Advanced Topic A14: *The Information Bottleneck*

Data Science with Python – BSc Advanced Lectures

Joerg Osterrieder

10 Minutes

Motivation: What Should a Representation Learn?

A good representation of inputs should retain prediction-relevant information only

- A neural network maps input X through layers to produce an output $\hat{Y}$
- Too much information about X retained: model memorises noise and overfits
- Too little information retained: model underfits and cannot predict Y well
- **The information bottleneck**: formalises this tradeoff using Shannon mutual information
- Goal: find the representation T that retains *maximal* information about Y with *minimal* information from X

The IB principle (Tishby, Pereira, Bialek 1999) unifies compression and generalisation theory

Shannon Mutual Information: Background

Mutual information quantifies statistical dependence

$$I(X;T) = \sum_{x,t} p(x,t) \log \frac{p(x,t)}{p(x)p(t)} = H(T) - H(T \mid X)$$

- $I(X;T) = 0$: X and T are independent; T contains no information about X
- $I(X;T) = H(X)$: T is a sufficient statistic of X (no compression)
- $I(T;Y)$: how much information T carries about the label Y
- The IB objective trades off $I(X;T)$ (compression) against $I(T;Y)$ (prediction power)
- In neural networks: each layer's activations form a representation T; IB measures its quality

Mutual information is the fundamental measure of dependency; it captures non-linear relationships

The Information Bottleneck Problem

Find the minimal sufficient statistic of X for predicting Y

$$\min_{p(t|x)} \; I(X;T) - \beta\, I(T;Y)$$

- T: the compressed representation; $p(t|x)$ is the stochastic encoder
- $\beta \geq 0$: Lagrange multiplier controlling the compression-prediction tradeoff
- Small β: compress aggressively, even at the cost of prediction accuracy
- Large β: retain full information about Y, allow large $I(X;T)$
- The **IB curve**: the Pareto front of $(I(X;T), I(T;Y))$ over all valid $p(t|x)$

The IB Lagrangian is an information-theoretic analogue of the bias-variance tradeoff

The Data Processing Inequality and Sufficient Statistics

Any processing of T cannot increase its information about X

$$X \to T \to \hat{Y}: \quad I(X;\hat{Y}) \le I(X;T) \le H(X)$$

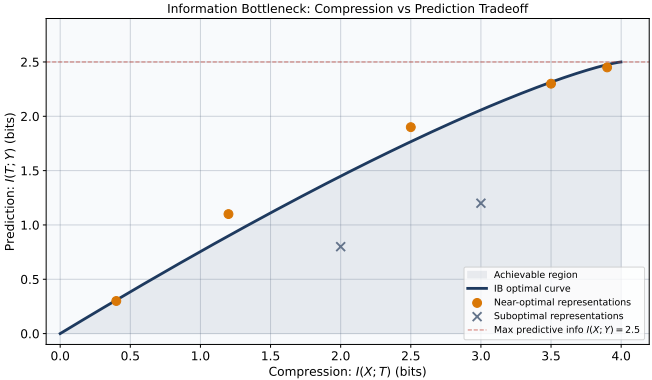
- **Data Processing Inequality (DPI)**: for any Markov chain $X \to T \to \hat{Y}$, processing cannot increase information; every layer is lossy
- **Sufficient statistic**: T is sufficient for Y given X if $I(T;Y) = I(X;Y)$; no information about Y is lost in the compression
- The IB optimum is a minimal sufficient statistic: sufficient (retains all Y-information) and minimal ($I(X;T)$ is as small as possible)
- In neural networks: the DPI implies that deeper layers cannot recover information discarded by shallower layers – architectural choices are irreversible compression steps
- Finance: a factor model that discards a signal early cannot recover it in later risk calculations – feature selection order matters

DPI: information can only decrease through a network; every layer is a one-way compression step

The Compression-Prediction Tradeoff

No representation can simultaneously minimise both objectives

- On the IB curve: $(I(X;T), I(T;Y))$ traces from $(0,0)$ to $(H(X), H(Y|X)^c)$
- Points on the curve are Pareto-optimal; points inside are achievable but suboptimal
- Points outside the curve are unachievable: you cannot get more prediction from less input

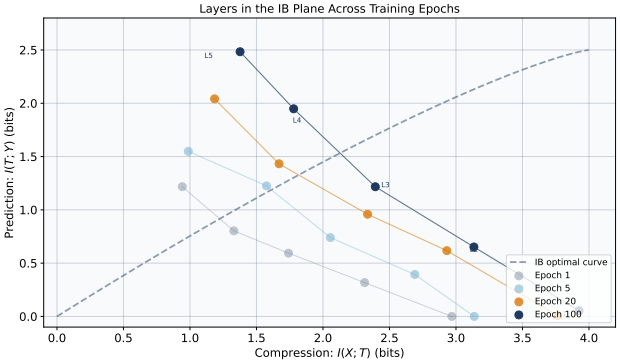


The IB curve is analogous to the efficient frontier in portfolio theory: maximum return for given risk

