We see that, for non-trivial latent spaces, term 2 (the mutual information between *z* we see that, for non-trivial fatelly spaces, term λ (the mutual information between λ and n) approaches its maximum value of log N. We also see that term 3 (the marginal KL) makes a significant contribution to the ELBO, confirming that a simple encoder-decoder model has a hard time matching the marginal $q(z)$ to $p(z)$. aximum value of log N. We also s Ω , *nanes* a *pignificant* controduon to the *DDD z*(*z*) decoder inouer has a nara three materials the line *^q*(*n*)*q*(*z*)] denotes the mutual information of *n* and *z* in *q*(*n, z*). To KL(*q*(*zⁿ [|] ^xn*) ^k *^p*(*zn*)) = ^X *n p*(*n*) makes a significant contribution

Term 2: This is the (negative) mutual information between *z* and the index *n*. It penalizes models in which we can determine which observations *x* are consistent with which *z* vectors. to overlap for distinct over \mathcal{L} over \mathcal{L} is the reconstructions \mathcal{L} over \mathcal{L} to the reconstructions \mathcal{L} over \mathcal{L} is the reconstructions of \mathcal{L} is defined as \mathcal{L} is defined as $\mathcal{$ $\text{I}\text{C}\text{I}\text{II}\text{I}$. Interesting the largest and below and below and below and below. $\frac{1}{2}$ M^{ch} validing λ are consistent $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

Term 3: This is the (negative) KL divergence between the average encoding distribution and the prior. $\frac{1}{2}$ average encounts

Observations 3 Qualitative perspectives of the perspectives of the perspectives of the perspectives of the perspectives of

The three terms above encode three desiderata:

Existing Perspectives on the ELBO Graphical Model: z_n \rightarrow (x_n) *N* We're interested in variational EM in models of the form $p_{\theta}(\boldsymbol{x}) = \int p_{\theta}(\boldsymbol{x} \,|\, \boldsymbol{z}) p(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}.$ Fit by maximizing log evidence lower bound (ELBO) *L*, $\log p_\theta(\boldsymbol{x}) = \log \int q_\phi(\boldsymbol{z} \,|\, \boldsymbol{x})$ $p_{\theta}(\bm{z}, \bm{x})$ $q_{\boldsymbol{\phi}}(\boldsymbol{z} \,|\, \boldsymbol{x})$ $\mathrm{d}\boldsymbol{z}\geq\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}\,|\, \boldsymbol{x})}\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{z},\boldsymbol{x})}{q_{\perp}(\boldsymbol{z}\,|\, \boldsymbol{x})}$ $q_{\boldsymbol{\phi}}(\boldsymbol{z} \,|\, \boldsymbol{x})$ \triangleq $\mathcal{L}(\theta, \phi)$. $p(\boldsymbol{z}) = \prod$ *N n*=1 $p(z_n), \qquad p_\theta(\bm{x}\,\vert\,\bm{z}) = \prod$ *N n*=1 $p_{\theta}(x_n | z_n)$ $q_{\boldsymbol{\phi}}(\boldsymbol{z} \,|\, \boldsymbol{x}) = \prod$ *N n*=1 $q_{\boldsymbol{\phi}}(z_n \,|\, x_n).$

ELBO Surgery: Yet another way to carve up the evidence lower bound Mathew D. Hofman (Adobe Research); Mathew J. Johnson (Google Brain)

Setup

• When term 3 is large, our choice of prior $p(z)$ is regularizing our model (whether or not we want it to). After optimization, we estimated the marginal KL term 3 via Monte Carlo: *K K p*(*z*) *k p*(*z*) *k p*(*z*) \mathbf{A} *S S p*(ˆ*zs*) *, z*ˆ*^s | n*ˆ*^s* ⇠ *q*(*z | xn*ˆ*^s*)*, n*ˆ*^s* ⇠ Unif(*{*1*,* 2*,...,N}*)*,* (19)

check this expression, write the check $p(n) \triangleq \frac{1}{n}$ Γ check this expression, we have N $\frac{1}{N},$ $\frac{1}{N}$.

$$
\mathcal{L}(\theta, \phi) = \left[\frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{q(z_n | x_n)} [\log p(x_n | z_n)] \right] - \underbrace{(\log N)}_{\textcircled{1}}
$$
\n(1)

\n(2)

Term 1: This is a traditional autoencoder objective, which encourages accurate reconstructions of *x* given *z*. α accurates α *q*(*n | z*) at each evaluation requires accessing all *N* observations (and the normalization also precludes

