On the Information Bottleneck

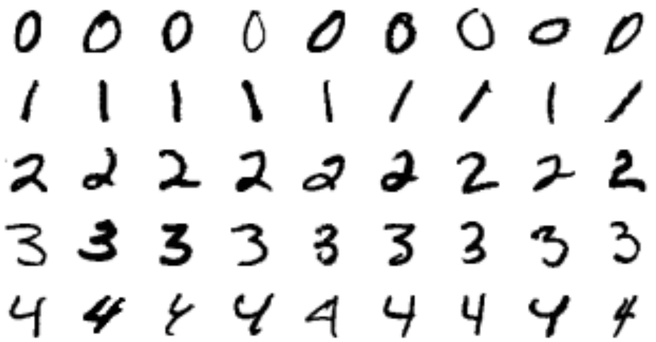
Wave Ngampruetikorn

Initiative for the Theoretical Sciences & Center for the Physics of Biological Function
The Graduate Center, CUNY

SIMONS FOUNDATION

THE GRADUATE CENTER
CITY UNIVERSITY OF NEW YORK

Extracting 'relevant' information from data underpins all forms of learning

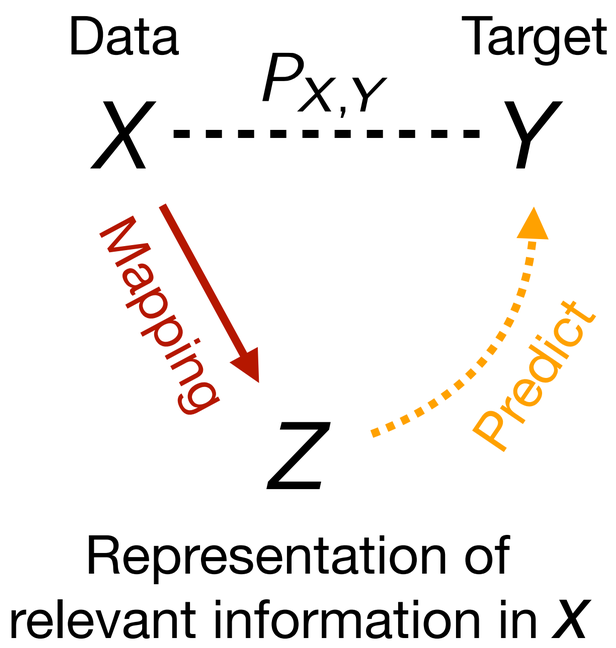


Classifying handwritten digits
requires extracting the ***right***
features from the space of pixels

But relevance is ill-defined without further qualification

The ***right*** features depend on what we want to predict,
eg, the digit, identity of writer, or brand of pen

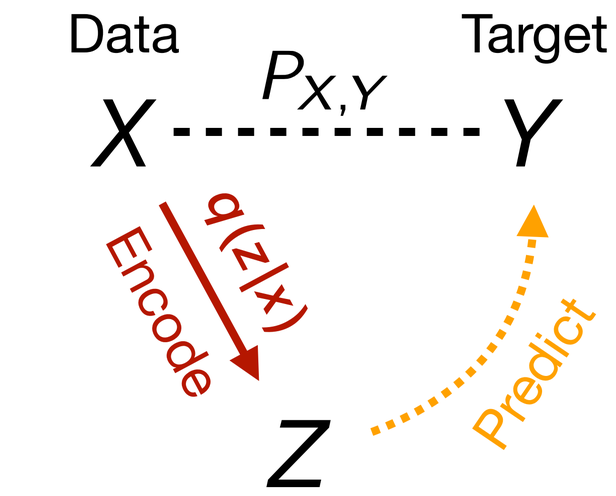
Information Bottleneck *defines* relevant information as the bits in data *X* that can predict 'target' *Y*



- Suppose we want to predict ***Y*** from some observable ***X***
- *Relevant* information in ***X*** is the features of ***X*** that are predictive of ***Y***
- We want to find a mapping from ***X*** to ***Z*** such that ***Z*** is
  1. **relevant** (predictive of ***Y***)
  2. **compact** (discards irrelevant bits)

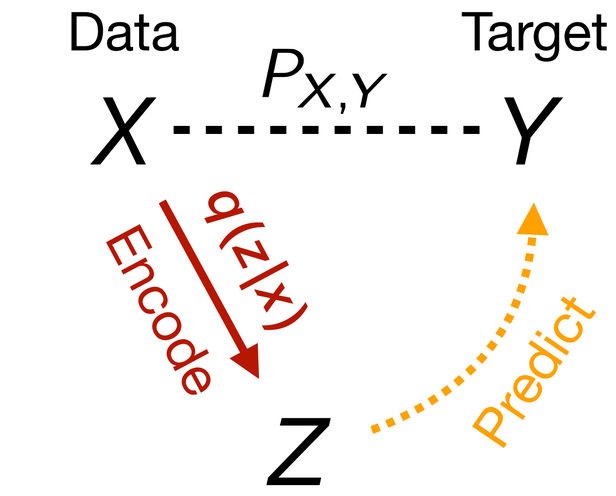
Tishby, Pereira & Bialek (37th Allerton Conference 1999)

Mapping is a conditional probability distribution and information is Shannon mutual information



- A conditional probability distribution—the encoder $q(z|x)$—defines the mapping from X to Z
- Measure relevant information with the mutual information $I(Z;Y)$—how predictive Z is of Y
- And irrelevant information with $I(Z;X)$—how predictive Z is of X

Extracting relevant information is optimization



Representation of relevant information in X

$$\min_{q(z|x)} L \quad \text{with} \quad L = I(Z;X) - \beta I(Z;Y)$$

Lagrange multiplier

$I(Z;X)$: favors compression

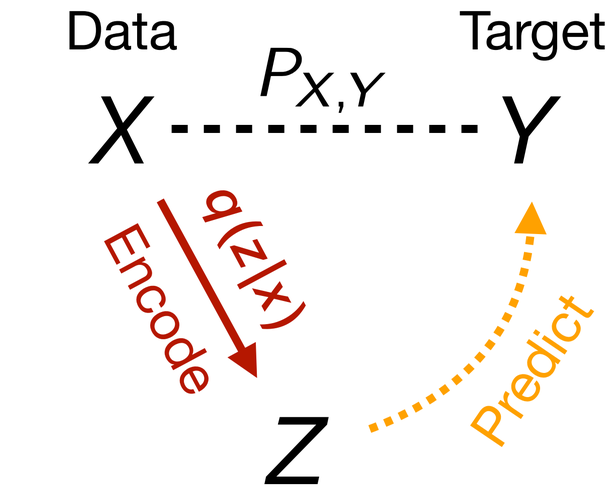
$\beta I(Z;Y)$: favors prediction

- We want to find an encoder that
  1. maximizes **relevant information**
  2. minimizes **irrelevant information**