Neural Network Layers in the IB Plane

Each layer computes a representation; the IB plane tracks layer quality

- Plot each layer of a trained network as a point: $(I(X;h_l), I(h_l;Y))$
- Lower layers are close to $(H(X), I(X;Y))$: retain most input information
- Upper layers (near output): closer to $(0, H(Y))$: compressed and predictive
- **Tishby & Schwartz-Ziv (2017)**: reported that layers move toward the IB curve during training
- Points on the IB curve represent optimal layer-wise representations

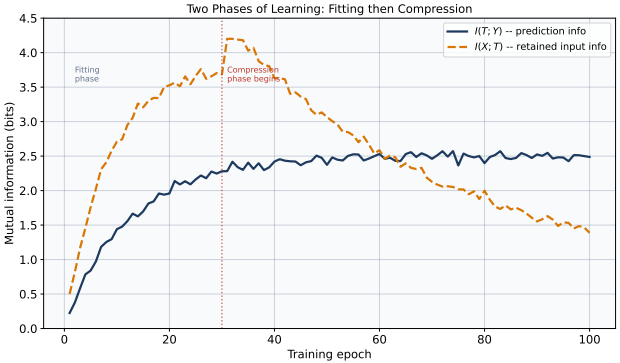


Each layer is a point in the IB plane; moving toward the curve means becoming more efficient

Two Phases of Learning: Fitting and Compression

Early training increases $I(T;Y)$; later training decreases $I(X;T)$

- **Phase 1 (fitting)**: mutual information with Y rises rapidly; the network memorises relevant patterns
- **Phase 2 (compression)**: mutual information with X decreases; the network discards irrelevant noise
- Tishby & Schwartz-Ziv (2017) reported this two-phase dynamics in simple networks



The compression phase is analogous to generalisation: discarding noise that does not help predict Y

Controversy: Does the Compression Phase Exist?

Saxe et al. (2018) challenged the universality of the two-phase finding

- Compression phase appears only for saturating activations (sigmoid/tanh): the mutual information estimator with binned activations artefactually produces the phase
- With ReLU activations: mutual information $I(X;T)$ does not decrease during training
- The compression is an artefact of activation saturation, not a general principle of deep learning
- IB as a *design principle* (Variational IB, see next slide) remains useful; IB as a *description* of neural network training is contested

IB is a powerful design principle; the empirical claim that all networks compress is not universally supported

Variational Information Bottleneck (VIB)

The IB objective as a learnable, end-to-end regulariser

$$\mathcal{L}_{\text{VIB}} = -I(T;Y) + \beta\, I(T;X) \approx -\mathbb{E}[\log q(y|t)] + \beta\, D_{\text{KL}}[p(t|x) \,\|\, r(t)]$$

- Alemi et al. (2017): parameterise $p(t|x) = \mathcal{N}(\mu_\phi(x), \sigma^2_\phi(x))$; use the reparameterisation trick
- The KL term penalises deviation from the prior $r(t) = \mathcal{N}(0, I)$: encourages compression
- VIB is essentially a VAE with a classification head; β controls regularisation strength
- Provides calibrated uncertainty estimates and is provably more robust to adversarial perturbations than cross-entropy training
- Finance: VIB-trained credit models produce compressed factor representations with well-calibrated default probabilities

VIB (Alemi et al. 2017): IB as a practical training objective; provides certified adversarial robustness

IB and Self-Supervised Contrastive Learning

Contrastive objectives are IB objectives over view-invariant features

- **SimCLR / CLIP**: train an encoder so that two augmented views of the same input produce similar representations; repel representations from different inputs
- IB interpretation: the contrastive loss maximises $I(T_1; T_2)$ (agreement between views) while implicitly minimising $I(T; X_{\text{augment}})$ (discarding augmentation-specific noise)
- **Tian et al. (2020)**: formally derive that optimal self-supervised representations are minimal sufficient statistics of the shared information between views – the IB optimum
- Finance application: two representations of the same asset (one from price history, one from news text) should share fundamental-value information and discard instrument-specific noise
- Open question: which augmentation policies for financial data (time warping, feature masking, noise injection) best isolate the underlying economic signal?

Contrastive SSL and IB are formally equivalent: contrastive training finds minimal sufficient statistics of view-shared information

Connections to Regularisation and Dropout

Standard regularisers can be interpreted as information compression

- **Weight decay** (ℓ_2 regularisation): penalises large weights; limits the capacity of each layer to store information; implicit compression
- **Dropout**: randomly zeroing units is equivalent to adding noise to the representation; increases $H(T|X)$, reducing $I(X;T)$ toward the IB target
- **VAE**: the ELBO objective is the VIB objective with $\beta = 1$; the latent space is an IB-optimal representation under Gaussian assumptions
- **Minimum description length (MDL)**: compressing the parameter description is equivalent to IB compression from an algorithmic information perspective

IB provides a theoretical umbrella for many regularisation methods: dropout, ℓ_2, VAE, and MDL