This decomposition sheds some new light on what the ELBO cares about: with sacrificial power by simply defining the prior to be \mathcal{L} . This choice would not be \mathcal{L} To get a sense for the new terms in (17), we fit a basic variational autoencoder to a binarized MNIST

follows from using *p*(*n*) = ¹

 $=$ KL($q(z)$ || $p(z)$) + $\mathbb{I}_{q(n,z)}[n,z]$ *,*

I*q*(*n,z*)[*n, z*] = E*q*(*z*)

1

N

• Terms 1 and 2 are in tension. To make term 1 large, we want *z* to tell us almost everything there is to know about *x*. But that often requires that *n* and *z* have high mutual information. $\frac{1}{2} \int_0^{\infty} \frac{1}{2} \, dx$ *a q i and <i>z* are in wission. To make with I large, we want *z* w will us all *no* all *nose* high Every uning mere is to know about x . But that onen requires that n and z have mgi $\frac{1}{2}$ k $\frac{1$ 1 X *S*

• Term 2 is bounded above and below: $0 \le \log N -$ In the case where reconstructions are very precise, we should expect term 2 to be near its maximum value of log N. $0 \leq \log N - \mathbb{E}_{q(z)}\mathbb{H}[q(n \,|\, z)] \leq \log N$

• $p(z)$ only appears in term 3, and term 3 can in principle be set to 0 for any model by setting $p(z) = q(z)$. This may be impractical or unwise, but it does imply that... T sense for the new terms in (17), we fit a basic variation α binarized MNIST α binarized MNIST α P $p(z)$ only appears in term 3, and term 3 can in principle be set to 0 for any mode

[1] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *International* Conference on Learning Representations (2014) *Conference on Learning Representations* (2014).

. Shakir Mohamed, and Daan Wierstra. "Stochastic backpropagation and approximate included approximational approximation of the entired in deep generative models". In: *International Conference on Machine*

, and Ruslan Salakhutdinov. "Importance weighted autoencoders". In: *In Learning Representations* (2016).

Duvenaud, Alex Wiltschko, Sandeep Datta, and Ryan Adams. "Comand neural networks for structured representations and fast inference". Interpretation Processing Systems 29. 2016.

Imy Ba. "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations*. 2015.

hen Shlens, Navdeep Jaitly, and Ian Goodfellow. "Adversarial Autoen-Conference on Learning Representations. 2016.

 $p(n, z)$ *p*(*n, z*)

 $=$ KL($q(z)$ || $p(z)$) + $\mathbb{E}_{q(z)}$ [KL($q(n|z)$ || $p(n)$)] \mathcal{L} 1 average reconstruction $=$ KL($q(z)$ || $p(z)$) + E_{$q(z)$}[KL($q(n | z)$ || $p(n)$)]

 $\log N - \mathbb{E}_{q(z)}[\mathbb{H}[q(n|z)]]) - \text{KL}(q(z) \| p(z))$.

(17)
<mark>(17)</mark>
(17)

where the first equality can be checked by expanding $p(n, z)$ and $q(n, z)$ $I^{(n)}$ and $I^{(n)}$ identices, the log and the last line lors, the second equal: splitting the log, and the last line follows from using $p(n) = \frac{1}{N}$. $\frac{dy}{dx}$ ionows from the endm rule and
 $\frac{dy}{dx}$ where the first equality can be checked by expanding $p(n, z)$ and $q(n, z)$ and canceling the *p*(*n*) and *q*(*n*) factors, the second equality follows from the chain rule and

Substituting this new KL expression into the ELBO (1) , we have

nonlinearities, and we fit them using the Adam optimizer [5]. For more details, see the code.

Google brain

Food for Thought and 2 are in tension with each other because to get a good average reconstruction score for 1 , we have for 1 typically need each encoding *zⁿ* to be specific to its corresponding observation *xⁿ* and hence *q*(*n | z*) should have low entropy. Term 2 acts as a regularizer, in that it encourages the encodings *q*(*z | xn*) $\frac{1}{N}$. The canonical observations about the ELBO expression given in (17). First, the two terms Γ \overline{a} are in tension with each other because to get a good average reconstruction score for 1 , we see that \overline{a} typically need to be specific to be specific to its corresponding observation \mathbf{z} and \mathbf{z} and \mathbf{z} for sample indices *s* = 1*,* 2*,...,S*, which requires total time proportional to *NS* to compute. We P estimated the mutual information term 2 by subtraction. As shown in Table 1, while the marginal KL term 3 could in principal in principal ϵ to be very small, it still contributes the significant lyingites the sign

want z to tell us almost ⇠ Unif(*{*1*,* 2*,...,N}*)*,* (19) that can "meet in the middle" with the middle "meet" with the middle "middle" with the encoder and decoder net
The encoder networks. The encoder networks and decoder networks. The encoder networks are also and decoder net

Emphasizes that, unlike maximum a posteriori (MAP), a good posterior approximation must not only assign its probability mass to regions of low energy (high joint probability) but also try to maximize the entropy of $q_{\phi}(z | x)$.

