The Neglected Sibling: Isotropic Gaussian Posterior for VAE

Lan Zhang  Wray Buntine  Ehsan Shareghi

Department of Data Science and AI, Monash University
Lan.Zhang@monash.edu

Abstract

Deep generative models have been widely used in several areas of NLP, and various techniques have been proposed to augment them or address their training challenges. In this paper, we propose a simple modification to Variational Autoencoders (VAEs) by using an Isotropic Gaussian Posterior (IGP) that allows for better utilisation of their latent representation space. This model avoids the sub-optimal behavior of VAEs related to inactive dimensions in the representation space. We provide both theoretical analysis, and empirical evidence on various datasets and tasks that show IGP leads to consistent improvement on several quantitative and qualitative grounds, from downstream task performance and sample efficiency to robustness. Additionally, we offer insights about the representational properties encouraged by IGP and also show that its gain generalises to image domain as well.\footnote{For code and data: https://github.com/lanzhang128/IGPVAE}

1 Introduction

Variational Autoencoders (VAEs) \cite{kingma2014auto-encoding} have been widely used in various areas of NLP, from representation learning for downstream tasks \cite{li2020understanding,wei2017variational,bowman2016generating}, to generation \cite{prokhoro2019independent}, and semi-supervised learning \cite{zhu2021semi,choi2019joint,yin2018generating,xu2017unsupervised}. In recent years, most of the developments around VAEs have focused on avoiding the commonly known posterior collapse problem \cite{bowman2016generating} which leads to learning sub-optimal representations \cite{havrylov2020activation,fu2019understanding,li2019understanding,dieng2019understanding,he2019hyperprior,higgins2017beta,yang2017semi,bowman2016generating}.

Despite the success of these techniques in avoiding the collapse, a non-collapsed VAE can still utilise the representation space sub-optimally \cite{prokhoro2019independent,he2019hyperprior,burda2016understanding}, as very commonly the learned representations do not fully utilise the latent space to encode information. Additionally, vanilla VAEs induce representation spaces as soft ellipsoidal regions \cite{bowman2016generating} and in such a space, the mean vectors are less representative, specifically as features for downstream tasks \cite{bosc2020latent}.

Figure \ref{fig:example} (Left) reports the distribution of mean and variance for VAE (collapse is avoided using a variant of $\beta$-VAE (Prokhorov et al. 2019) with a standard Diagonal Gaussian Posterior. The empirical observation in Figure 1 suggests the uni-dimensional posterior distribution on inactive dimensions (a.k.a. Inactive Units) \cite{burda2016understanding} are at least a subset of dimensions being active. However, in a non-collapsed VAE, it is required from the posterior to encode information in the latent code, leading to at least a subset of dimensions being active.

1 A dimension $u$ is defined to be active if $A_u = $ Cov$_\mathcal{N}(\mathbb{E}_{u=q(u|\mathbf{x})}[u])$ is larger than 0.01, where Cov denotes covariance.

Figure 1: An observation of the latent space in VAEs trained on SNLI Corpus \cite{bowman2015large}. Boxplots of posteriors’ mean (top) and variance (bottom) are produced on the test set. Left plot corresponds to a non-collapsed VAE (using $C = 5$, Eq. \ref{eq:beta}), right plots correspond to the same setting but with an isotropic posterior. Blue and red boxes denote results of active units (AU) and inactive units (IAU), respectively.
with the exception that all dimensions share the same unified variance.

Our experimental results across text and image modalities indicate that the proposed modification consistently outperforms the vanilla VAE, and strong baseline such as IWAE (Burda, Grosse, and Salakhutdinov 2016) on several quantitative and qualitative bases. We also illustrate that the proposed simple modification improves representational robustness to input perturbation, as well as sample efficiency in downstream tasks. Finally, we provide a theoretical insight about the efficiency of Isotropic VAE, as well as empirical evidence to highlight its optimisation merits.