Limitations of the Information Bottleneck

IB is theoretically elegant but empirically difficult

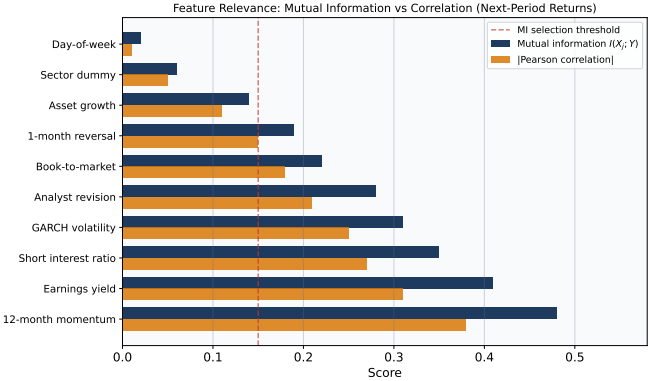
- Estimating mutual information in high dimensions is hard: most estimators have high variance and are computationally expensive
- MINE (Mutual Information Neural Estimation) and KNIFE are tractable but biased; results depend on estimator choice
- The compression phase controversy: IB dynamics vary with activation function and do not generalise to all architectures
- The IB framework assumes a fixed label Y; it does not naturally handle multi-task or self-supervised objectives
- For very high-dimensional inputs ($d > 1000$): the IB curve is hard to compute even in principle

IB is more useful as a design principle (VIB) than as an empirical theory of training dynamics

Finance Application: Feature Selection via Mutual Information

Ranking input features by their information about returns

- Mutual information: $I(X_j; Y)$ quantifies how much feature j predicts the return
- Unlike correlation, MI captures non-linear dependencies (e.g., skew, tail effects)
- IB-based selection: keep features with high $I(X_j; Y)$ and low redundancy $I(X_j; X_k)$



MI-based feature selection captures non-linear signal that correlation-based methods miss

Finance Application: Information-Theoretic Portfolio Construction

Applying the IB principle to factor model compression

- A factor model maps hundreds of raw signals to a small number of portfolio weights
- IB view: find the minimal compression T of all signals that maximally predicts next-period returns
- **Maximum entropy portfolio**: choose weights that maximise $H(R_p)$ subject to target return; equivalent to minimising $I(R_p; \text{noise})$
- **IB-regularised predictor**: train a VIB model on cross-sectional return data; the bottleneck forces the model to identify robust factors
- Key insight: a factor that is highly predictive but shares information with many other factors is redundant – IB-based selection removes it

IB for portfolio construction: find the minimal sufficient statistic of market signals for future returns

IB for Model Interpretability

Compression as a measure of representation quality

- A layer with low $I(X;T)$ has discarded irrelevant variation: its representation is more interpretable
- **Information-theoretic saliency**: the most informative input dimensions are those that contribute most to $I(X;T)$ at the last layer
- **SHAP and IB**: SHAP values approximate the Shapley information allocation – how each feature contributes to $I(X_{\text{all}};Y)$
- Compressed representations generalize better and are less sensitive to adversarial perturbations: IB compression is a natural defence
- Finance: regulators value interpretable models; IB-compressed representations identify the key drivers without noise

IB compression and interpretability are linked: a maximally compressed layer retains only causally relevant features

Code Pointer

Information bottleneck and mutual information in Python

- **VIB implementation**: parameterise the encoder as a Gaussian; KL divergence from PyTorch's `kl_divergence`; add $\beta \cdot \text{KL}$ to the cross-entropy loss
- **MINE** (Mutual Information Neural Estimation): adversarial MI estimator; `pip install mine-pytorch`
- **scikit-learn**: `mutual_info_classif(X, y)` for feature selection; based on k-nearest-neighbour MI estimator
- **Information-theoretic feature selection**: `from sklearn.feature_selection import mutual_info_regression`
- Finance: compute $I(X_j; Y_{\text{return}})$ for all candidate factors; select top-k non-redundant features

Start with sklearn's mutual_info_classif for feature ranking; use VIB for end-to-end regularised training

Summary and Open Questions

Three things to remember

- The IB principle: compress the input ($\downarrow I(X;T)$) while retaining predictive power ($\uparrow I(T;Y)$); the optimal representation is a minimal sufficient statistic of X for Y
- Variational IB (VIB) is the practical instantiation: a VAE-like encoder with a β-KL regulariser; provides robustness, calibration, and compression simultaneously
- Finance applications: IB-based feature selection captures non-linear signal; VIB-trained credit models produce well-calibrated and interpretable default predictions

Open questions:

1. How do we estimate mutual information reliably in high-dimensional continuous spaces?
2. Does the IB principle extend naturally to self-supervised and contrastive learning objectives?
3. Can IB-optimal representations satisfy regulatory interpretability requirements in finance?

Further reading: Tishby et al. (1999) IB; Alemi et al. (2017) VIB; Saxe et al. (2018) IB controversy