Master's Thesis Artificial Intelligence
University of Amsterdam
July 12th, 2021

Taking a step back: assessing the TransformerVAE as a latent variable model first

Author Claartje Barkhof

Academic Supervisors
David Stap
Dr. Wilker Ferreira Aziz

Supervisors at DPG Media
Joris Baan
Lucas de Haas

Outline

1. Introduction into latent variable modelling of language
2. Problem statement & research goals
3. A bit more background & related work
4. An alternative mode of analysis: focusing on the quality of approximate posterior inference
5. Conclusion

1 Introduction to modelling language with a latent variable model

Learning representations of language (written text)

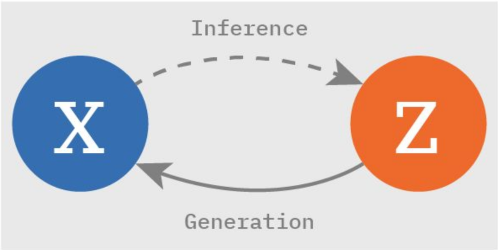
- Extracting ‘relevant’ features from a piece of text for a downstream task
  - e.g. for news personalisation, search, summarisation, etc.
- Black-box, distributed representations coming from large Transformer[1] neural networks are shown to be effective in the context of transfer learning
  - e.g. BERT for NLP benchmark tasks[2]
- ***but,*** **not necessarily:**
  - global or high-level (scattered throughout the network)
  - easy to control / manipulate (often only by textual prompts)
  - generalising outside the data distribution (interpolation is not possible)

[1] Vaswani et al., 2017
[2] Devlin et al., 2019

Deep generative latent variable modelling of language

Latent variable model

- **Explicit representation**
  - Hierarchical statistical model in which unobserved variables z (latent representation) can be used to explain observed variables x (data)
- **Global representation**
  - The latent variables may govern global, high level and abstract features of language such as topic, stylistic features or sentiment
- **Latent space structure**
  - Latent representations are defined by smooth distributions that are pressured to exploit neighborhood in an efficient way
  - Smooth manipulation in the latent space may induce complex patterns of variation in the data space
  - Generalisation outside of the data distribution



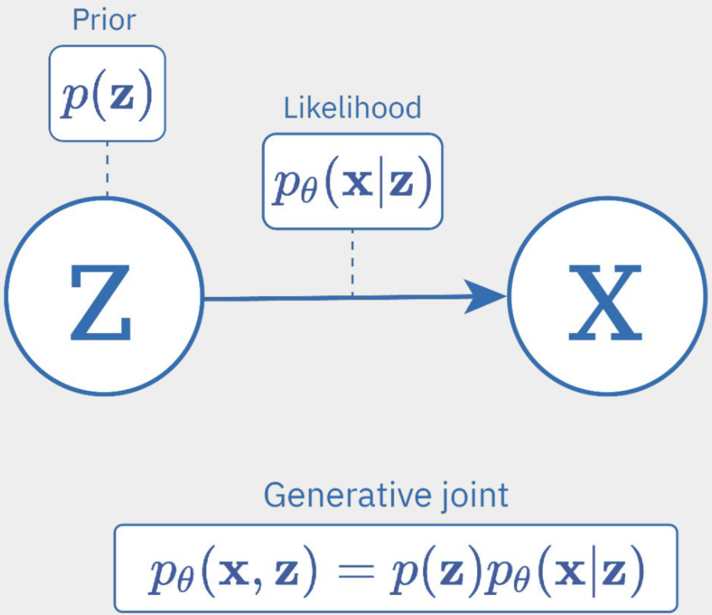
Variational Autoencoder[1] for modelling language

How to learn such a generative model from data?

- Optimising the log likelihood of the model $L_D(\theta) = P(X|\theta)$ entails integrating out z, which is generally intractable

How to "use" such a model?

- The induced posterior $p_\theta(z/x)$ lets us infer representations from the data, but is also generally intractable.



[1] Kingma & Welling, 2014

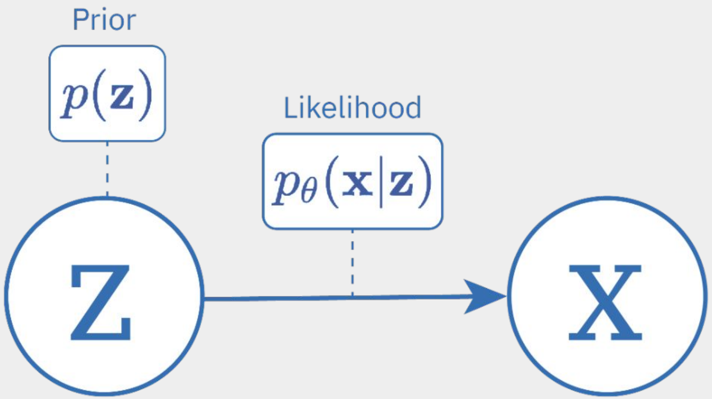
Variational Autoencoder[1] for modelling language

We can use a **Variational Autoencoder,** a framework that prescribes how to learn a latent variable model, that:

- Uses variational inference for optimisation, **optimising evidence lower bound (ELBO)**
  → *circumventing intractable likelihood*

- Thereby introducing an **approximate posterior distribution** that allows for approximate inference
  → *circumventing the intractable true posterior p(z|x)*

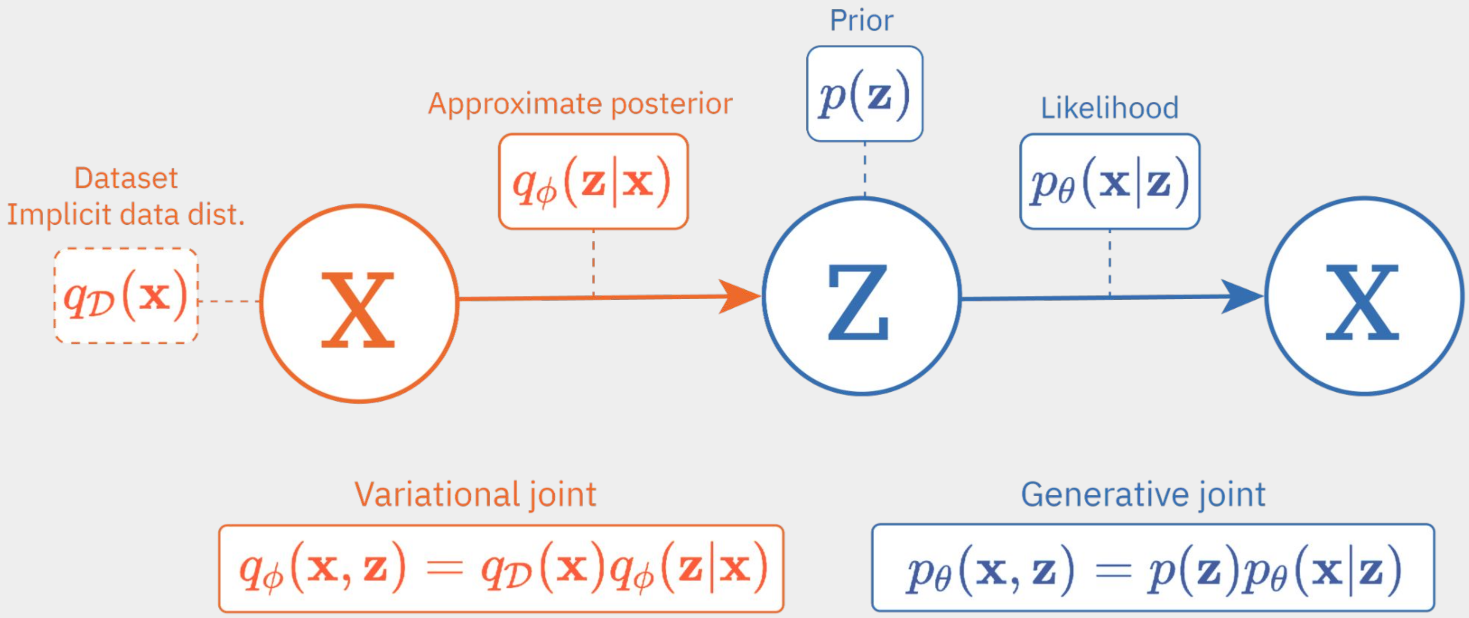
- Leverages **(deep) neural networks to model complex probability distributions** and efficiently sample *x* given *z* and *z* given *x*
  → *using stochastic gradient descent for optimisation*



