specifically, there exists a constant K such that for any $\psi_1, \psi_2 \in \Psi$:

$$\sup_{x,t} |f_{\psi_1}(x,t) - f_{\psi_2}(x,t)| \leq K \|\psi_1 - \psi_2\|_2.$$

Because an ε/K-covering of the parameter space Ψ induces an ε-cover of the function class $\mathcal{F}_\psi$ in the $L:\infty$ norm, and therefore in the $L_2(Q)$ norm for any measure Q, we obtain the following bound on the uniform covering number:

$$\sup_Q \mathcal{N}(\varepsilon, \mathcal{F}_\psi, L_2(Q)) \leq \mathcal{N}(\varepsilon/K, \Psi, \|\cdot\|_2) \leq \left(\frac{CK}{\varepsilon}\right)^w.$$

We then verify the uniform entropy condition by evaluating the uniform entropy condition:

$$\int_0^1 \sup_Q \sqrt{\log \mathcal{N}(\varepsilon, \mathcal{F}_\psi, L_2(Q))}\, d\varepsilon \leq \int_0^1 \sqrt{w \log\left(\frac{CK}{\varepsilon}\right)}\, d\varepsilon = \sqrt{w} \int_0^1 \sqrt{\log \frac{CK}{\varepsilon}}\, d\varepsilon.$$

Since $\sqrt{\log(1/\varepsilon)}$ is integrable at the origin, this integral is finite, confirming that $\mathcal{F}_\psi$ is a Donsker class. An entirely analogous argument applies to the decoder function class $\mathcal{F}_\theta$, where the compactness of Θ and the Lipschitz nature of the softmax-coupled network ensure a finite uniform entropy integral.

Delta method. Finally, we apply the functional delta method to establish bootstrap consistency. Since $\hat{I}_\psi(X;T)$ is a Hadamard differentiable transformation $\mathcal{T}$ of the empirical mean of f_ψ, Theorem 3.2

guarantees that the bootstrap empirical process $\sqrt{N}(\hat{I}^*_\psi(X;T) - \hat{I}_\psi(X;T))$ asymptotically recovers the sampling distribution of $\sqrt{N}(\hat{I}_\psi(X;T) - \mathbb{E}[\hat{I}_\psi(X;T)])$, and similarly for the sufficiency bound $\hat{I}_\theta(Y;T)$ to recover the distribution of $\sqrt{N}(\hat{I}_\theta(Y;T) - \mathbb{E}[\hat{I}_\theta(Y;T)])$.

□

A.3 Proof of Corollary 3.5

Proof. By Theorem 3.4, the bootstrap confidence region $\mathcal{E}_\beta$ captures the true expected variational bounds $(\mathbb{E}[\hat{I}_\psi(X;T_\beta)], \mathbb{E}[\hat{I}_\theta(Y;T_\beta)])$ with probability approaching $1-\delta$ as $N \to \infty$. Therefore, whenever the true expected bounds fall inside $\mathcal{E}_\beta$, they are bounded by $\hat{I}_{\min}(Y;T_\beta)$ and $\hat{I}_{\max}(X;T_\beta)$. □

B BIB Algorithm

Algorithm 1 Selecting the optimal β using nonparametric bootstrapped confidence regions