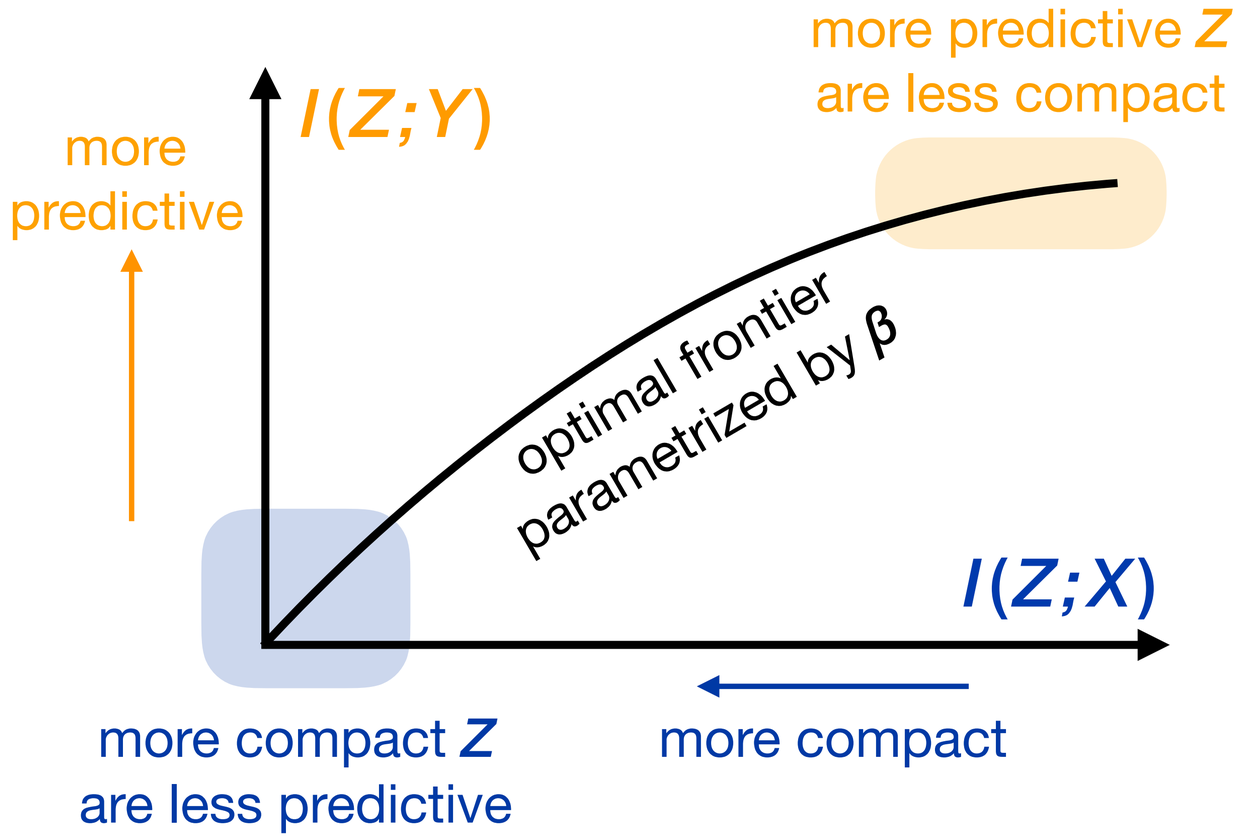
Tishby, Pereira & Bialek (37th Allerton Conference 1999)

Trade-off exists between compression and prediction

$$\min_{q(z|x)} L \quad \text{with} \quad L = I(Z;X) - \beta I(Z;Y)$$



Representation of relevant information in X



IB has a formal solution

$$\mathcal{L}[q(z|x)] = \overbrace{I(Z;X) - \beta I(Z;Y)}^{\text{IB loss}} - \overbrace{\sum_x \lambda(x) \sum_z q(z|x)}$$

This term ensures that *q*(*z*|*x*) is normalized for all *x*

Each term is a function of the encoder ***q*(*z*|*x*)**

$$I(Z;X) = \sum_x p(x) \sum_z q(z|x) \ln \frac{q(z|x)}{q(z)}$$

$$I(Z;Y) = \sum_y p(y) \sum_z q(z|y) \ln \frac{q(z|y)}{q(z)}$$

From Markov constraint *Z*—*X*—*Y*

$$q(z|y) = \sum_x q(z|x)p(x|y)$$

$$q(z) = \sum_x q(z|x)p(x)$$

Minimizing the loss function gives

$$\frac{\delta \mathcal{L}[q(z|x)]}{\delta q(z|x)} = 0 \quad \Rightarrow \quad q(z|x) = \frac{1}{\mathcal{N}} q(z) e^{-\beta \mathrm{D_{KL}}[p(y|x)||q(y|z)]}$$

normalization factor

Tishby, Pereira & Bialek (1999)

IB is versatile with applications in many areas

- Neural coding (Palmer et al *PNAS* 2015)
- Evolutionary population dynamics (Sachdeva et al *bioRxiv 2020.04.29.069179*)
- Statistical physics (Gordon et al *arXiv:2012.01447*)
- Clustering (Strouse & Schwab *Neural Comput.* 2019),
- Deep learning (Alemi et al *ICLR* 2017; Achille & Soatto *JMLR* 2018)
- Reinforcement learning (Goyal et al *ICLR* 2019)
- …

PNAS

Predictive information in a sensory population

Stephanie E. Palmer[a,b], Olivier Marre[c,d], Michael J. Berry II[c,d], and William Bialek[a,b,1]

[a]Joseph Henry Laboratories of Physics and [b]Lewis–Sigler Institute for Integrative Genomics, and [c]Department of Molecular Biology and [d]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544

Contributed by William Bialek, April 13, 2015 (sent for review January 19, 2014)

Guiding behavior requires the brain to make predictions about the future values of sensory inputs. Here, we show that efficient predictive computation starts at the earliest stages of the visual system. We compute how much information groups of retinal ganglion cells carry about the future state of their visual inputs and show that nearly every cell in the retina participates in a group of cells for which this predictive information is close to the physical limit set by the statistical structure of the inputs themselves. Groups of cells in the retina carry information about the future state of their own activity, and we show that this information can be compressed further and encoded by downstream predictor neurons that exhibit feature selectivity that would support predictive computations. Efficient representation of predictive information is a candidate principle that can be applied at each stage of neural computation.