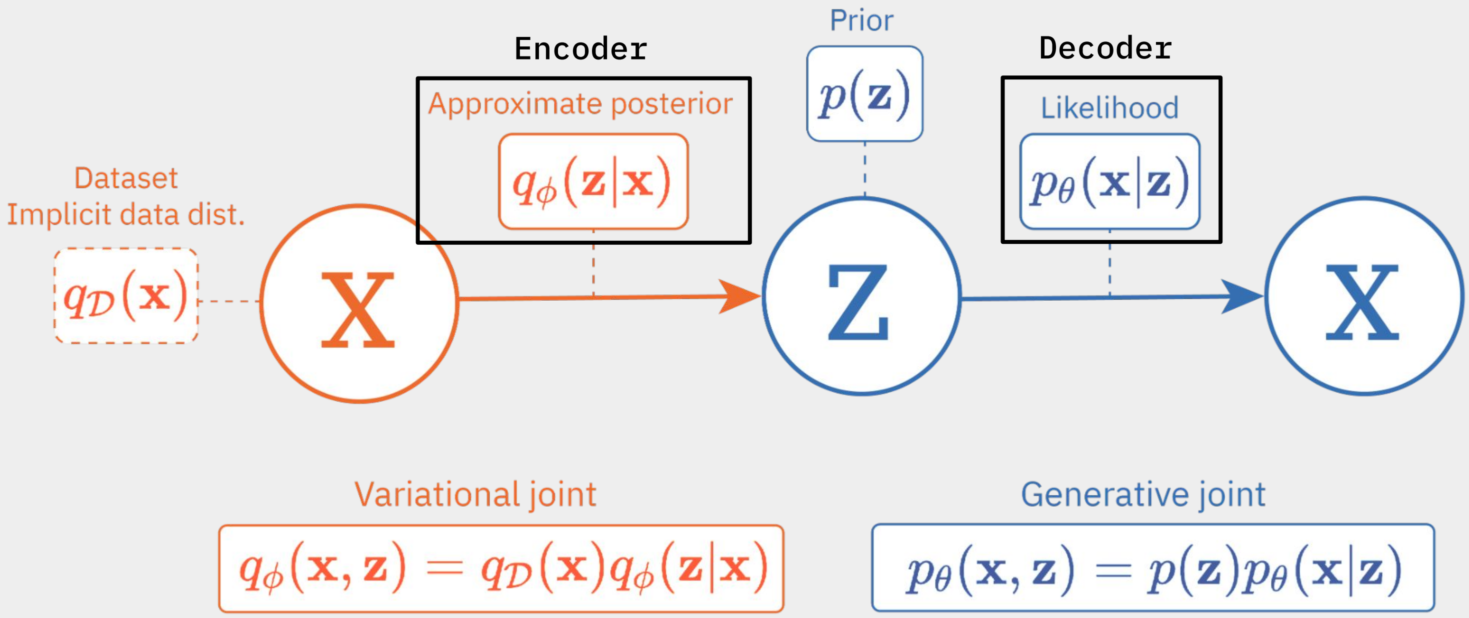
Generative joint

$$p_\theta(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$$

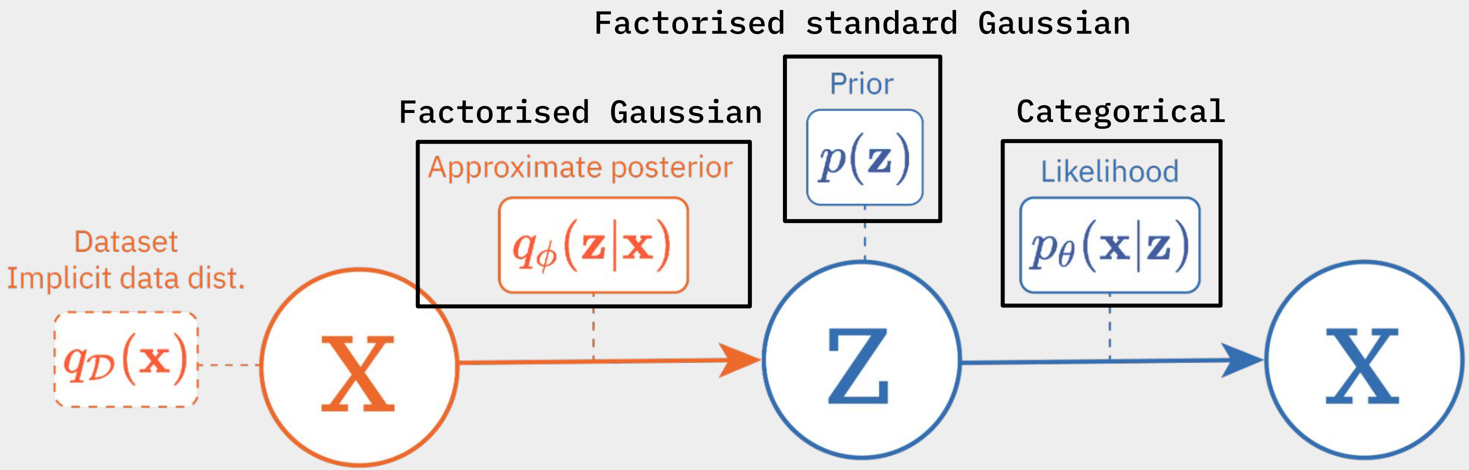
Variational Autoencoder[1] for modelling language



Variational Autoencoder[1] for modelling language



Variational Autoencoder[1] for modelling language



Variational joint

$$q_\phi(\mathbf{x}, \mathbf{z}) = q_\mathcal{D}(\mathbf{x}) q_\phi(\mathbf{z}|\mathbf{x})$$

Generative joint

$$p_\theta(\mathbf{x}, \mathbf{z}) = p(\mathbf{z}) p_\theta(\mathbf{x}|\mathbf{z})$$

Balancing two views of the joint between x and z

Balancing joints $$\boxed{\min_{\phi,\theta} D_{KL}\left(q_\phi(\mathbf{x},\mathbf{z})||p_\theta(\mathbf{x},\mathbf{z})\right)} \quad (2.7)$$

$$= \max_{\phi,\theta} -D_{KL}\left(q_\phi(\mathbf{x},\mathbf{z})||p_\theta(\mathbf{x},\mathbf{z})\right) \quad (2.8)$$

$$= \max_{\phi,\theta} \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}\left[\log q_\phi(\mathbf{z}|\mathbf{x})p_\mathcal{D}(\mathbf{x}) - \log p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})\right] \quad (2.9)$$

$$= \max_{\phi,\theta} \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}\left[p_\theta(\mathbf{x}|\mathbf{z})\right] - \mathbb{E}_{p_\mathcal{D}(\mathbf{x})}\left[D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right)\right] + \mathbb{E}_{p_\mathcal{D}(\mathbf{x})}\left[\log p_\mathcal{D}(\mathbf{x})\right] \quad (2.10)$$

$$\boxed{= \max_{\phi,\theta} \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z})}\left[p_\theta(\mathbf{x}|\mathbf{z})\right]} - \boxed{\mathbb{E}_{p_\mathcal{D}(\mathbf{x})}\left[D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right)\right]} + \text{constant} \quad (2.11)$$

$$\max_{\phi,\theta} \boxed{\mathbb{E}_{p_\mathcal{D}(\mathbf{x})}\left[\mathcal{L}_\mathbf{x}^{\text{ELBO}}(\phi,\theta)\right]}, \text{ where } \mathcal{L}_\mathbf{x}^{\text{ELBO}}(\phi,\theta) \leq \log p_\theta(\mathbf{x}) \quad (2.12)$$

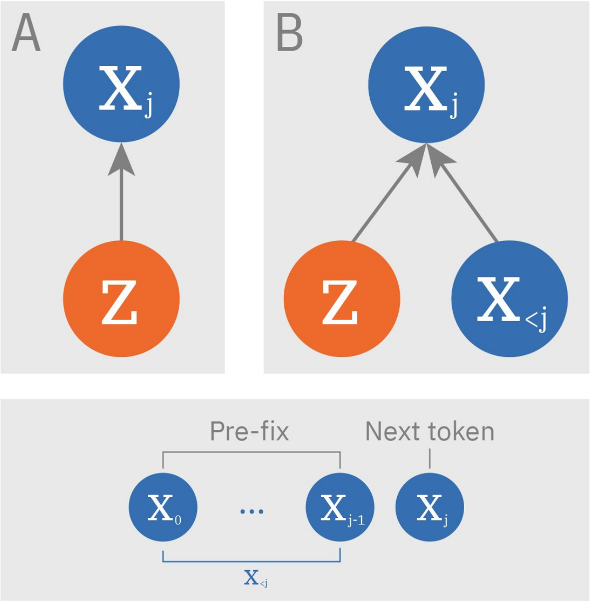
ELBO = lower bound on the model evidence
= negative reconstruction loss + KL divergence from prior to approximate posterior

expected ELBO = - (Distortion + Rate)