Input: Train set $\mathcal{S}_{\text{tr}}$, validation set $\mathcal{S}_{\text{val}}$, grid $[\beta_1, \ldots, \beta_K]$, and target sufficiency α
Create bootstrapped sets $\mathcal{S}^*_1, \ldots, \mathcal{S}^*_M$ using $\mathcal{S}_{\text{val}}$
for $k = 1$ **to** K **(in parallel) do**
  Train IB model T_{β_k} **using** $\mathcal{S}_{\text{tr}}$
  if k = 1 then
    Set $T_{\beta^*} = T_{\beta_k}$
  end if
  for $m = 1$ **to** M **(in parallel) do**
    Compute $\hat{I}_{\mathcal{S}^*_m}(X;T_{\beta_k})$ **and** $\hat{I}_{\mathcal{S}^*_m}(Y;T_{\beta_k})$
  end for
  Compute the bootstrap mean $\boldsymbol{\mu}_k = \frac{1}{M}\sum_{m=1}^M \mathbf{z}_{k,m}$ **with** $\mathbf{z}_{k,m} = (\hat{I}_{\mathcal{S}^*_m}(X;T_{\beta_k}), \hat{I}_{\mathcal{S}^*_m}(Y;T_{\beta_k}))$
  Compute the bootstrap covariance matrix $\boldsymbol{\Sigma}_k = \frac{1}{M-1}\sum_{m=1}^M (\mathbf{z}_{k,m} - \boldsymbol{\mu}_k)(\mathbf{z}_{k,m} - \boldsymbol{\mu}_k)^\top$
  Compute squared Mahalanobis distances for all samples: $d^2_{k,m} = (\mathbf{z}_{k,m} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{z}_{k,m} - \boldsymbol{\mu}_k)$
  Compute 95% bootstrap ellipse: $\mathcal{E}_k = \{\, \mathbf{z} : (\mathbf{z} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{z} - \boldsymbol{\mu}_k) \le d^2_{k,(\lceil 0.95M \rceil)} \,\}$
  Compute a lower bound on sufficiency: $\hat{I}_{\min}(Y;T_{\beta_k}) = \min_{\mathbf{z} \in \mathcal{E}_k} z_Y$
  Compute an upper bound on compression: $\hat{I}_{\max}(X;T_{\beta_k}) = \max_{\mathbf{z} \in \mathcal{E}_k} z_X$
  if $\hat{I}_{\min}(Y;T_{\beta_k}) \ge \alpha$ **and** $\hat{I}_{\max}(X;T_{\beta_k}) \le \hat{I}_{\max}(X;T_{\beta^*})$ **then**
    $T_{\beta^*} \leftarrow T_{\beta_k}$
  end if
end for
return T_{β^*}

C Plug-in estimator

C.1 Comparison to high probability bounds

Setup. To evaluate our bootstrap confidence intervals against a baseline, we compare them to the finite-sample analysis of the information bottleneck done in Shamir et al. (2010).

Theorem C.1 (Shamir et al. (2010))**.** *Let T be determined by a fixed and known conditional probability distribution $p(t|x)$. Let $\mathcal{S}_N$ be a sample of size N drawn from the joint probability distribution $\mathbb{P}(x,y)$. For any confidence parameter $\delta \in (0,1)$, it holds with a probability of at least $1-\delta$ over the sample $\mathcal{S}_N$ that:* $|I(X;T) - \hat{I}_S(X;T)| \le (|\mathcal{T}|\log(N) + \log(|\mathcal{T}|))\sqrt{\log(4/\delta)}/\sqrt{2/N} + (|\mathcal{T}|-1)/N$ *and* $|I(Y;T) - \hat{I}_S(Y;T)| \le (3|\mathcal{T}|+2)\log(m)\sqrt{\log(4/\delta)}/\sqrt{2N} + \big((|\mathcal{Y}|+1)(|\mathcal{T}|+1) - 4\big)/N$ *where $\hat{I}_{\mathcal{S}_N}$ is the plug-in estimate of mutual information based on $\mathcal{S}_N$.*