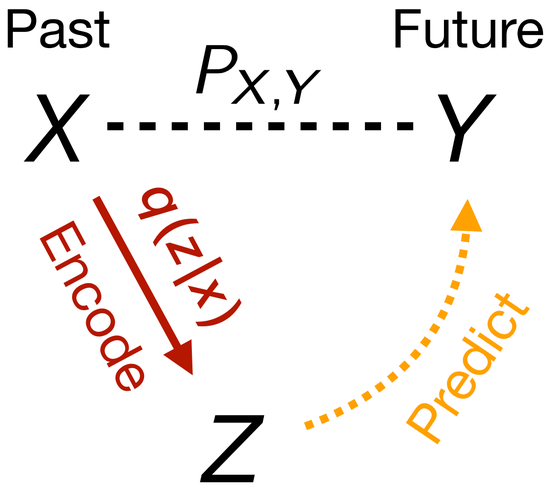
becomes a binary "word" $w_t \equiv \{\sigma_1(t),\ \sigma_2(t),\ \cdots,\ \sigma_N(t)\}$. If we (or the brain) observe the pattern of activity w_t at time t, how much do we know about the position of the moving object? Neurons are responding to the presence of the object, and to its motion, but there is some latency in this response, so that w_t will be maximally informative about the position of the object at some time in the past, $x_{t'<t}$. On the other hand, we know that the brain is capable of predicting the future position of moving objects and that these ganglion cells provide all of the visual data on which such predictions are based, so it must be true that w_t also provides some information about $x_{t'>t}$.

We can make these ideas precise by estimating, in bits, the information that the words w_t provide about the position of the object at time t' (5–8):

Predictive information in a sensory population

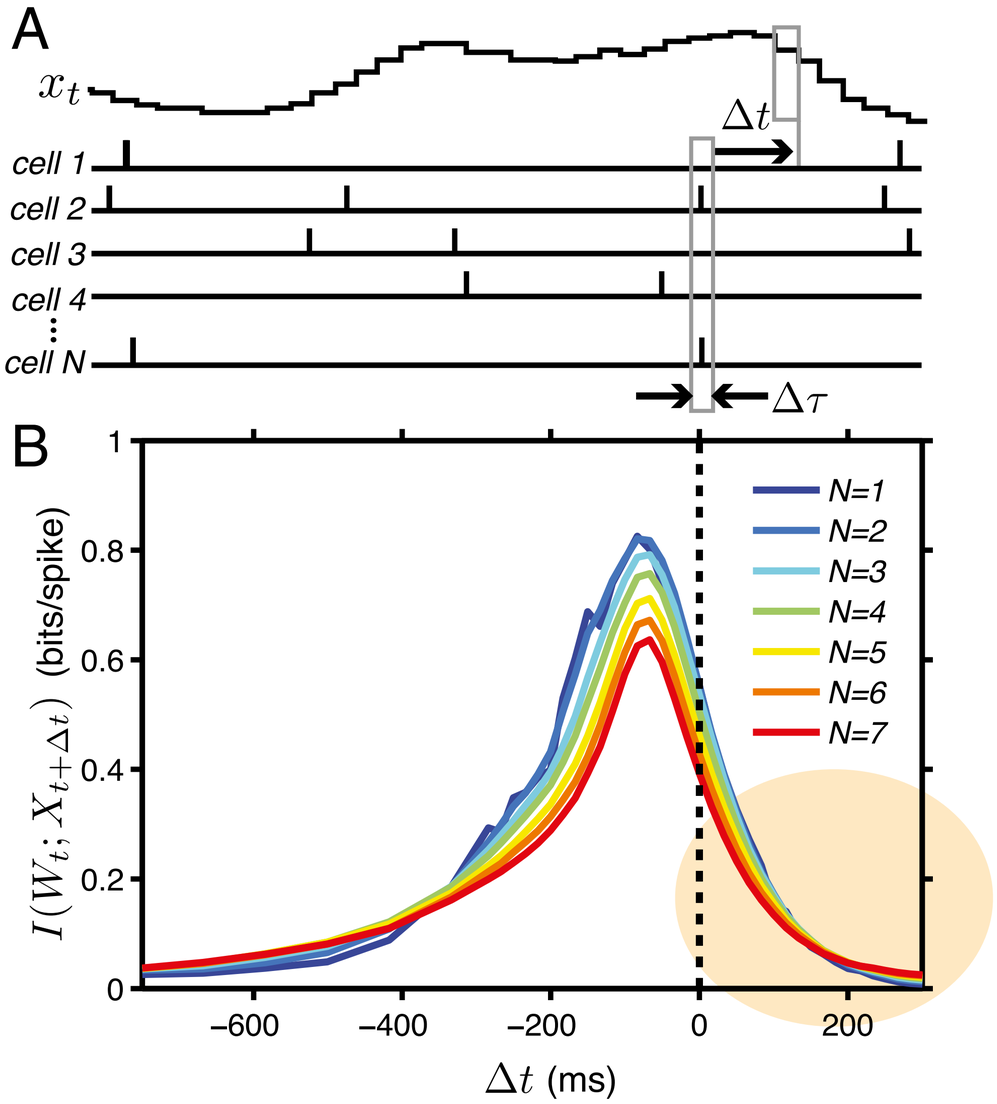
Stephanie E. Palmer[a,b], Olivier Marre[c,d], Michael J. Berry II[c,d], and William Bialek[a,b,1]

[a]Joseph Henry Laboratories of Physics and [b]Lewis–Sigler Institute for Integrative Genomics, and [c]Department of Molecular Biology and [d]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544