Posterior collapse & strong decoder problem

- Numerical goals of latent variable modelling are not inherently aligned with the qualitative goals of representation learning

- Many definitions of the generative model may model the data distribution, including ones that do not make use of the latent variables[1]

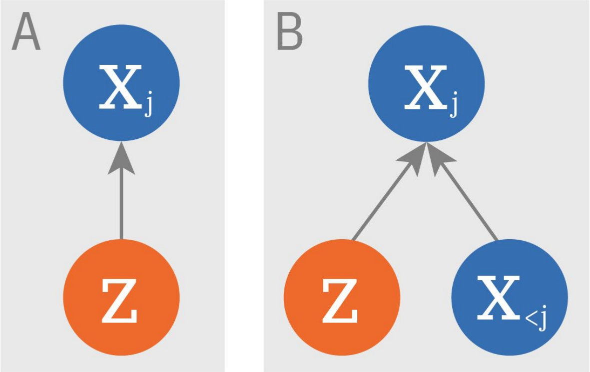


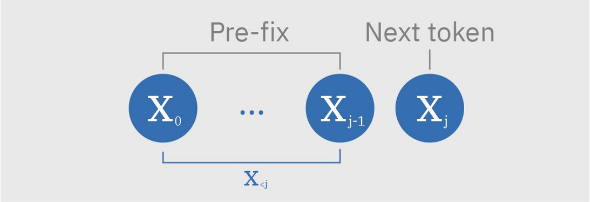
[1] Chen et al., 2017

Posterior collapse & strong decoder problem

Order of events

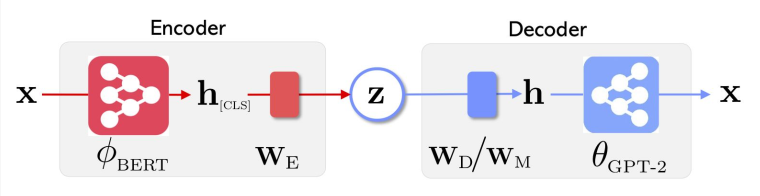
- When the decoder $p_\Theta(x|z)$ can use structure internal in x to model $p_D(x)$ it may approach $p_D(x)$ without making use of z
- As a consequence by laws of probability calculus:
    - The induced generative posterior $p_\Theta(z|x)$ collapses to the prior $p(z)$ as x and z are independent under p_Θ
- To trivially balance the two joints: $q_\phi(z|x)$ mimics this behaviour and collapses to the prior as well: x and z are independent under q_ϕ as well
- The KL from the prior to the approximate posterior, or the Rate, vanishes
    - Recall that ELBO = - (Distortion + Rate)
    - If ELBO can be minimised without making use of Rate, it will!
- **Strong decoder:** Auto-regressive language models are very well suited to model internal structure in x!





(Pre-trained) TransformerVAE[1]

Instantiating the neural networks in the VAE with large pre-trained Transformer networks



In this paper, we propose OPTIMUS, the first large-scale pre-trained deep latent variable models for natural language. OPTIMUS is pre-trained using the sentence-level (variational) auto-encoder objectives on large text corpus. This leads to a universal latent space to organize sentences (hence named OPTIMUS). OPTIMUS enjoys several favorable properties: (*i*) It combines the strengths of VAE, BERT and GPT, and supports both natural language understanding and generation tasks. (*ii*) Comparing to BERT, OPTIMUS learns a more structured

[1] Li et al., 2020

Summary so far

- Latent variable modelling makes for an interesting alternative perspective on representation learning
    - Global, designated stochastic representations that allow for smooth manipulation
    - Generalising outside of the data distribution
- **VAE** is a framework that prescribes a way to learn such a model at scale using deep neural networks
- Goals of latent variable modelling and representation learning are not inherently aligned: **posterior collapse**
- Especially in the case of **strong decoders** the latent variables may be ignored by the generative model
- A pre-trained TransformerVAE is a *very* powerful decoder VAE, so: how can we leverage that strength while also learning meaningful representations?

2 Problem statement & research goals

Problem statement & research goals

- Two lines of research collide
    1. using **latent variable models** for representation learning
    2. using **powerful density estimators**, such as large Transformer networks, to model language

- Li et al. (2020) showcase the performance of TransformerVAE along axes that are common in NLP
    - global statistics (such as perplexity values)
    - performance on downstream tasks
    - illustrative evidence in the form of text samples

- We aim to take a step back and assess the model as a latent variable model first
    - Setting the goals clear
    - Analysing along alternative axes of evaluation

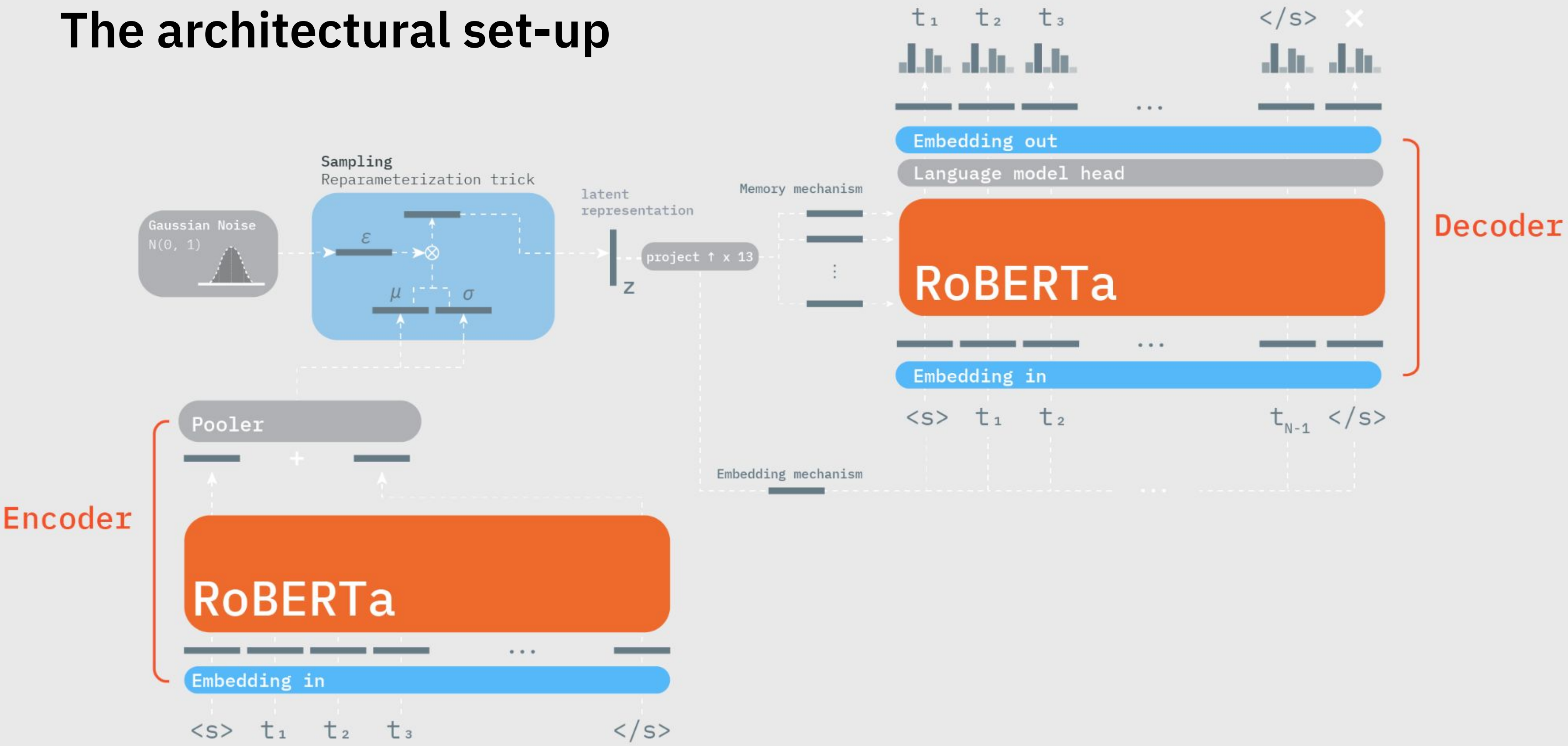
Setting our goals clear

Learning a generative latent variable model in the context of strong decoders such as large pre-trained Transformer networks