Since the high-probability bounds of Theorem C.1 hold with probability of at least $1 - \delta$, we can interpret them as confidence intervals (or rectangles in the $I(X;T), I(Y;T)$-plane) around the plug-in estimates. We compare them to the $1 - \delta$ bootstrap confidence intervals of BIB by empirically measuring the width and coverage each provides. Note, however, that these two confidence interval approaches offer different guarantees: the bounds of Shamir et al. (2010) hold with a probability of *at least* $1 - \delta$ for any underlying data distribution $\mathbb{P}(x, y)$, whereas the bootstrap confidence intervals are *data-dependent*, and only provide *asymptotic* $1 - \delta$ coverage of the true mutual information according to Theorem 3.3. Thus, we should expect the high-probability bounds to yield more conservative intervals, and that the bootstrap ellipses are generally tighter in practice. Note also that the high-probability bounds only apply to the plug-in mutual information estimator in discrete data settings, whereas the bootstrap confidence intervals give guarantees in more general settings. In our experiments, we set up an IB setting with discrete, simulated data with known ground truth mutual information, and we optimize the IB Lagrangian the Blahut-Arimoto algorithm (Tishby et al., 2000). We then compute the width and empirical coverage of both the $1 - \delta$ high-probability bounds and the BIB confidence ellipses over multiple independent trials.

Table 3: Joint uncertainty regions in the information plane across values of β. Joint coverage requires the true pair $(I(X;T), I(Y;T))$ to be inside the ellipse, while Union coverage requires at least one marginal estimate to fall within its respective bound. Configuration: $n = 200$, $|\mathcal{T}| = 2$, $|\mathcal{X}| = |\mathcal{Y}| = 3$, $\delta = 0.05$, $B_{\text{BIB}} = 200$, $R_{\text{reps}} = 20000$.

	Bootstrap ellipse			Shamir rectangle	
β	Joint Coverage	Union Coverage	Mean area	Joint Coverage	Area
0.60	0.9319	1.0000	7.52×10^{-3}	1.0000	21.25
0.90	0.9488	1.0000	1.08×10^{-2}	1.0000	21.25
0.99	0.9471	1.0000	1.25×10^{-2}	1.0000	21.25

Results. In Table 3, we observe, as expected, that for every β-level, the bootstrap confidence ellipses have a smaller area, and a coverage of about 93% to 95%, while the more conservative high-probability bounds achieve a 100% coverage. On average, the confidence regions induced by the high-probability bounds occupy only 0.03% of the area covered by the high-probability bounds, reflecting the fundamentally different nature of the two uncertainty quantification approaches. Although the high-probability bounds provide strong, distribution-independent guarantees, they can be overly conservative in finite-sample regimes. By adapting to the geometry of the empirical information plane, BIB allows for much tighter joint uncertainty regions.

D ADDITIONAL PLOTS

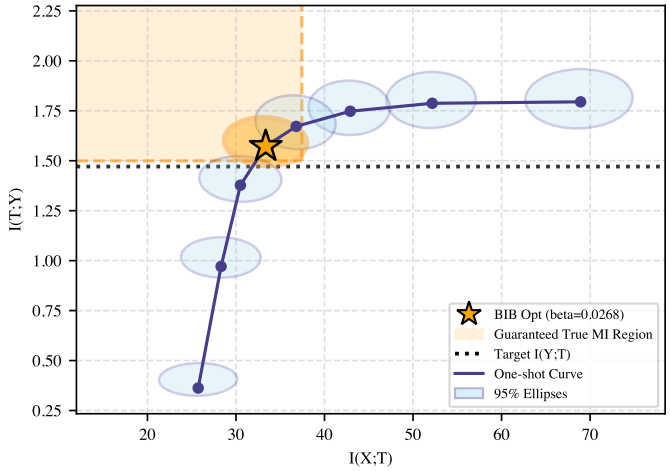
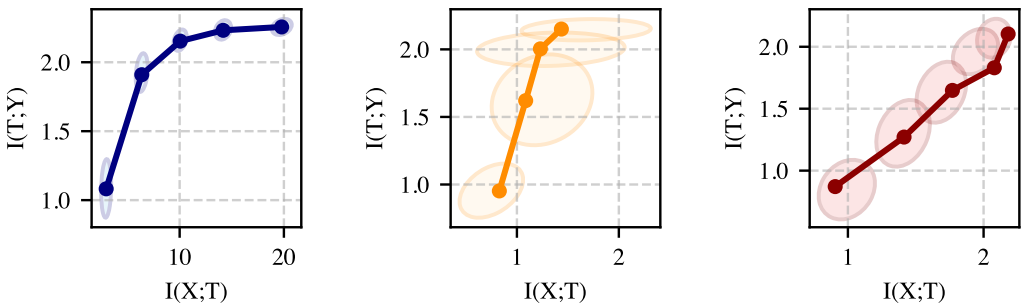