2 Variational Autoencoder (VAE)

Let \( x \) denote datapoints in data space and \( z \) denote latent variables in the latent space, and assume the datapoints are generated by the combination of two random processes: The first random process is to sample a point \( z(i) \) from the latent space in VAEs with prior distribution of \( z \), denoted by \( p(z) \). The second random process is to generate a point \( x(i) \) from the data space, denoted by \( p(x|z(i)) \). VAE uses a combination of a probabilistic encoder \( q_{\phi}(z|x) \) and decoder \( p_{\theta}(x|z) \), parameterised by \( \phi \) and \( \theta \), to learn this statistical relationship between \( x \) and \( z \). VAE is trained by maximizing the lower evidence lower bound (ELBO), \( L(\phi, \theta; x) \):

\[
E_{q_{\phi}(z|x)}[\log p_{\theta}(x|z)] - KL(q_{\phi}(z|x)||p(z))
\]

(1)

The first term of objective function is the expectation of the log of data likelihood under the prior distribution of \( z \). The second term is KL-divergence, measuring the distance between the recognition distribution \( q_{\phi}(z|x) \) and the prior distribution \( p(z) \) and can be seen as a regularisation.

2.1 Posterior Collapse

In the presence of auto-regressive and powerful decoders, a common optimisation challenge of training VAEs in text modelling is called posterior collapse, where the learned posterior distribution \( q_{\phi}(z|x) \), collapses to the prior \( p(z) \). Posterior collapse results in the latent variables \( z \) being ignored by the decoder. Several collapse strategies have been proposed to alleviate this problem from different angles such as choice of decoders (Yang et al. 2017) (Bowman et al. 2016), adding more dependency between encoder and decoder (Dieng et al. 2019), adjusting the training process or objective function (Havrylov and Titov 2020) (Fu et al. 2019) (He et al. 2019) (Bowman et al. 2016), and imposing direct constraints to the KL term (Pelsmaeker and Aziz 2020) (Kazavi et al. 2019) (Higgins et al. 2017). In this work, we apply constraint optimisation (Burgess et al. 2018) (Prokhorov et al. 2019):

\[
L(\phi, \theta; x) = E_{q_{\phi}(z|x)}[\log p_{\theta}(x|z)] - \beta |KL(q_{\phi}(z|x)||p(z))| - C
\]

(2)

where \( C \) is a positive real value which represents the target KL-divergence term value. We set \( \beta = 1 \) to make sure the weights of the two terms balance, noting that it acts as a Lagrange Multiplier (Boyd and Vandenberghe 2004). This also has an information-theoretic interpretation, where the placed constraint \( C \) on the KL term is seen as the amount of information transmitted from a sender (encoder) to a receiver (decoder) via the message \( z \) (Alemi et al. 2018).

3 Isotropic Gaussian Posterior (IGP)

A common behaviour of VAEs is the presence of inactive representation units across the entire dataset, causing the number of utilised dimensions to be even far smaller than the number of potential generative factors behind any real-world dataset. The soft ellipsoidal representation space of VAEs is known to lead to less representative mean vectors (Bosc and Vincent 2020).

We argue that tying the variance of dimensions on the posterior will avoid the aforementioned issue since the encoder of VAEs would be forced to either use all dimensions or none and the learned latent space is soft spherical. In the Gaussian case, this corresponds to using an Isotropic Gaussian, a subclass of diagonal Gaussian distribution \( \{N(\mu, \sigma^2 I) | \mu \in \mathbb{R}^n, \sigma \in \mathbb{R}^+ \} \), as the posterior. Although an AE can have the same effect, tying variances to a constant value 0 rather than a learned value makes the regularizer no use and the generation quality worse.

Tying the variances in IGP could impose a different pathological pattern on units’ activation as illustrated in the right panels of Figure 1. As demonstrated the use of IGP in a non-collapsed VAE encourages the number of active units to be pushed towards the maximum (i.e., matching the representation dimensionality). These observations generalise to other datasets and configurations of \( C \).

Additionally, the use of IGP allows the estimation of variance more accurately. Suppose we have \( N \) samples with the same posterior. For a K-dimension diagonal Gaussian posterior, we will have an estimate of variance with standard deviation approximately \( \frac{\hat{\sigma}_k^2}{\sqrt{\frac{2}{NK}}} \) for each dimension \( k \), whereas for an isotropic Gaussian posterior, we will have a unified estimation of variance with standard deviation approximately \( \frac{\hat{\sigma}^2}{\sqrt{\frac{2}{NK}}} \), where \( \hat{\sigma}_k^2 \) and \( \hat{\sigma}^2 \) denote the estimates of the variance. Moreover, with \( K \) different \( \hat{\sigma}_k^2 \) estimates, a few may differ substantially from their best values by chance.