Neural activity

population of ganglion cells in the salamander retina



Trajectory of a moving bar X_t

Neural spiking responses W_t

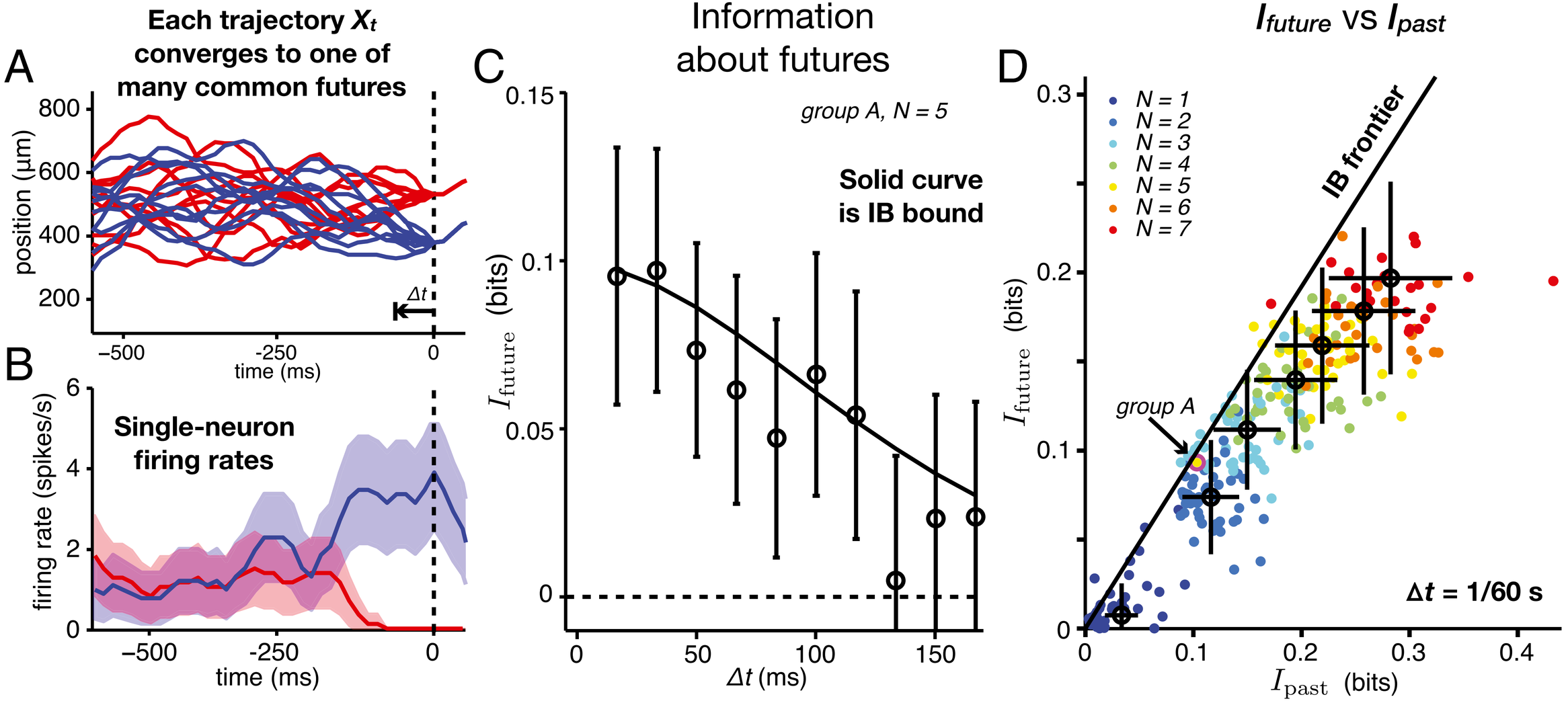
Encoded information per neuron $I(W_t ; X_{t+\Delta t})$

Information about future position

Predictive information in a sensory population

Stephanie E. Palmer[a,b], Olivier Marre[c,d], Michael J. Berry II[c,d], and William Bialek[a,b,1]

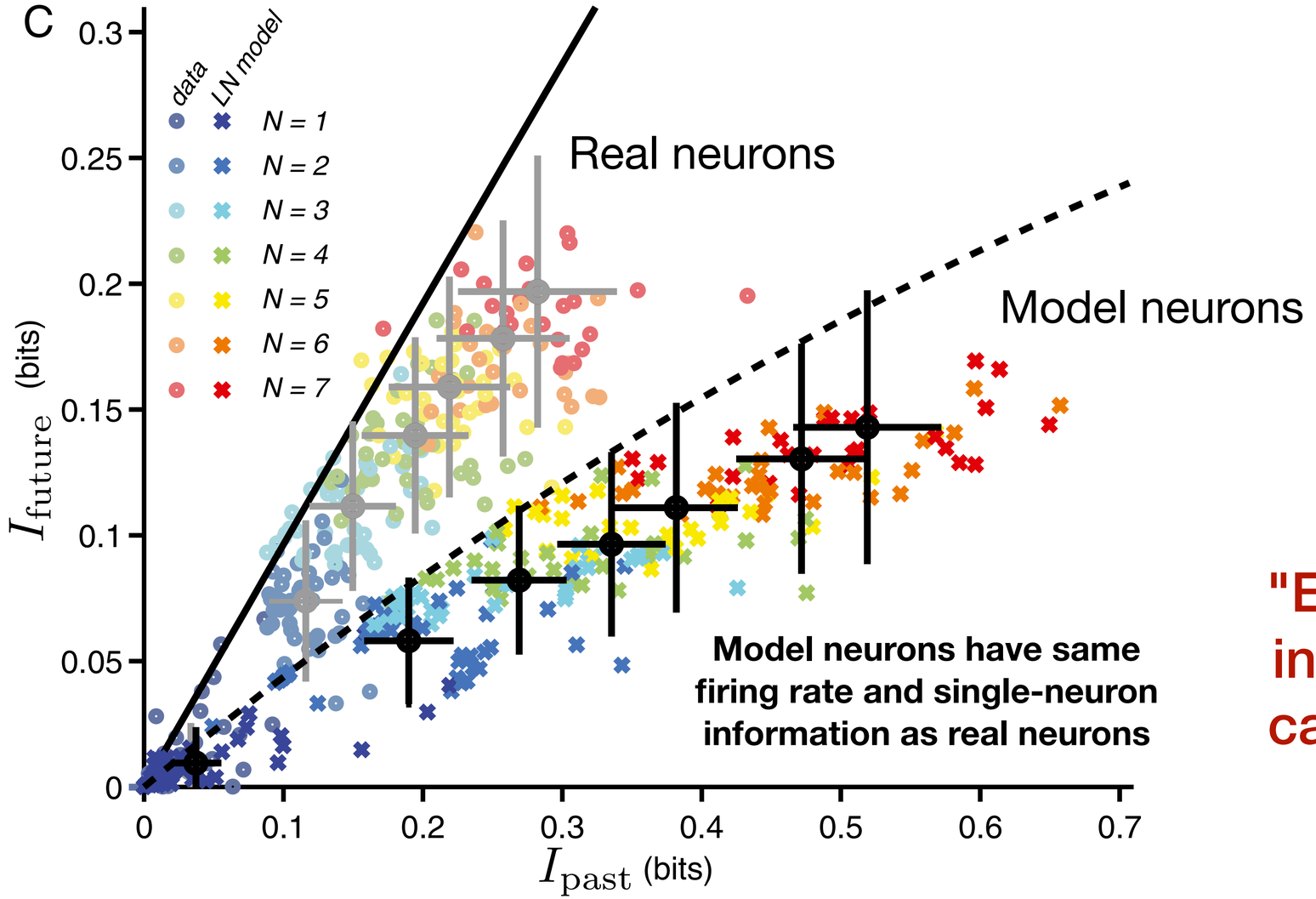
[a]Joseph Henry Laboratories of Physics and [b]Lewis–Sigler Institute for Integrative Genomics, and [c]Department of Molecular Biology and [d]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544

