The IB Lagrangian is given by:

$$\min_{q(z|x)} L = H(y \mid z) + \lambda I(z; x)$$

Introduce an auxiliary variable and consider the minimization problem:

$$\min_{q(z|x), p(z)} L = H(y \mid z) + \lambda \mathbb{E}_{p(x)}[KL(q(z \mid x) \parallel p(z))]$$

Notice that if the task is reconstruction (i.e., $y = x$) then this is the loss function of a VAE (with an extra coefficient in front of the KL term).

**Recall:** The VAE loss can be derived from variational inference and can be thought as a two part code: structure of the data + reconstruction error.
Learning disentangled representations
(Higgins et al., 2017, Burgess et al., 2017)

Start with very high $\beta$ and slowly decrease during training.

**Beginning:** Very strict bottleneck, only encode most important factor

**End:** Very large bottleneck, encode all remaining factors

Think of it as a non-linear PCA, where *training time* disentangles the factors.
Learning disentangled representations

(Higgins et al., 2017, Burgess et al., 2017)

Each component of the learned representation corresponds to a different semantic factor.

Components of the representation $z$

Image seed

Higgins et al., $\beta$-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, 2017
Burgess et al., Understanding Disentangling in beta-VAE, 2017

Pictures courtesy of Higgins et al., Burgess et al.
Learning invariant representations for a task

Deeper layers filter increasingly more nuisances

Stronger bottleneck = more filtering

Only informative part of the image

Other information is discarded

The catch

What if we just represent an image by its index in the training set (or by a unique hash)?

<table>
<thead>
<tr>
<th>X</th>
<th>Z</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>24,576 bits</td>
<td>16 bits</td>
<td>4 bits</td>
</tr>
<tr>
<td><img src="https://via.placeholder.com/150" alt="Image" /></td>
<td><img src="https://via.placeholder.com/150" alt="Image" /></td>
<td><img src="https://via.placeholder.com/150" alt="Image" /></td>
</tr>
<tr>
<td></td>
<td>0000000000000000</td>
<td>0100</td>
</tr>
<tr>
<td></td>
<td>0000000000000001</td>
<td>0001</td>
</tr>
<tr>
<td></td>
<td>0000000000000010</td>
<td>0010</td>
</tr>
<tr>
<td></td>
<td>0000000000000011</td>
<td>0101</td>
</tr>
</tbody>
</table>

It is a sufficient representation and it is close to minimal.
This Information Bottleneck is wishful thinking

The IB is a statement of desire for future data we do not have:

\[
\min_{q(z|x)} \mathcal{L} = H_{p,q}(y|z) + \beta I(z; x)
\]

What we have is the data collected in the past.

What is the best way to use the past data in view of future tasks?
Training data

\{ (car, horse, deer, ...) \}

Weights

Invariant representation

Testing
The Kolmogorov Structure of a Task

The space of solutions can be explored using the Lagrangian:

\[ L = L(D; M) + \beta K(M) \]

\[ K(M) \leq KL(q(z|x) \parallel p(z)) \]

The coding length of the weights can be approximated by Variational Inference.
The Local Information Bound

Let \( w^* \) be a local minimum. The optimal amount of gaussian noise is to add is:

\[
\Sigma = \left( I + \frac{2\lambda^2}{\beta} F(w_0) \right)^{-1},
\]

where \( F(w^*) \) is the Fisher Information Matrix (equiv. Hessian) computed in \( w^* \).

\[
I(w; \mathcal{D}) \leq \frac{\|w\|^2}{\lambda^2} + \log |2\lambda^2 NF(w^*) + I|
\]

Weight Information is bounded by the geometry of the loss landscape*

Flat minima have low information in the weights.

* Hochreiter and Schmidhuber, Flat Minima, Neural Computation, 1997
For random labels, at $\beta = 1$ (the VLBO value) there is a phase transition between overfitting and underfitting.
Bias-variance tradeoff

Information is a better measure of complexity

Parametrizing the complexity with information in the weights, we recover bias-variance trade-off trend.

Achille and Soatto, *Emergence of Invariance and Disentanglement in Deep Representations*, JMLR 2018
PAC-Bayes bound (Catoni, 2007; McAllester 2013).

\[ L_{\text{test}}(q(w|\mathcal{D})) \leq \frac{1}{N(1 - 1/2\beta)} \left[ H_{p,q}(y|x,w) + \beta \ KL(q(w|\mathcal{D})||p(w)) \right] \]

Corollary. Minimizing the IB Lagrangian for the weights minimizes an upper bound on the test error (Dziugaite and Roy, 2017; Achille and Soatto, 2017)

This gives non-vacuous generalization bounds! (Dziugaite and Roy, 2017)
A new Information Bottleneck

Weights IB Overfitting

\[ D \xrightarrow{w} p(y|x) \]

\[ \min_w \mathcal{L} = H_{p,q_w}(y|z) + \beta I(D; w) \]

Activations IB Invariance

\[ x \xrightarrow{z} y \]

\[ \min_{q(z|x)} \mathcal{L} = H_{p,q}(y|z) + \beta I(z; x) \]