1. **Modelling the data distribution**, defining a generative model that explains the observed data well
2. Finding a (meaningful) **relationship between x and z**
3. Learning a statistically healthy model and performing **accurate approximate posterior inference**: $q(x, z) \sim p(x, z)$

→ Mostly goal 3 is underexposed in literature

The architectural set-up



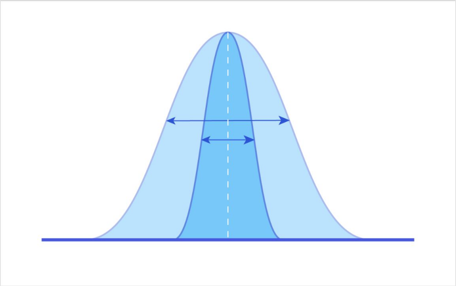
3 A bit more background & related work

Encoding information in a Gaussian VAE[1]

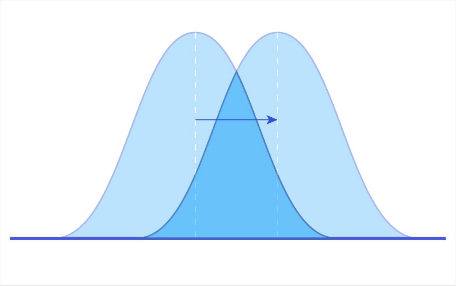
- Encoding information with factorised Gaussian approximate posterior distributions
  - Dispersing means
  - Decreasing variance

- Distinguishability comes at a *cost* expressed as rate
  - We want to learn smooth encodings, not point estimates as in autoencoders
  - We want data points that share features to overlap in their encodings to organise the latent space in an efficient way

Decreasing variance



Dispersing means

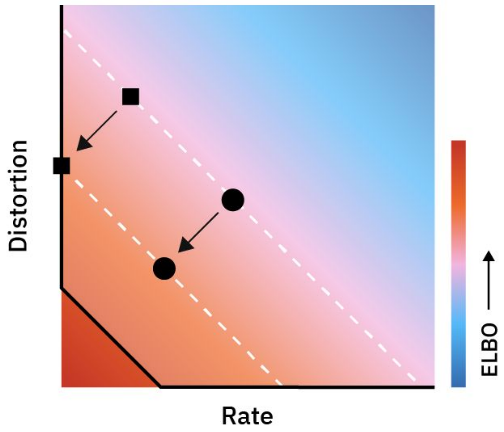


[1] Burgess et al. (2018)

Rate-distortion perspective on VAEs

- **Balancing compression & information retainment**
  - ELBO = - (R + D)
- **ELBO optimisation is not sensitive to rate-distortion trade-offs** and thus not directly to mutual information trade-offs[3]
- **Bits back decoding perspective[1]**
  - if the code coming from *q* is an inefficient one and the information can be modelled locally in *p* it will prefer to do so (information preference property[2])
- **Variational bounds on mutual information[3]**
  - $H - D \leqslant I_q(X; Z) \leqslant R$

[1] Chen et al. (2017)
[2] Zhao et al. (2019)
[3] Alemi et al. (2017)



Counteracting posterior collapse

- **Recall that a collapsed posterior is *p(z|x)* collapsing to *p(z)* due to *p(x|z)* not making use of *z***
  - A consequence, or symptom, is *q(z|x)* collapsing to *p(z)*
  - Can be diagnosed with observing vanishing KL term (vanishing rate)
- **Annealing the weight of the rate:** linear[1] or cyclical [2] schemas
- **Weakening the decoder:** drop-out of context at the decoder[2] or limiting receptive field
- **Richer priors:** mismatch by design[4, 5]
- **Information maximisation:** change the objective to directly maximise lower bound on mutual information[6, 7]

[1] Bowman et al. (2016)
[2] Fu et al. (2019)
[3] Chen et al. (2017)
[4] Pelsmaeker & Aziz (2019)
[5] Razavi et al. (2019)
[6] Zhao et al. (2017)
[7] Zhao et al. (2018)

Counteracting posterior collapse

Targeting a specific rate

Minimum Desired Rate[1]

Optimise ELBO with constraint on rate

$$\max_{\theta,\phi} \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathcal{L}_{\mathbf{x}}^{\text{ELBO}}(\phi,\theta)\right] \text{ s.t. } \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right)\right] \geq r$$

Free bits[2]

Cap the KL term if below a certain threshold

$$\mathcal{L}_{\mathbf{x}}^{\text{FB}}(\phi,\theta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] + \max\left(\lambda, D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right)\right)$$

We will focus on those two as they:

- Are direct in there purpose of achieving non-zero rate solutions (as opposed to for example richer priors or annealing)
- Simple in implementation and close to original ELBO formulation
- Used by Li et al. (2020) in the TransformerVAE, being successful at that
- Comparing these two: one may violate the ELBO by capping the KL term, can we observe differences?

[1] Pelsmaeker & Aziz (2019)
[2] Kingma et al. (2016)

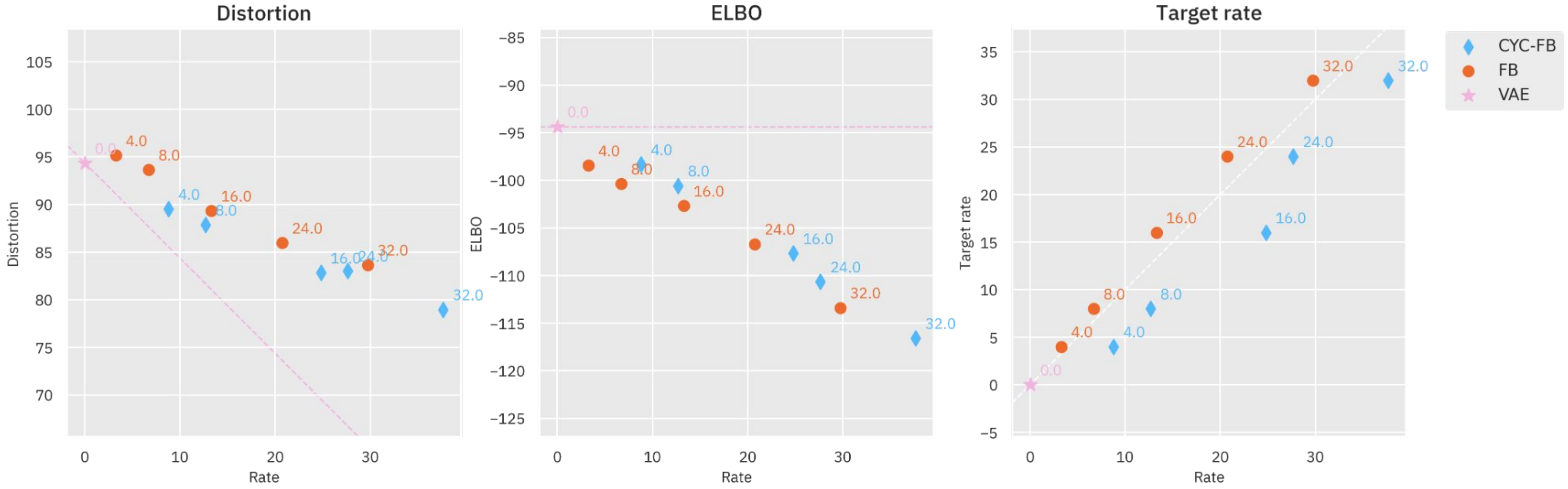
Summary so far

- The **cost of encoding** is expressed as rate
- **Rate upper bounds the mutual information** between *x* and *z* under *q*
- ELBO optimisation itself is not sensitive to rate-distortion trade-offs
- Amongst other techniques, we can try to **target a specific rate** to achieve non-zero mutual information between *x* and *z*
- But, if we manage to obtain non-zero rate solutions, what happens with the **quality of our approximate inference**?
    - Can we say something about the balance between p(x, z) and q(x, z)?
    - How is the latent space organised?