E*q*(*ⁿ [|] ^z*)

E*q*(*zⁿ [|] ^xn*)[log *p*(*xⁿ | zn*)]

 $\overbrace{}^{x }$ θ index-code mutual info.

 $\overbrace{\hspace{2.5cm}}^{2}$ $3)$ marginal KL to prior

We fit basic variational autoencoder models with 2, 10, and 20 latent dimensions to a binarized 60000-image MNIST dataset, computed the average KL divergence from eq. 7, estimated term 3 by Monte Carlo, and estimated term 2 by subtracting \vert our estimate of term 2 from the average KL. The results are plotted below.

$$
\frac{1}{N} \sum_{n=1}^{N} \text{KL}(q(z_n | x_n) \| p(z_n)) = \sum_{n} q(n, z) \log \frac{q(n, z)}{p(n, z)} \n= \text{KL}(q(z) \| p(z)) + \mathbb{E}_{q(z)}[\text{KL}(q(n | z) \| p(n))] \n= \text{KL}(q(z) \| p(z)) + (\log N - \mathbb{E}_{q(z)}[\mathbb{H}[q(n | z)]]),
$$

log *^q*(ˆ*zs*)

p(ˆ*zs*)

also computed the average KL ¹ *N*

 $\left\{ \begin{array}{c} |z| \leq |z| \leq |z| \leq |z| \leq |z| \end{array} \right\}$ The results above suggest that deep latent Gaussian models have a hard time $\frac{127117}{200}$ $\frac{12717}{200}$ $\frac{12717}{200}$ $\frac{12717}{200}$ producing unimodal marginal posteriors. Perhaps we should investigate learning multimodal priors for $p(z)$ that meet $q(z)$ halfway? to overlap for distinct observations *n*, but this effect is likely to be weak relative to the reconstruction Empirically, we have found that reconstructions are very precise and, correspondingly, *q*(*z | n*) is 0 log *N* E*q*(*z*)H[*q*(*n | z*)] log *N.* (18) These results confirm that our current encoder and decoder models (and optimizers) find it difficult $t(x)$) – $KL(q(z) || p(z))$. \blacksquare The results above suggest that deep fatelly to The results above succest that deen latent Gaussian models have a hard tin ine results above suggest that accp fatent classian inouchs have a hard this roqueing unimodal marginal posteriors. Perhaps we should i

• We could set $p(z) = q(z)$, but this choice is computationally impractical for large datasets, and may overfit badly. What's the right level of power for $p(z)$? very concentrated relative to *q*(*z*), resulting in 2 is close to its maximum value of log *N*. $\frac{1}{\sqrt{2}}$ appears in a property $\frac{1}{\sqrt{2}}$ on $\frac{1}{\sqrt{2}}$. The $\frac{1}{\sqrt{2}}$ the *m*ight level of new constant priors **prior priors** α **priors** α , only the ELBO, α is a first set α is the α is β to α is β datasets and may overfit hadly. What's the right level of nower for $p(z)$ Second, while *q*(*z*) appears in all terms, *p*(*z*) only appears in 3 . Thus when considering choosing We could set $p(z) = q(z)$, but this choice is computationally impractical for large a aldscis, and may over the baday. What state upsite it very believer for $p(z)$.

• DLGMs are powerful density estimators. Shouldn't a deeper DLGM be able to $\text{match } q(z)?$ without sacrificing model power by simply defining the prior to be prior to be $q(\lambda)$. This choice would not be $q(\lambda)$ priors *p pl pl<i>c***Ms** are nowerful density estimators. Showever the α without sacrificing model power by simply defining the prior to be *q*(*z*). This choice would not be and the scalable computation $q(z)$.

• This analysis also applies to the non-variational case where $q(z | n) = p(z | x_n)$. ndex *n*. It \vert What can this analysis tell us about latent-variable density estimation in general? $\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1}{\sqrt{1-\frac{1$ $\frac{1}{2}$ can interest in principal because $\frac{1}{2}$ in the set to $\frac{1}{2}$ in the strong it is large it in the strong strong it is large it in the strong strong it is large it in the strong indicate it is large in the s is from many managers in the mon-variation. Setting 1 to 2 this analysis also applies to the non-variation. to the index n It \vert index \vert what can this analysis tell us about latent variable This analysis also annlies to the non-variational case where $a(z \mid n) = p(z \mid r)$ the data distribution distribution. In the absorption of $q(z \mid w)$ into $q(z \mid w)$ Vhat can this analysis tell us about latent-variable density estimation in general?