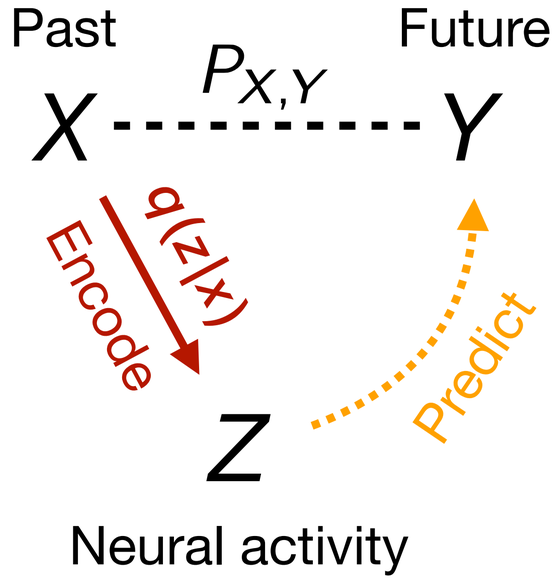


Predictive information in a sensory population

Stephanie E. Palmer[a,b], Olivier Marre[c,d], Michael J. Berry II[c,d], and William Bialek[a,b,1]

[a]Joseph Henry Laboratories of Physics and [b]Lewis–Sigler Institute for Integrative Genomics, and [c]Department of Molecular Biology and [d]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544





"Efficient coding of predictive information is a principle that can be applied at every stage of neural computation."

DEEP VARIATIONAL INFORMATION BOTTLENECK

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, Kevin Murphy
Google Research
{alemi,iansf,jvdillon,kpmurphy}@google.com

ABSTRACT

We present a variational approximation to the information bottleneck of Tishby et al. (1999). This variational approach allows us to parameterize the information bottleneck model using a neural network and leverage the reparameterization trick for efficient training. We call this method "Deep Variational Information Bottleneck", or Deep VIB. We show that models trained with the VIB objective outperform those that are trained with other forms of regularization, in terms of generalization performance and robustness to adversarial attack.

DEEP VARIATIONAL INFORMATION BOTTLENECK

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, Kevin Murphy
Google Research

Estimating mutual information from high dimensional data is *very* difficult

$$I(Z;X) = \sum_x p(x) \sum_z q(z|x) \ln \frac{q(z|x)}{q(z)}$$

$$I(Z;Y) = \sum_y p(y) \sum_z q(z|y) \ln \frac{q(z|y)}{q(z)}$$

From Markov constraint *Z—X—Y*

$$q(z|y) = \sum_x q(z|x)p(x|y)$$

$$q(z) = \sum_x q(z|x)p(x)$$

Require summing over all data
every time we update the encoder!

DEEP VARIATIONAL INFORMATION BOTTLENECK

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, Kevin Murphy
Google Research

Estimating variational bounds is easier

$$
\begin{aligned}
I(Z;X) &= \sum_x p(x) \sum_z q(z|x) \ln \frac{q(z|x)}{\boxed{q(z)}} \\
&= \sum_x p(x) \sum_z q(z|x) \ln \frac{q(z|x)}{\boxed{r(z)}} \\
&\quad - \boxed{\sum_z q(z) \ln \frac{q(z)}{r(z)}} \quad \textbf{KL divergence is non-negative} \\
&\leq \sum_x p(x) \sum_z q(z|x) \ln \frac{q(z|x)}{\boxed{r(z)}} \\
&\equiv \hat{I}_{\mathrm{VU}}(Z;X)[q(z|x), \boxed{r(z)}]
\end{aligned}
$$

Variational Upper bound

$$
\begin{aligned}
I(Z;Y) &= \sum_z q(z) \sum_y q(y|z) \ln \frac{\boxed{q(y|z)}}{p(y)} \\
&= \sum_z q(z) \sum_y q(y|z) \ln \frac{\boxed{s(y|z)}}{p(y)} \\
&\quad + \sum_z q(z) \boxed{\sum_y q(y|z) \ln \frac{q(y|z)}{s(y|z)}} \quad \textbf{KL divergence is non-negative} \\
&\geq \sum_{x,y} p(x,y) \sum_z q(z|x) \ln \frac{\boxed{s(y|z)}}{p(y)} \\
&\equiv \hat{I}_{\mathrm{VL}}(Z;Y)[q(z|x), \boxed{s(y|z)}]
\end{aligned}
$$

Variational Lower bound

DEEP VARIATIONAL INFORMATION BOTTLENECK

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, Kevin Murphy
Google Research

Estimating variational bounds is easier

$$I(Z;X) \leq \hat{I}_{\mathrm{VU}}(Z;X)[q(z|x), r(z)] \qquad I(Z;Y) \geq \hat{I}_{\mathrm{VL}}(Z;Y)[q(z|x), s(y|z)]$$

$$\begin{aligned} L[q(z|x)] &= I(Z;X) - \beta I(Z;Y) \\ &\leq \hat{I}_{\mathrm{VU}}(Z;X) - \beta I_{\mathrm{VL}}(Z;Y) \\ &\equiv L_{\mathrm{VIB}}[q(z|x), s(y|z), r(z)] \end{aligned}$$

VIB Loss upper bounds IB Loss

Use neural nets to approximate ***q***, ***s*** and ***r***

VIB improves generalization and adversarial robustness in classification tasks

(reduce overfitting)