4 Extending the information theoretic view to account for approximate inference

Inefficient encoding

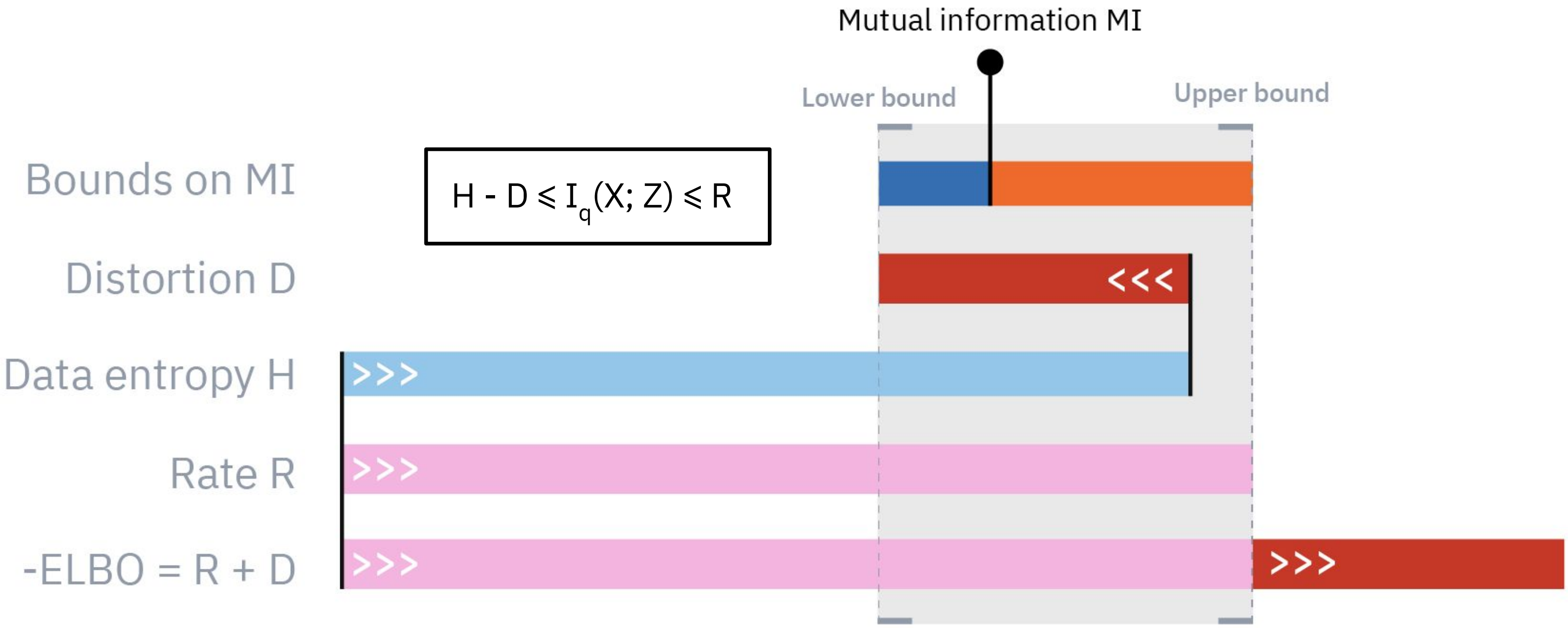
- The information preference property at work: we can not ‘trade’ distortion for rate
- Higher target rates result in worse / lower ELBO values

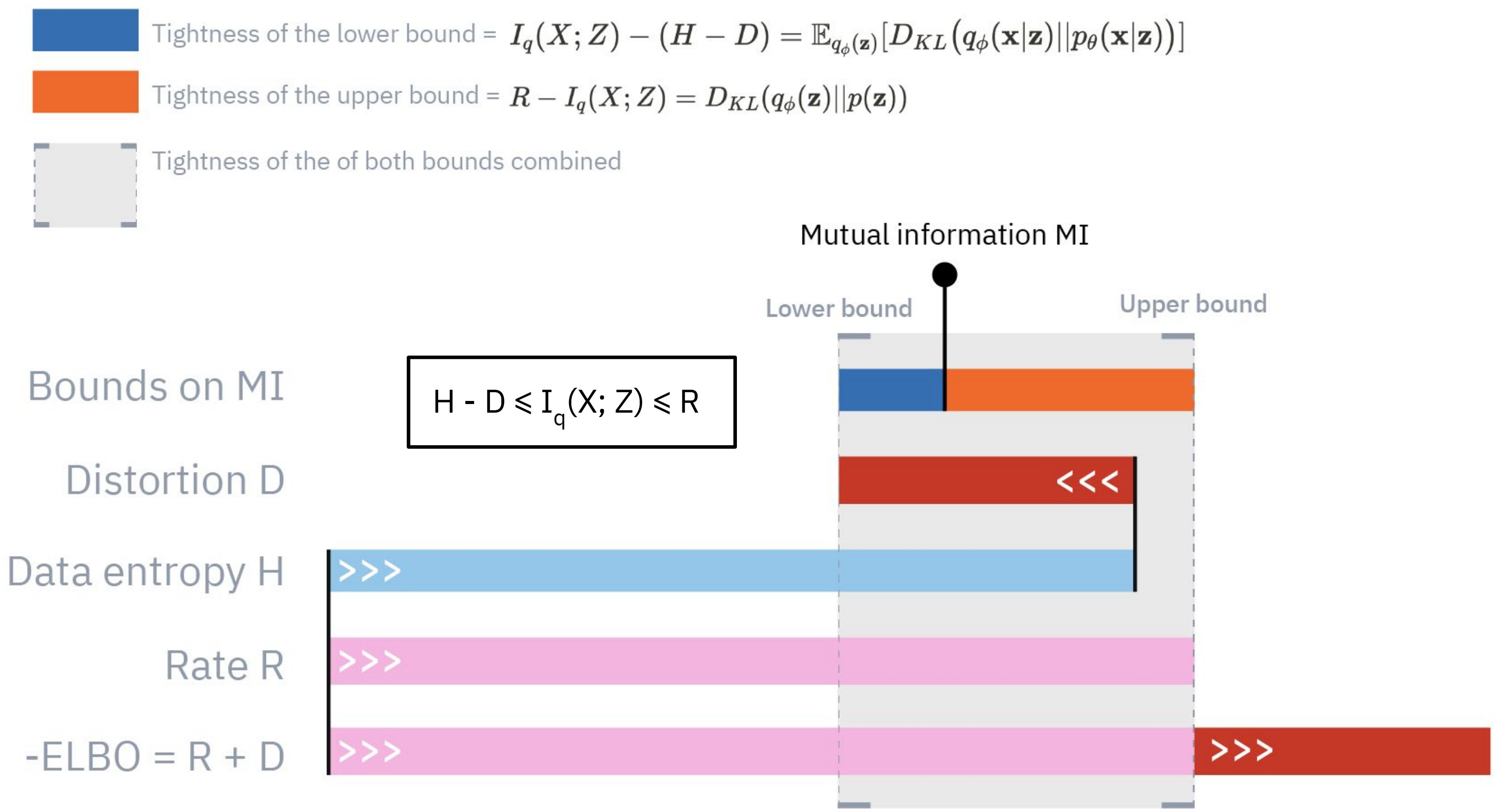


Exploring the variational bounds on mutual information

→ We know MI is relevant to representation learning, but what else is there?