Figure 3: Information plane trajectory highlighting the BIB-selected optimal model. The shaded orange region contains the true $I(X;T), I(Y;T)$ with 95% probability. As established in Corollary 3.5, since $(\mathbb{E}[\hat{I}(X;T)], \mathbb{E}[\hat{I}(Y;T)])$ lies within the confidence ellipse with 95% confidence, the true $(I(X;T), I(Y;T))$ consequently lies in the top-left quadrant w.r.t. the ellipse's worst-case extremes with 95% confidence.



(a) VIB (b) MINE (c) InfoNCE

Figure 4: Information plane trajectories and 95% BIB confidence ellipses for VIB, MINE, and InfoNCE on CIFAR-10. Since MINE and InfoNCE produce lower bounds on $I(X;T)$, they do not give guarantees on compression and thus not on the IB objective. This noticeably degrades the IB optimization, as the plots for MINE and InfoNCE do not trace out a convex information curve when varying β. However, BIB still gives 95% guarantees on the prediction sufficiency in these cases, because $\mathbb{E}[\hat{I}(Y;T)]$ remains a valid lower bound across all three architectures.

E Extended Robustness Results across Estimators

Table 4: Robustness comparison across mutual information estimators: NIB, MINE, and InfoNCE. We compare β selection via single-point estimates (β^*_{point}) vs. BIB (β^*_{BIB}). Success is defined as meeting the sufficiency constraint $I(Y;T) \geq \alpha$ on 20 held-out datasets.

NIB Estimator

Dataset	Selection	β^*	Rob. (%)	Acc. (%)
MNIST	Point	0.724	70.0%	62.60%
	BIB	0.433	90.0%	77.00%
ImageNet	Point	0.750	65.0%	54.20%
	BIB	0.410	95.0%	62.80%

MINE Estimator

Dataset	Selection	β^*	Rob. (%)	Acc. (%)
MNIST	Point	1.584	60.0%	67.20%
	BIB	0.631	100.0%	79.50%
ImageNet	Point	1.620	55.0%	56.80%
	BIB	0.615	100.0%	68.40%

InfoNCE Estimator

Dataset	Selection	β^*	Rob. (%)	Acc. (%)
MNIST	Point	0.857	85.0%	67.10%
	BIB	0.464	100.0%	76.60%
ImageNet	Point	0.880	75.0%	58.10%
	BIB	0.445	100.0%	69.20%

Table 4 shows that across all mutual information estimators, the BIB selection method is significantly more reliable in satisfying sufficiency constraints on unseen data compared to standard single-point estimates. While point-based selection often fails to meet the target sufficiency α on held-out test sets, BIB consistently achieves robustness of 90% to 100%. The incorporation of uncertainty via bootstrap confidence regions successfully identifies a safety margin in the trade-off parameter β, leading to models that generalize their performance and maintain higher classification accuracy.

The utility of BIB is especially pronounced in the large-scale ImageNet regime, where the increased variance of neural mutual information estimators (such as MINE and InfoNCE) makes point estimates highly over-optimistic. By incorporating uncertainty via bootstrap confidence regions, BIB successfully identifies a conservative safety margin in the trade-off parameter β. This leads to models that not only meet target sufficiency constraints with high probability but also maintain significantly higher classification accuracy on unseen data.

F Representation characteristics

To understand the practical impact of robust calibration, we investigate the qualitative differences between representations selected via deterministic point estimates versus those selected by the BIB framework. Standard point-estimate selection is highly sensitive to the inherent variance of mutual information estimators, often choosing a trade-off parameter β that appears optimal on a specific validation set but fails to generalize. This typically results in over-compression, where important information about the labels ($I(Y;T)$) is removed. In contrast, BIB selects a compression-sufficiency trade-off that is robust to estimator noise by utilizing the 95% confidence lower bound of the information estimates. As shown in Figure 5, this approach ensures the model maintains a robust manifold that preserves relevant information.