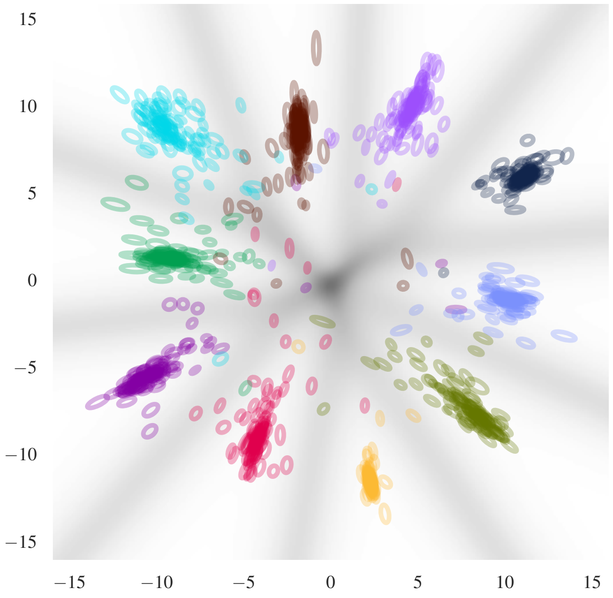
see original paper

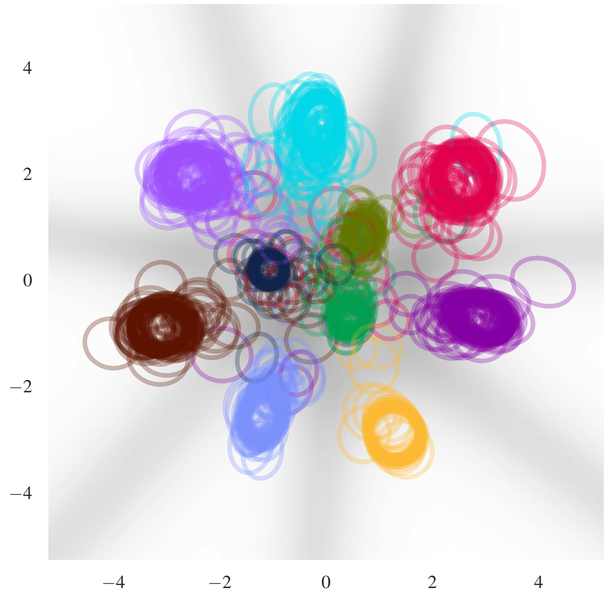
DEEP VARIATIONAL INFORMATION BOTTLENECK

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, Kevin Murphy
Google Research

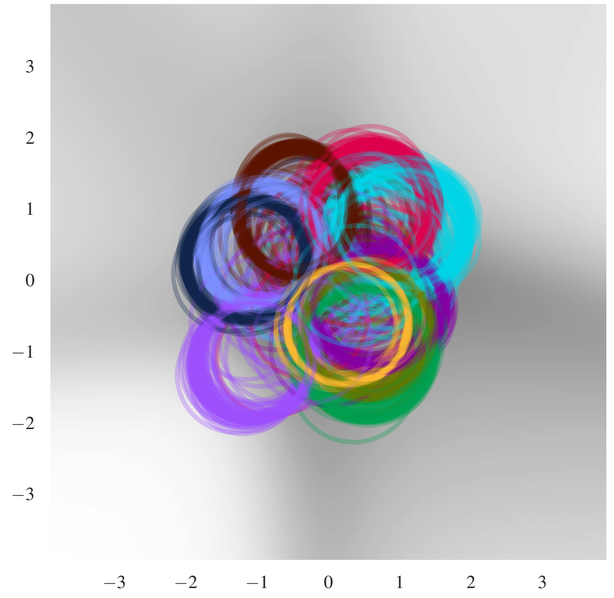
2D VIB embedding of MNIST

Less compression, more predictive





More compression, less predictive



(a) $\beta = 10^{-3}$, $\text{err}_{\text{mc}} = 3.18\%$, $\text{err}_1 = 3.24\%$

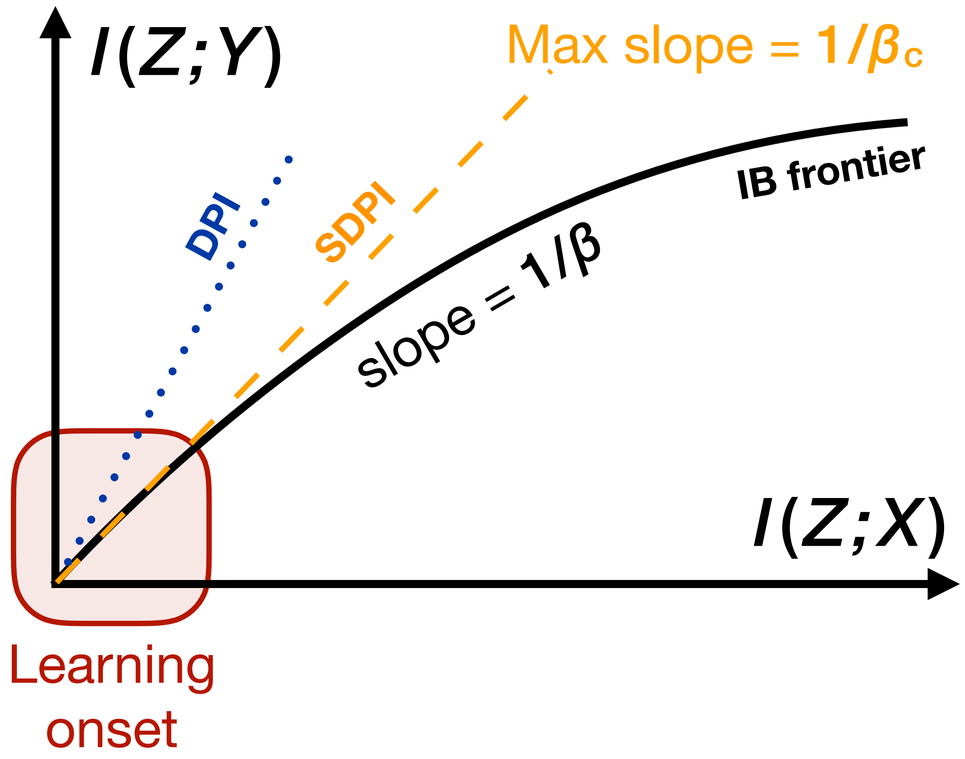
(b) $\beta = 10^{-1}$, $\text{err}_{\text{mc}} = 3.44\%$, $\text{err}_1 = 4.32\%$

(c) $\beta = 10^{0}$, $\text{err}_{\text{mc}} = 33.82\%$, $\text{err}_1 = 62.81\%$.

IB framework is precise and appealing but analytically intractable in general

Except for Gaussian variables (Chechik et al JMLR 2005)

We consider the limiting case of learning onset where relevant information per extracted bit is maximum



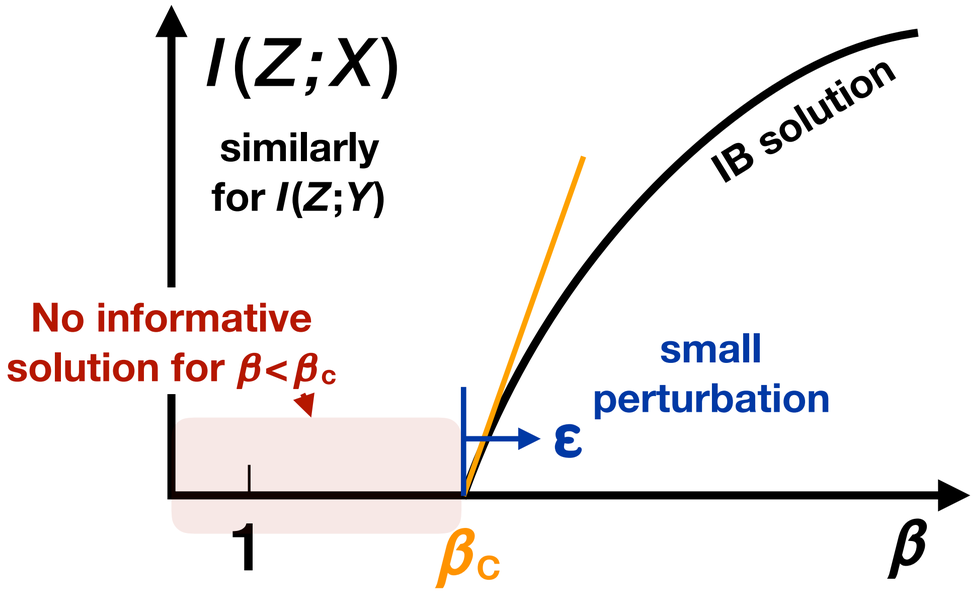
Data processing inequality
$I(Z;Y) \le I(Z;X)$ [DPI]