$$H - D \leqslant I_q(X; Z) \leqslant R$$

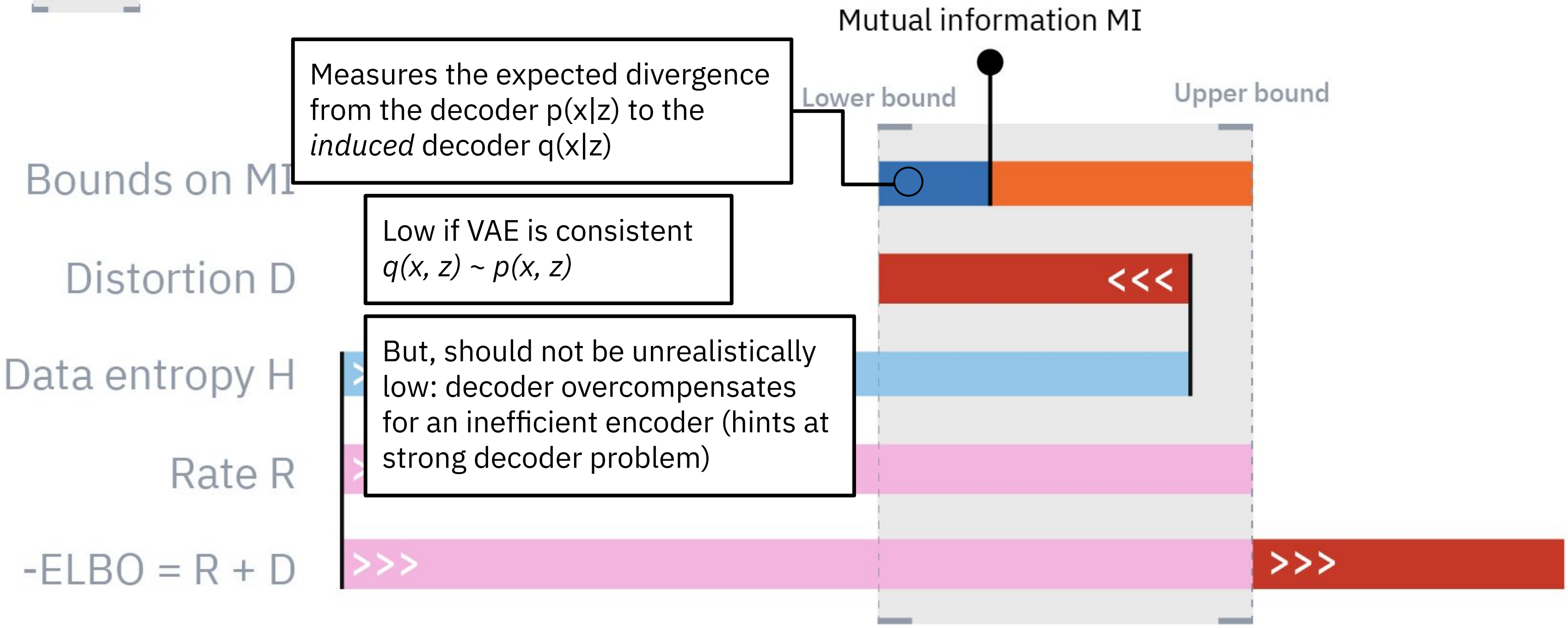



Tightness of the lower bound = $I_q(X;Z) - (H - D) = \mathbb{E}_{q_\phi(\mathbf{z})}[D_{KL}\left(q_\phi(\mathbf{x}|\mathbf{z})||p_\theta(\mathbf{x}|\mathbf{z})\right)]$
Tightness of the upper bound = $R - I_q(X;Z) = D_{KL}(q_\phi(\mathbf{z})||p(\mathbf{z}))$
Tightness of the of both bounds combined
Mutual information MI
Lower bound
Upper bound
Bounds on MI
$H - D \leqslant I_q(X; Z) \leqslant R$
Distortion D
Data entropy H
Rate R
-ELBO = R + D

Tightness of the lower bound = $I_q(X;Z) - (H - D) = \mathbb{E}_{q_\phi(\mathbf{z})}[D_{KL}(q_\phi(\mathbf{x}|\mathbf{z})||p_\theta(\mathbf{x}|\mathbf{z}))]$

Tightness of the upper bound = $R - I_q(X;Z) = D_{KL}(q_\phi(\mathbf{z})||p(\mathbf{z}))$

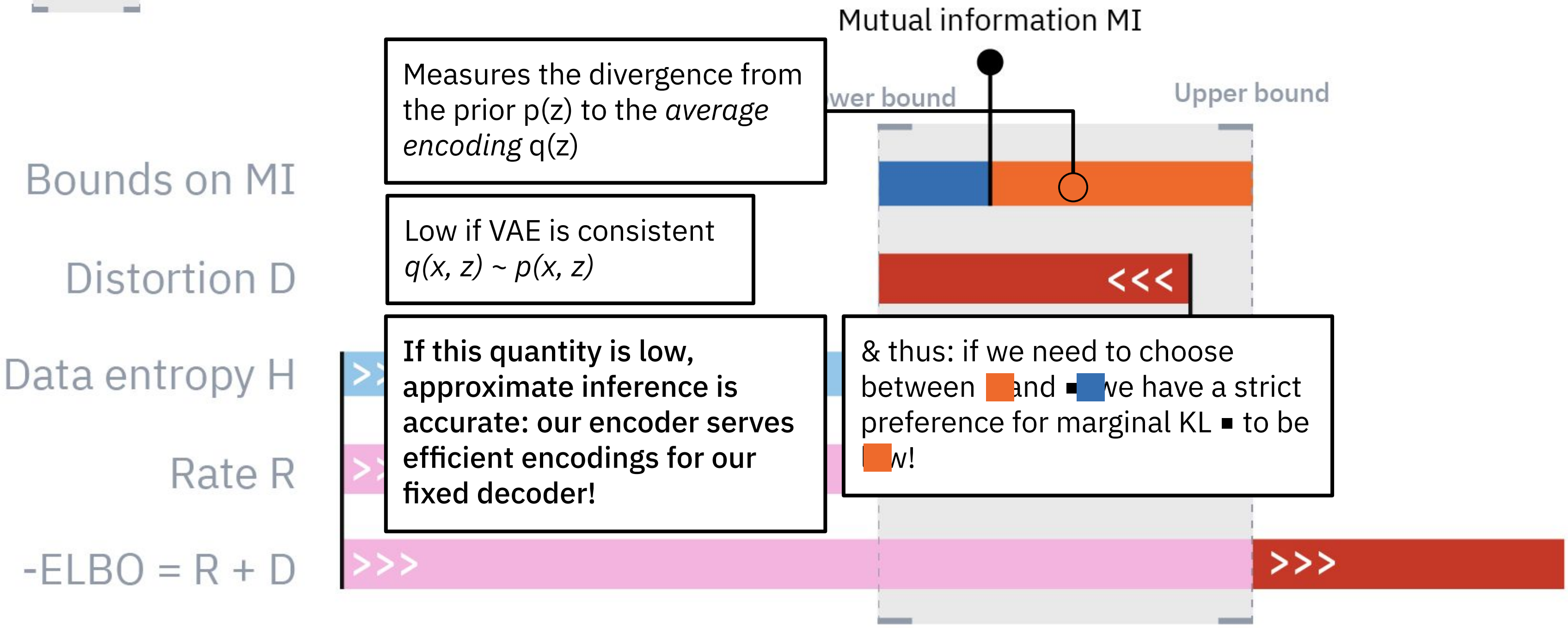
Tightness of the of both bounds combined



Tightness of the lower bound = $I_q(X;Z) - (H - D) = \mathbb{E}_{q_\phi(\mathbf{z})}[D_{KL}(q_\phi(\mathbf{x}|\mathbf{z})||p_\theta(\mathbf{x}|\mathbf{z}))]$

Tightness of the upper bound = $R - I_q(X;Z) = D_{KL}(q_\phi(\mathbf{z})||p(\mathbf{z}))$