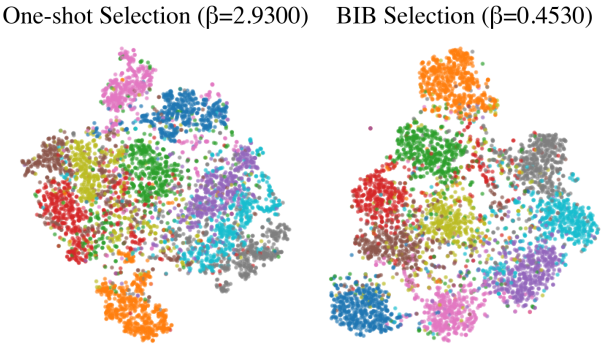


Figure 5: Latent space visualization (t-SNE) for VIB models on MNIST with a target sufficiency $I(Y;T) \geq 0.35$. The one-shot selection ($\beta = 2.9300$) suffers from significant cluster overlap due to over-compression. The BIB selection ($\beta = 0.4530$) identifies a more conservative trade-off that achieves clearer cluster boundaries and superior class disentanglement.

As seen in Figure 5, representations selected by BIB exhibit stable cluster geometry with clearer boundaries between classes. The point-estimate selection demonstrates significant cluster overlap, indicating that the over-optimistic choice of β has led to the loss of discriminative features. Through this visualization, we confirm that the safety margin provided by BIB confidence regions fundamentally improves the quality and separability of the learned representations.

G Ablation Study

G.1 Setup

We want to study how our bootstrap analysis is affected by: the size of the training and validation datasets, the number of model parameters, and the Lagrangian parameter β. Specifically, we investigate whether the reliability gains of BIB are consistent across different scales of data and model complexity. IN addition to MNIST, experiment on the medium-complexity learning task CIFAR-10, extracting publicly available pretrained checkpoints of ResNet-18, provided by `torchvision.models` subpackage, which is part of the official PyTorch project. We expect both the dataset size and model capacity to significantly influence the variance of mutual information estimates, as higher-dimensional representations often require more samples to achieve stable density estimation. To isolate the effect of a single variable, we vary it while keeping all other learning parameters fixed. All fixed learning parameters follow the configuration described in Appendix I. We let the total dataset size $N_{\text{tot}} = N_{\text{tr}} + N_{\text{val}}$ vary across $\{1000, 2000, 5000, 10000\}$, we let the dimensionality of the encoder in the VIB vary across $\{8, 32, 128, 512\}$, and we let β vary across $\{0.001, 0.01, 0.1, 0.99\}$.

G.2 Results

Table 5: Comparison of bootstrap ellipse areas ($Area_{95}$) across IB methods ($N = 1000, d_t = 128$). Smaller areas indicate lower estimator variance and higher reliability.

(a) MNIST dataset

Beta (β)	**VIB**	**NIB**	**MINE**	**InfoNCE**
0.001	5.4624	0.1327	2.6064	0.1441
0.010	3.3813	0.1206	2.4085	0.1143
0.100	2.6778	0.0821	1.5294	0.1106
0.990	2.0899	0.0005	1.0267	0.9552

(b) CIFAR-10 dataset

Beta (β)	**VIB**	**NIB**	**MINE**	**InfoNCE**
0.001	1.1197	0.0416	0.2906	0.0400
0.010	0.4935	0.0679	0.2031	0.0746
0.100	0.2793	0.0492	0.2873	0.1651
0.990	0.3519	0.0006	0.2198	0.6606

Table 6: Comparison of bootstrap ellipse areas ($Area_{95}$) across IB methods for varying bottleneck dimensionality ($\beta = 10^{-2}, N = 1000$). Smaller areas indicate lower estimator variance and higher reliability.