Relevant information ratio $= \dfrac{I(Z;Y)}{I(Z;X)} \le \beta_c^{-1} \le 1$

Strong data processing inequality
$I(Z;Y) \le (1/\beta_c) \times I(Z;X)$ [SDPI]

$$\min_{q(z|x)} L \quad \text{with} \quad L = I(Z;X) - \beta I(Z;Y)$$

We develop perturbation theory around the learning onset



We expand the encoder around the learning onset

$$q(z|x) = q_0(z) + \varepsilon q_1(z|x) + \varepsilon^2 q_2(z|x) + \dots$$

At learning onset (ε=0) *any* uninformative encoder [$q(z|x)=q(z)$] is IB optimal (no informative solution exists)

The extracted information expands as follows

$$I_{Z;X}[q] = \cancel{I^{(0)}_{Z;X}[q_0]} + \varepsilon I^{(1)}_{Z;X}[q_1] + \varepsilon^2 I^{(2)}_{Z;X}[q_1, q_2] + \dots$$

always vanish for uninformative q_0

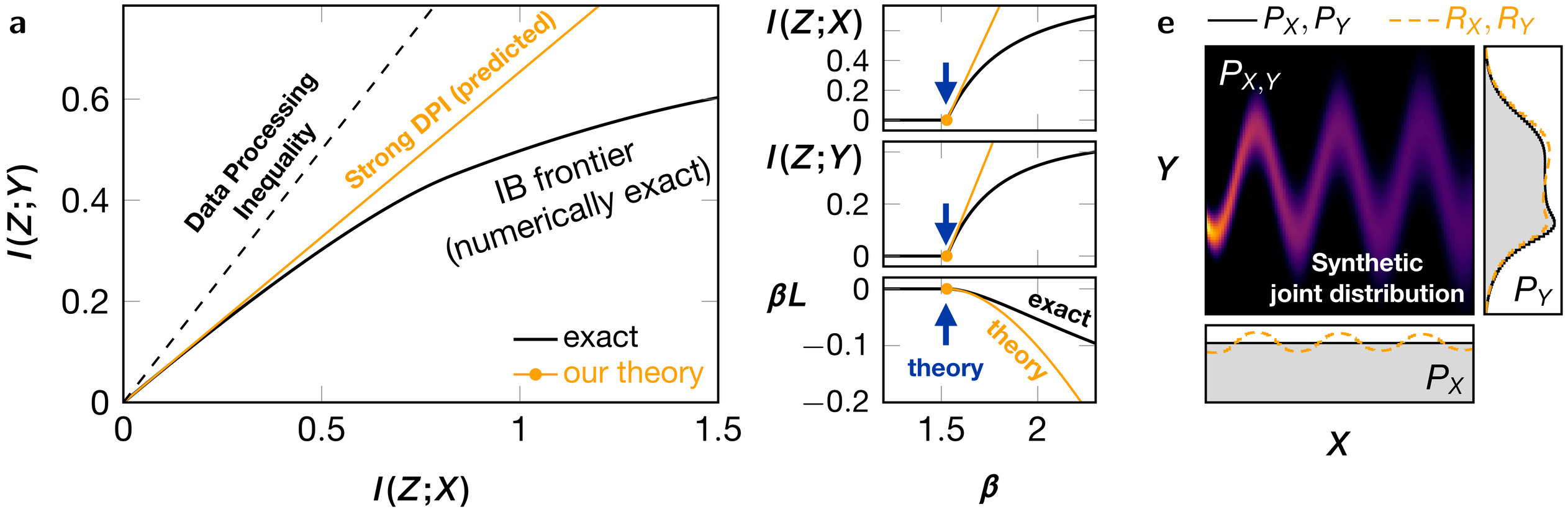
vanish in previous works* which assume z cannot take values outside support of q_0

2nd order term is required to fix scale of 1st-order response

**Gedeon, Parker & Dimitrov (Entropy 2012)*
**Wu, Fischer, Chuang & Tegmark (UAI 2019)*
**Wu & Fischer (ICLR 2020)*

See, Ngampruetikorn & Schwab (Machine Learning and the Physical Sciences Workshop at NeurIPS 2020)

Our theory predicts maximum relevant information ratio ($1/\beta_c$) and completely characterizes learning onset



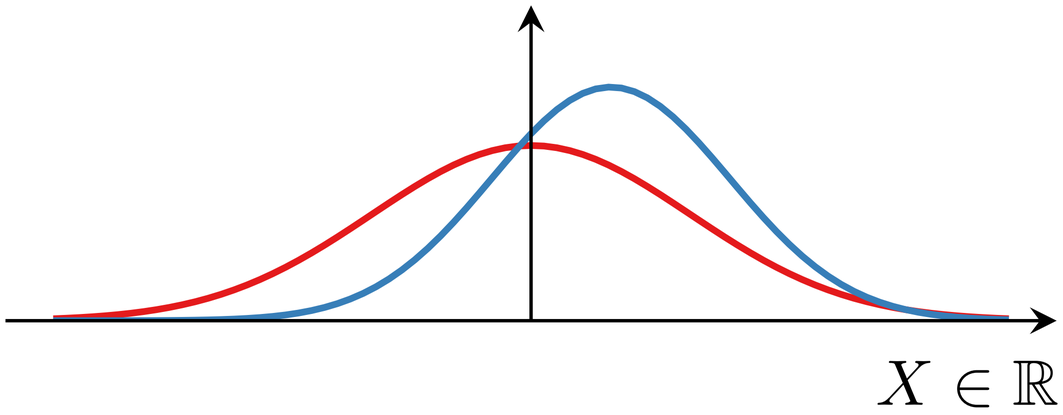
See, Ngampruetikorn & Schwab (Machine Learning and the Physical Sciences Workshop at NeurIPS 2020)