Tightness of the of both bounds combined



Analysing optimisation techniques w.r.t. marginal KL

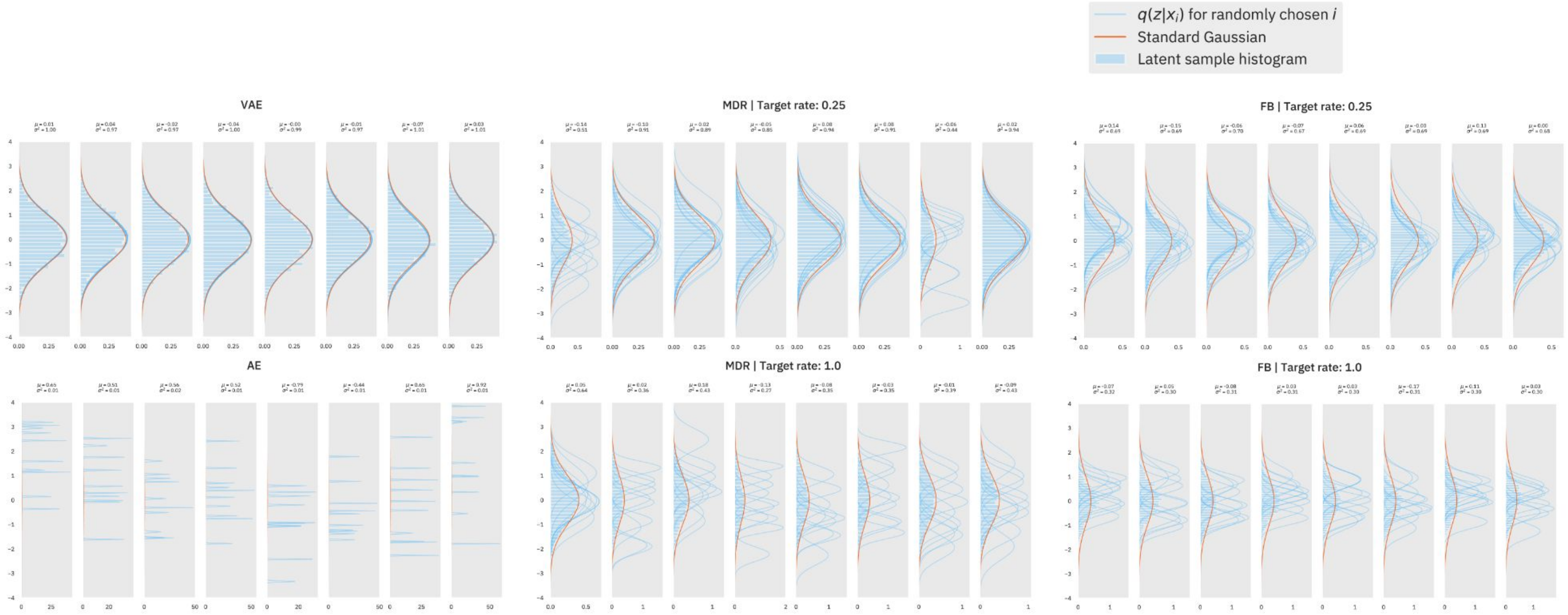
Marginal KL $$D_{KL}\left[q_\phi(\mathbf{z})||p(\mathbf{z})\right]$$

- Intractable to compute
- Computationally expensive to estimate
- Bounded in estimation

For analysis purposes, we want to compare sample groups from $q_\phi(z)$ and $p(z)$, which we can obtain by ancestral sampling and by sampling from the known prior

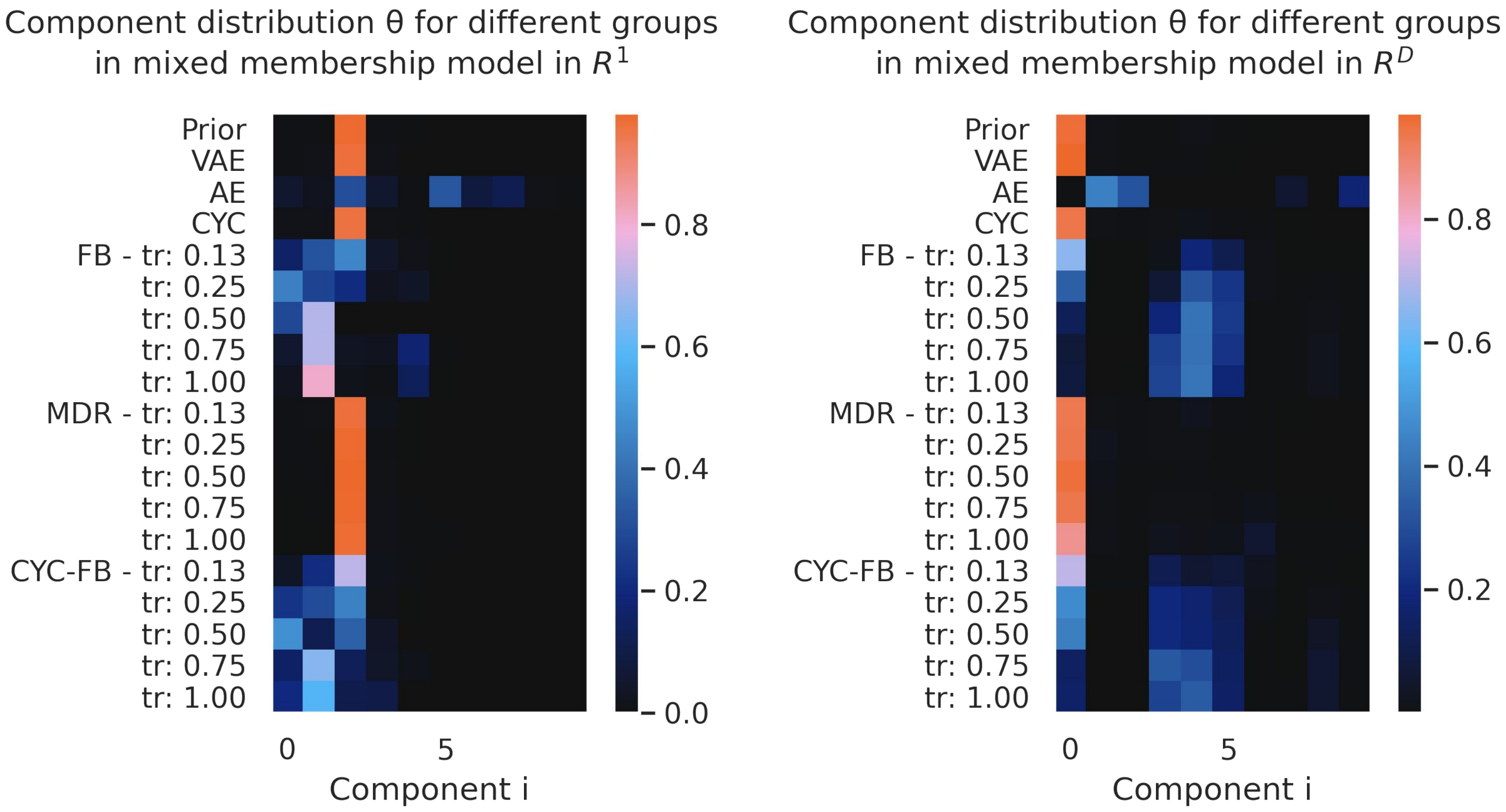
We turned to: Statistical tests (e.g. MMD), visual inspection and Mixed Membership Model