(a) MNIST dataset

dim(T)	**VIB**	**NIB**	**MINE**	**InfoNCE**
8	0.1507	0.1324	1.6893	0.1121
32	0.4304	0.1291	2.1434	0.1380
128	2.5971	0.0723	1.5382	0.1181
512	9.5668	0.0032	1.7411	0.1377

(b) CIFAR-10 dataset

dim(T)	**VIB**	**NIB**	**MINE**	**InfoNCE**
8	0.0381	0.0265	0.3302	0.5862
32	0.1102	0.0203	0.3399	0.4488
128	0.5936	0.0010	0.2205	0.3210
512	1.4381	0.0005	0.3288	0.4654

Table 7: Comparison of bootstrap ellipse areas ($Area_{95}$) across IB methods for varying sample sizes ($\beta = 10^{-2}, d_t = 128$). Smaller areas indicate lower estimator variance and higher reliability.

(a) MNIST dataset

N	**VIB**	**NIB**	**MINE**	**InfoNCE**
1000	3.0652	0.0724	1.3468	0.1342
2000	0.5206	0.0680	1.0040	0.0650
5000	0.1065	0.0386	0.3702	0.2608
10000	0.0207	0.0129	0.1802	0.1787

(b) CIFAR-10 dataset

N	**VIB**	**NIB**	**MINE**	**InfoNCE**
1000	0.5471	0.0044	0.2900	0.3134
2000	0.1695	0.012	0.3763	0.3451
5000	0.0163	0.003	0.1781	0.1955
10000	0.0044	0.006	0.1044	0.2504

The results across all three ablation studies confirm that BIB provides a consistent and adaptive measure of uncertainty. In Table 5, we see that the Lagrangian parameter β significantly shifts the uncertainty profile; as β approaches 1 (maximum compression), the bootstrap areas generally decrease as the representation becomes lower-dimensional and less complex.

In Table 6, we observe that VIB shows a sharp increase in estimator variance as the bottleneck dimensionality $\dim(T)$ grows. This highlights a critical use case for BIB: for high-capacity models, point estimates of mutual information become increasingly unstable, making the safety margin provided by our bootstrap ellipses essential for reliable IB calibration. The neural methods MINE and InfoNCE are, however, rather stable across dimensionality, perhaps because these estimators also have their own critic network architectures that are decoupled from the encoder's dimensionality. This effect is visualized in Figure 6.

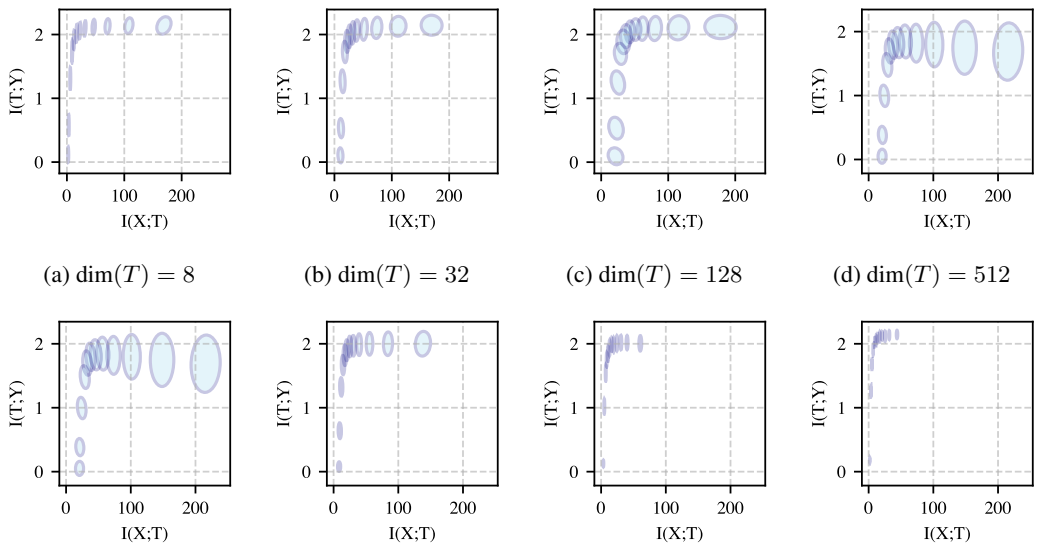
Finally, Table 7 demonstrates the expected asymptotic behavior: as the sample size N increases, the bootstrap ellipse areas shrink across all methods. Notably, even at $N = 10000$, the neural estimators (MINE and InfoNCE) maintain larger uncertainty regions than VIB, suggesting that BIB is particularly valuable for complex, non-parametric information estimators. This effect is visualized in Figure 6.