> • Do the marginal posteriors of classical models (e.g., factor analysis) and more powerful flat models (e.g., mixtures of factor analyzers, latent Dirichlet allocation) \int do a better or worse job of matching their priors? do a better of worse job of materials then priors. Γ the morning is patentes of elegated models (e.g. fector emplared end me D_0 the marginal posteriors of classical models (e.g., factor analysis) and more o a better or worse job of matching their priors? σ and the evaluate σ may be caused by a rigid prior that the encoder and decoder an

p(*z*)

*q*avg

 $\frac{avg}{\phi}(z)$

Indeed, unlike $KL(q(z_n | x_n) || p(z_n))$, the marginal KL $KL(q_{\phi}^{\text{avg}}(z) || p(z))$ can in principle be made arbitrarily small, with $q_{\phi}^{\text{avg}}(z) \approx p(z)$.

Consider the average encoding distribution,

$$
q_{\phi}^{\text{avg}}(z) \triangleq \frac{1}{N} \sum_{n=1}^{N} q_{\phi}(z \mid x_n).
$$

We would expect that if $x_n \sim p_\theta(x)$ and $q_\phi(z | x_n) \approx p_\theta(z | x_n)$, then for large N,

$$
p(z) = \int p_{\theta}(z \mid x) p_{\theta}(x) dx = \mathbb{E}_{x \sim p_{\theta}(x)} p_{\theta}(z \mid x) \approx \frac{1}{N} \sum_{n} p_{\theta}(z \mid x_n) \approx \frac{1}{N} \sum_{n} q_{\phi}(z \mid x_n).
$$

In practice, however, there may be large gaps or holes in the average encoding distribution. The cartoon at right shows one scenario where this could happen. As the latent dimension grows, we should expect it to get harder to fill up a large spherical space with many small blobs, for the same reason kernel density estimation scales poorly with dimension.

A New Rewrite of the ELBO An Experiment

The average encoding distribution is hidden in the ELBO. To simplify notation, treat the index *n* as a random variable and define

$$
q(n,z) \triangleq q(n)q(z \mid n), \qquad q(z \mid n) \triangleq q(z \mid x_n), \qquad q(n) \triangleq \frac{1}{N}
$$

$$
p(n,z) \triangleq p(n)p(z | n), \qquad p(z | n) \triangleq p(z), \qquad p(n) \triangleq \frac{1}{N}
$$

Note that $q^{\text{avg}}(z) = \sum_{n=1}^{N} q(z, n)$.

Evidence minus posterior KL

$$
\mathcal{L}(\theta, \phi) = \log p_{\theta}(\boldsymbol{x}) - \mathrm{KL}(q_{\phi}(\boldsymbol{z} \,|\, \boldsymbol{x}) \, \| \, p_{\theta}(\boldsymbol{z} \,|\, \boldsymbol{x}))
$$

Emphasizes that ELBO is a lower bound that becomes tighter as the variational distribution better approximates the posterior.

Average negative energy plus entropy

$$
\mathcal{L}(\theta, \phi) = \mathbb{E}_{q_{\phi}(\boldsymbol{z} \mid \boldsymbol{x})}[\log p_{\theta}(\boldsymbol{z}, \boldsymbol{x})] + \mathbb{H}[q_{\phi}(\boldsymbol{z} \mid \boldsymbol{x})]
$$

Average term-by-term reconstruction minus KL to prior

$$
\mathcal{L}(\theta,\phi) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{q_{\phi}(z_n | x_n)} [\log p_{\theta}(x_n | z_n)] - \text{KL}(q_{\phi}(z_n | x_n) \| p(z_n)) \tag{1}
$$

Emphasizes that the ELBO has an autoencoder's average reconstruction term as well as a KL divergence from each encoding distribution to the prior.

The Average Encoding Distribution

We can rewrite the KL to the prior in (1) as

$$
\frac{1}{N} \sum_{n=1}^{N} \text{KL}(q(z_n | x_n) \| p(z_n)) = \text{KL}(q(z) \| p(z_n))
$$

$$
= \mathrm{KL}(q(z) \, \| \, p(z))
$$

where $\mathbb{I}_{q(n,z)}[n,z]$ denotes the mutual information of n and z in $q(n,z)$.

To check this expression, write