Visual differences between MDR & Free bits optimised latent spaces

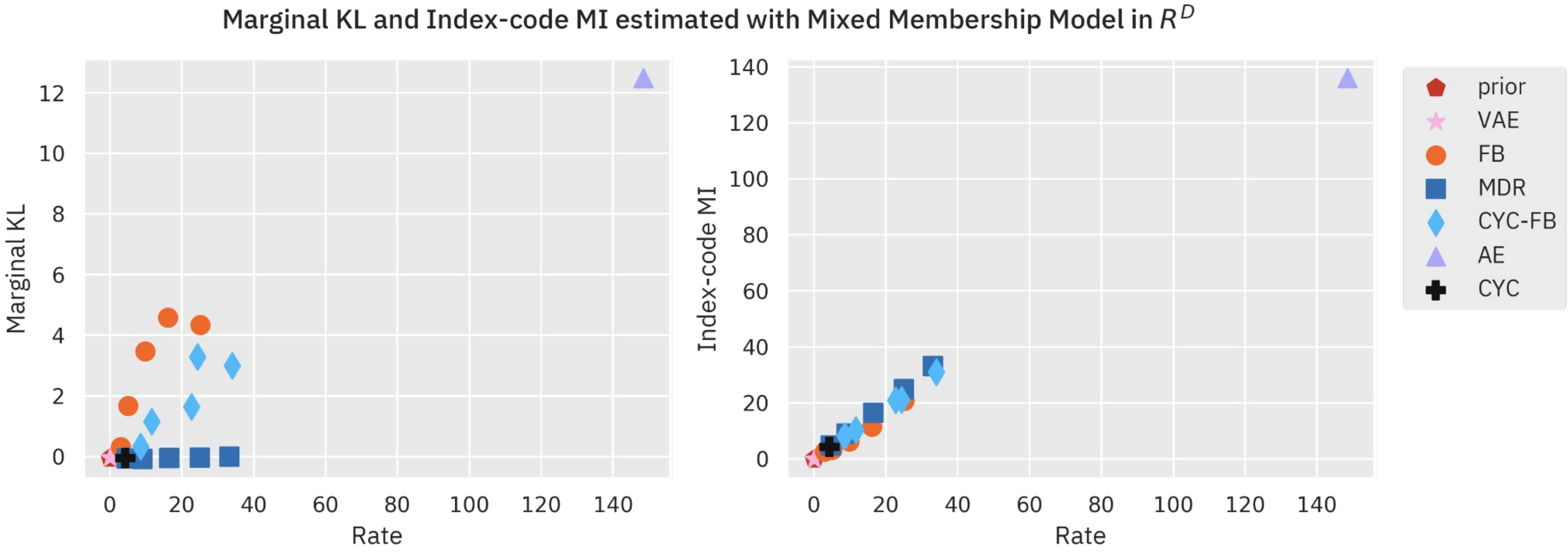


→ Plotting 8 out of 32 dimensions

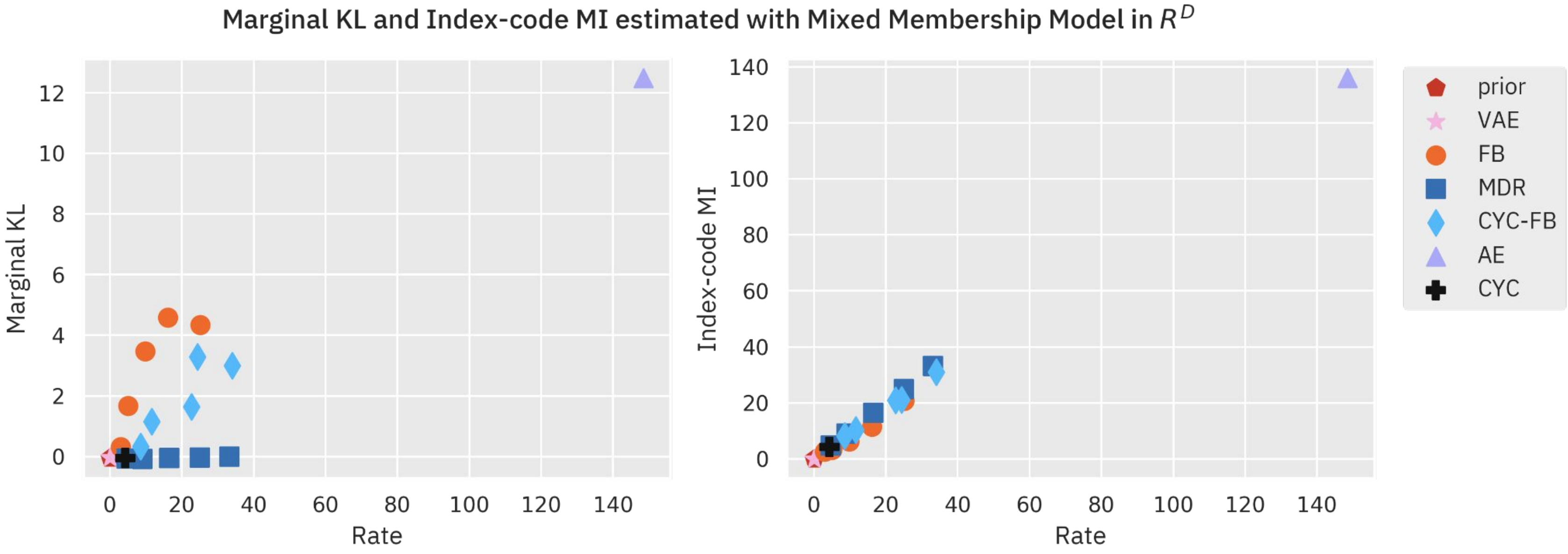
Analysing optimisation techniques w.r.t. marginal KL



Analysing optimisation techniques w.r.t. marginal KL



Analysing optimisation techniques w.r.t. marginal KL



→ For some techniques (Free bits), increased MI (relevant to representation learning) comes at the cost of increased marginal KL (compromised approximate posterior inference)

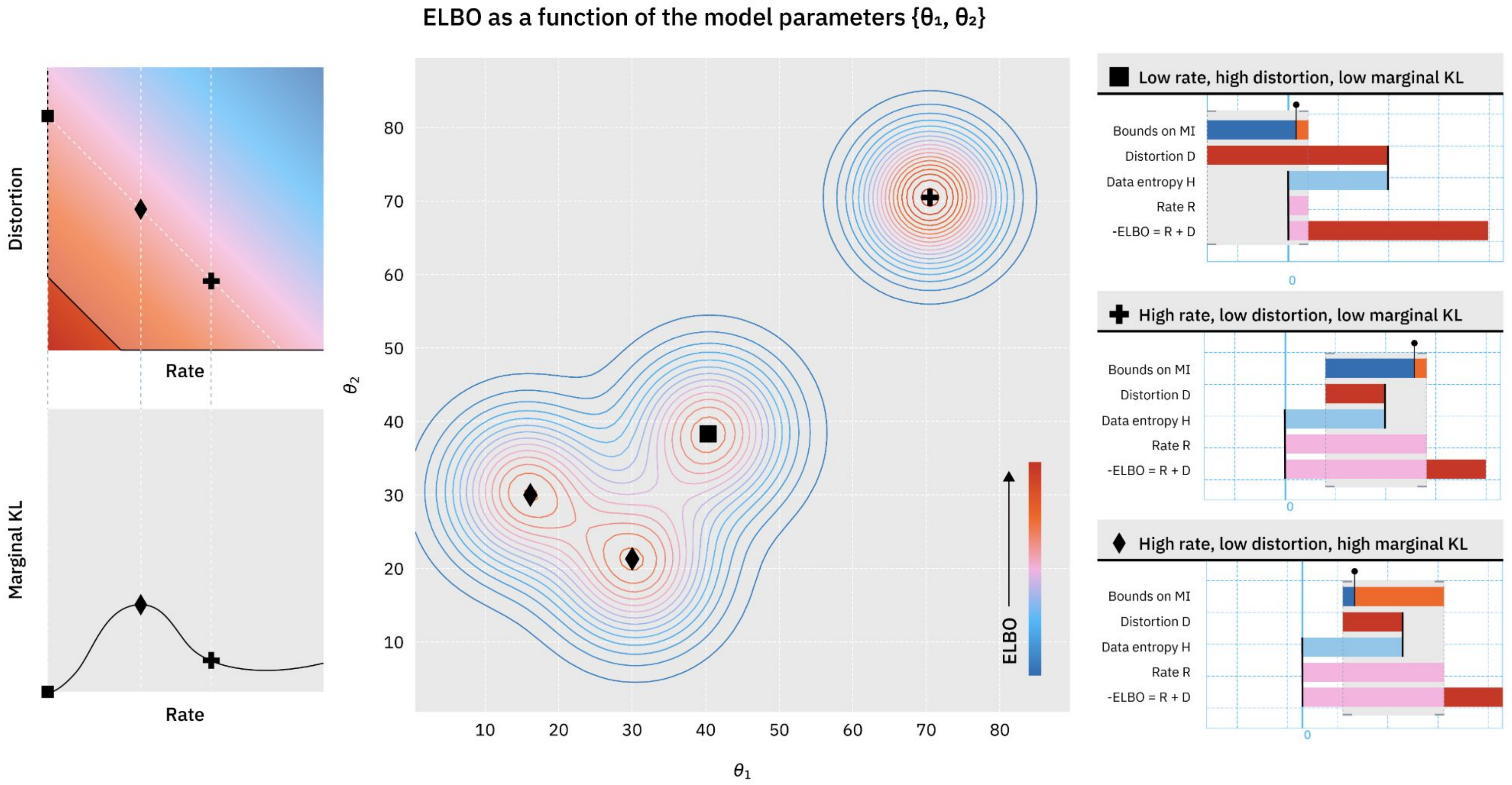
Potential optimisation pathological directions

We knew already that Rate-distortion trade-off quantitatively gives an indication of qualitative goals of representation learning

We now also know that Bounds on MI trade-off quantitatively gives an indication of qualitative goals of variational inference

And that ELBO insensitive *not only* to rate-distortion trade-offs *but also* to relative magnitude of marginal KL

Potential pathological directions during ELBO optimisation



5 Conclusion

Conclusion

- **Setting the goal of this field straight:**
  - learning statistically healthy latent variable models in the context of powerful density estimators such as large transformer networks
  - Interested in learning a meaningful latent space, approximate posterior inference is important!
- **Adopt other modes of evaluation than benchmark tasks, even visual inspection of latent space can show a lot, mixed membership model is also insightful**
- **Marginal KL is an alternative quantity of interest that is potentially more important to watch than mutual information**
- **Breaking the ELBO has serious consequences, MDR is a very suitable alternative to Free bits**

Many thanks to David Stap, Wilker Aziz, Joris Baan, Lucas de Haas & Vlad Niculae!