H Computational overhead

To quantify the computational requirements of BIB, we measured the execution time for the two primary phases of the algorithm using a VIB model on both the MNIST and ImageNet datasets. As shown in Table 8, the bootstrap phase introduces a minimal computational burden relative to the standard training process. As detailed in Table 8, uncertainty quantification via BIB adds only a 5 to 11 percent overhead to the standard training pipeline.

Table 8: Computational cost of VIB training vs. the BIB bootstrap phase ($n = 200$). The overhead represents the additional time required for uncertainty quantification relative to the training phase.

Dataset	**Training Phase (s)**	**Bootstrap Phase (s)**	**Overhead** $\%$
MNIST (100 epochs)	25.87	2.96	11.44 $\%$
ImageNet (50 epochs)	1806.88	100.37	5.50 $\%$



(a) $\dim(T) = 8$ (b) $\dim(T) = 32$ (c) $\dim(T) = 128$ (d) $\dim(T) = 512$

(e) $N = 1000$ (f) $N = 2000$ (g) $N = 5000$ (h) $N = 10000$

Figure 6: Ablation study of the BIB uncertainty regions for VIB on MNIST. The top row (a-d) illustrates the effect of increasing the bottleneck dimensionality $\dim(T)$ from 8 to 512 while keeping $N = 1000$ fixed. The 95% confidence ellipses grow significantly as the capacity increases, reflecting the rising estimator variance in high-dimensional latent spaces. The bottom row (e–h) shows the effect of increasing the total dataset size N from 1,000 to 10,000 with $\dim(T) = 128$. As expected, the uncertainty regions contract as more samples become available, demonstrating the asymptotic consistency of the bootstrap estimates. Each plot displays the information plane $I(X;T)$ vs. $I(Y;T)$ for a sweep of β values, with individual points representing one-shot estimates and ellipses representing the bootstrap-mapped joint uncertainty.

This efficiency is due to the post-hoc nature of BIB. Unlike traditional ensemble methods that require re-training models for each resample, BIB performs its analysis on fixed, pre-trained representations. The "overhead" consists solely of forward passes through the encoder and decoder on resampled validation sets, followed by the calculation of confidence ellipses. Notably, as the scale of the task increases (moving from MNIST to ImageNet), the relative overhead of the bootstrap phase decreases significantly. This confirms that BIB is not only computationally feasible but scales gracefully to high-capacity models and large-scale datasets where traditional uncertainty quantification might be prohibitive.

I Implementation Details

This appendix provides the technical specifications, architectural choices, and hyperparameter configurations for each experiment presented in the main text. All experiments are run on NVIDIA A1000 GPUs.

I.1 Comparison to High-Probability Bounds

We evaluate the BIB and Shamir's bounds using a discrete 3×3 joint distribution, training representations via the Blahut-Arimoto algorithm with a representation cardinality of $|\mathcal{T}| = 2$. For the uncertainty quantification, we use $B = 200$ bootstrap resamples to construct the BIB ellipses and compare them against Shamir rectangles across $R = 20,000$ Monte Carlo repetitions with $n = 200$ samples and $\delta = 0.05$.