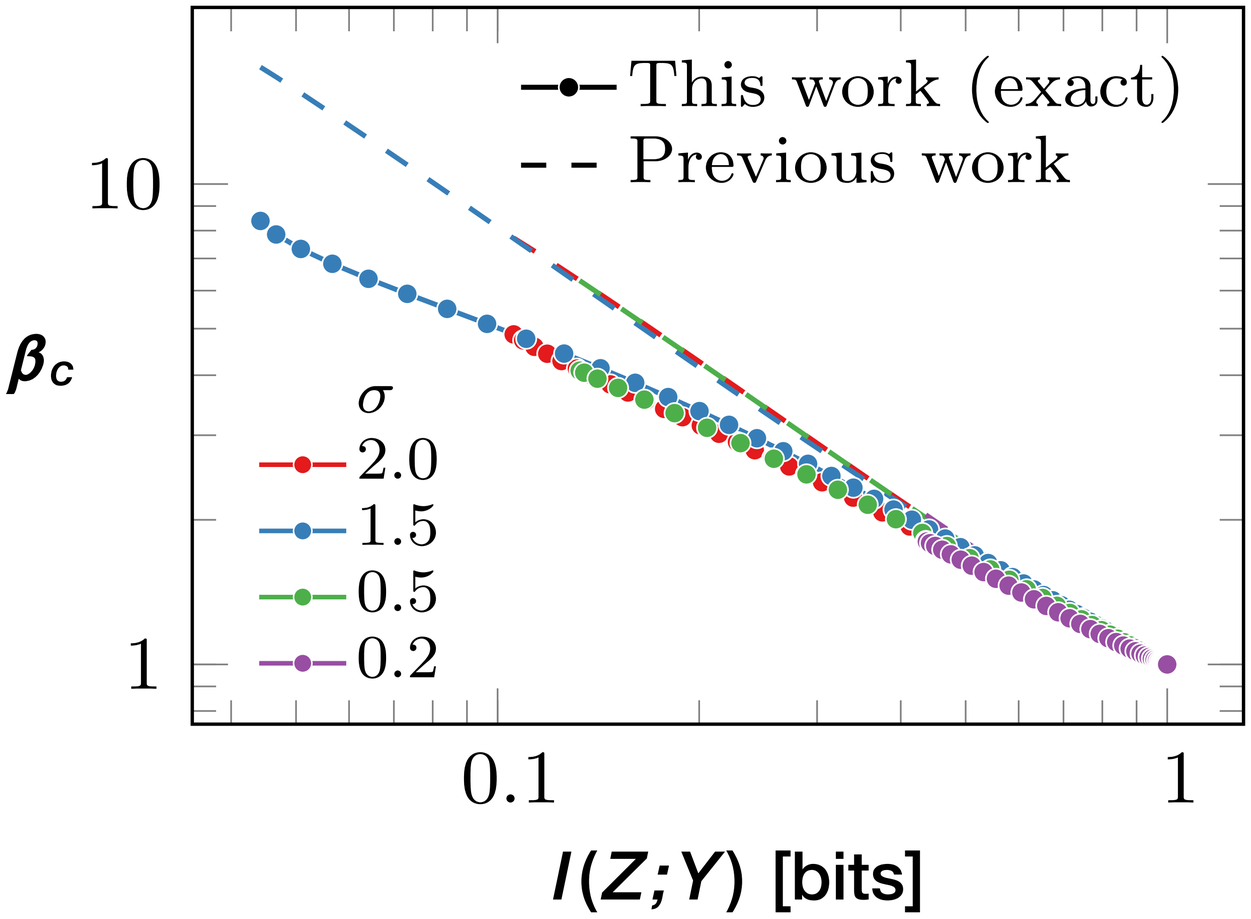
Previous work yields only an upper bound for β_c

Learning to predict binary label *Y*

$$Y \in \{y_1, y_2\},\ p(y_1) = p(y_2) = 1/2$$

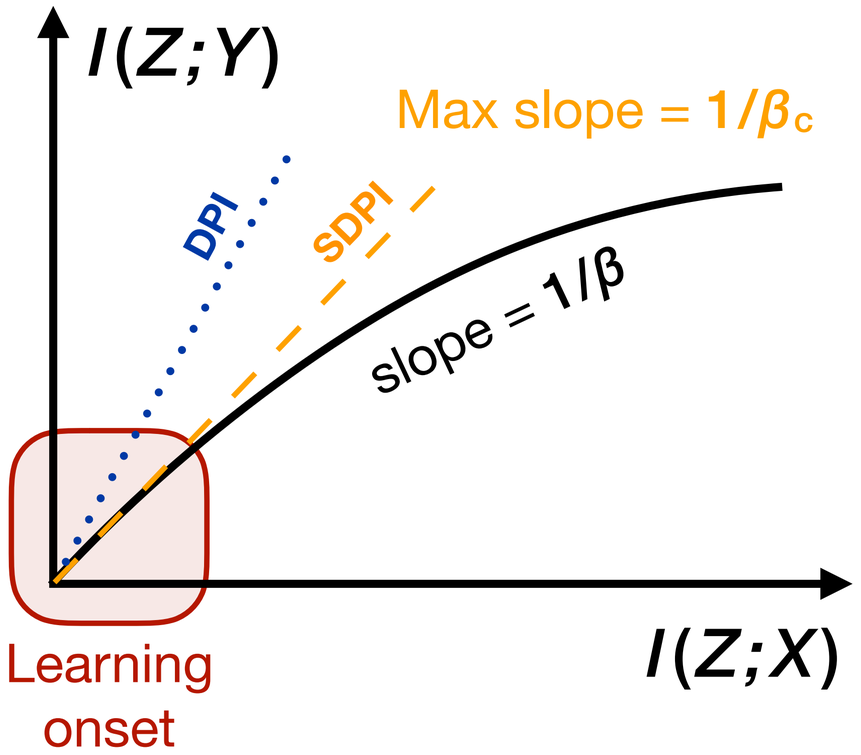
$$p(x|y_1) \sim \mathcal{N}(0, 1)$$
$$p(x|y_2) \sim \mathcal{N}(\mu, \sigma^2)$$





Ngampruetikorn & Schwab 2021 (under review)

Our results have potential implications for fundamental research and in practice



Maximum relevant information ratio ($1/\beta_c$)

- Related to contraction coefficient in information theory [Anantharam etal *arXiv:1304.6133*]
- Tight bound of the thermodynamic efficiency in predictive systems [Still etal *PRL* 2012]
- Useful measure of correlations [Kim etal *NeurIPS* 2017]
- Might help tune hyperparameters in deep learning techniques such as VIB [Wu etal *UAI* 2019]

Thank you!

CUNY NIH NSF SIMONS FOUNDATION

Ngampruetikorn & Schwab
(Machine Learning and the Physical Sciences Workshop at NeurIPS 2020)