I.2 USING BIB FOR IB CALIBRATION

In this experiment, we implement the Deep Variational Information Bottleneck (VIB) on the MNIST and ImageNet datasets using a latent representation dimension of $\dim(T) = 256$. The ImageNet experiments utilize pre-extracted 2048-dimensional features. These features were extracted using the ResNet-50 model provided by the `torchvision.models` subpackage, which is part of the official PyTorch project. The model architecture and its weights are available through the PyTorch Hub and the official GitHub repository (`pytorch/vision`). For our experiments, we utilized pre-trained ResNet-50 weights and extracted the final 2048-dimensional activations from the global average pooling layer to serve as input to our IB framework. The model is trained for 100 epochs using the Adam optimizer with a learning rate of 10^{-4} and a batch size of 64, employing a single Monte Carlo sample for the reparameterization trick. We evaluate the model's performance over a grid of eight β values log-spaced between $10^{-3.0}$ and $10^{-0.5}$. The training set consists of $N_{\text{train}} = 1000$ samples, while the information curves and uncertainty regions are computed on a validation set of $N_{\text{val}} = 400$. For the BIB framework, we generate $B = 200$ bootstrap resamples of the validation set to estimate the joint sampling distribution of the mutual information quantities. These samples are used to construct 95% confidence ellipses based on the Mahalanobis distance. To assess the robustness of our IB calibration, we split the remaining test data into 20 independent held-out subsets of 100 samples each. IB calibration is performed against a target sufficiency threshold of $\alpha = 0.70 \cdot H(Y)$ (approximately 1.55 nats), comparing a single-point selection strategy against the conservative lower-bound criterion provided by the BIB ellipses. The experiments in Appendix E, are set up using the same data, but with different grids over β to explore the information curve for each mutual information estimator.

I.3 BEHAVIORS OF DIFFERENT IB METHODS ON MNIST AND IMAGENET

For the comparative analysis of IB estimators on MNIST and Imagenet, we evaluate the stability of VIB, NIB, MINE, and InfoNCE by visualizing their uncertainty regions in the information plane. For MNIST, we employ a three-layer MLP with 1024 hidden units, while the ImageNet experiments extract features from a ResNet-50 model. All models use a latent dimension of $\dim(T) = 128$ and are trained with the Adam optimizer (learning rate 10^{-4}) for 300 epochs to ensure convergence across the different objectives. We evaluate each method over a log-spaced grid of five β values, using $N_{\text{trn}} = 9,000$ and $N_{\text{val}} = 1000$ for Imagenet. For neural estimators like MINE and InfoNCE, we use a four-layer MLP critic network and perform additional optimization steps on the validation set to ensure tight MI bounds. The BIB uncertainty regions are generated using $B = 200$ bootstrap resamples; for the neural methods, the critic is re-optimized for each resample to capture the variance inherent in the estimation process. The resulting 95% confidence ellipses allow for a direct comparison of how each estimator's specific geometry and training objective influence the precision of the resulting information curve.

I.4 ABLATION STUDY

In our ablation study, we analyze how the reliability of the BIB framework scales across different data regimes, bottleneck capacities, and compression constraints. Using the MNIST and CIFAR-10 datasets, we sweep across total sample sizes $N \in \{1000, 2000, 5000, 10000\}$ with a fixed 90/10 train-validation split and vary the latent dimension $\dim(T) \in \{8, 32, 128, 512\}$. The models are trained for 100 epochs using the Adam optimizer with a learning rate of 10^{-4} and evaluated across a range of trade-off parameters $\beta \in \{0.001, 0.01, 0.1, 0.99\}$. For each configuration, we quantify the estimation uncertainty by calculating the area of the 95% confidence bootstrap ellipse derived from $B = 100$ resamples. For neural estimators like MINE and InfoNCE, we employ a statistics network optimized for 20 additional steps per bootstrap resample to ensure a tight variational lower bound. This setup allows us to rigorously assess how the statistical efficiency of the BIB regions is influenced by the interaction between sample size, model complexity, and the specific information-theoretic objective used.