4 Experiments

We trained our models on four datasets, CBT (Hill et al. 2016), Yahoo Question and DBpedia (Zhang, Zhao, and Le Cun 2015), and SNLI Hypothesis (Bowman et al. 2015). See Table 1 for data statistics. We compare the choice of isotropic Gaussian posterior (IGP) with vanilla diagonal Gaussian posterior (DGP) on various quantitative and qualitative settings, from reconstruction and representation utilisation (§4.1) to downstream classification task and sample efficiency (§4.2). We also provide additional analysis about the representational properties encouraged by IGP, showing it works well with extremes where all dimensions are either active or inactive. In the Gaussian case, such property could be encouraged by using an Isotropic Gaussian, corresponding to Figure 1 (Right). This only involves a simple modification of VAE by replacing its posterior with an Isotropic Gaussian distribution \( \mathcal{N}(\mu, \sigma^2 I) \), which is similar to vanilla VAE's posterior with the exception that all dimensions share the same unified variance.
other priors, and could also be used as a warm initialiser to boost performance of vanilla VAEs (§5).

**Model Configurations.** We use the VAE architecture of (Bowman et al. 2016) and concatenate the latent code with word embedding at every timestamp as the input of the decoder. For VAE with IGP, we just produce one variance value and assign it to be the variance of posterior for all dimensions. At decoding phase, we use greedy decoding. The dimensions for word embedding, encoder-decoder LSTMs, and latent code are (200, 512, 32) for CBT, Yahoo and DBpedia, and (50, 128, 8) for SNLI. Three different values of $C$ (see Eq. 2) is used on each dataset to explore the impact of the amount of information transmitted by the code. We also adopt Autoencoder (AE) as a baseline.

**4.1 Basic Findings**

Figure 2 reports the reconstruction loss, active units (AU; Burda, Grosse, and Salakhutdinov 2016) and BLEU-2 (Papineni et al. 2002) for 3 settings of $C = 5, 15, 50$ on 4 datasets. KL in all cases match the set target $C$. For full results, including KL, BLEU-4 and ROUGE-2/4 (Lin 2004), see Supplementary Material.

**Table 1: Statistics of datasets.**

| Data       | Train | Dev  | Test | Vocab | # Classes |
|------------|-------|------|------|-------|-----------|
| CBT        | 192K  | 10K  | 12K  | 12K   | -         |
| DBpedia    | 140K  | 14K  | 14K  | 12K   | 14        |
| Yahoo Question | 100K  | 10K  | 10K  | 12K   | 10        |
| SNLI       | 100K  | 10K  | 10K  | 7K    | -         |

**Target $C$.** The usage of $C$ can effectively control the KL-divergence to a certain level. The reconstruction loss generally drops with the increase of $C$ value. We observe the same pattern, comparing the loss for VAE (DGP) and VAE with IGP under each $C$ value. While the increase of $C$ is interpreted as the amount of information transmitted through the representation (Alemi et al. 2018; Piot et al. 2019), the use of IGP allows for a better utilisation of the latent space and consequently a more accurate estimation of the posterior.

**Active unit (AU).** AU measures the activity of dimensions in the representation space. Intuitively, if the changes of the mean value of posterior distribution on a dimension across all data points in the training set are significant, the corresponding dimension is active and carries information. As can be observed, compared to the diagonal Gaussian counterpart, IGP can activate significantly more dimensions, almost all dimensions in many cases.

**BLEU-2.** VAEs with IGP significantly and consistently outperform vanilla VAEs. Having a comparable BLEU-2 on CBT dataset for IGP ($C = 5$) vs. DGP ($C = 15$), highlights the potential of IGP even at a lower $C$ value: 15.45 vs 14.76.

**Does IGP work only with $C$?** We chose setting a target KL via $C$ as our strategy to avoid posterior collapse. While this strategy offers an information-theoretic interpretation of VAE behaviours, we still want to examine if our proposed IGP could work with another commonly used method, $\beta$-VAE (Higgins et al. 2017) which incorporates a $\beta$ coefficient for the KL term. We train VAEs on SNLI with the same training configuration as before and three different $\beta$. The results are reported in Table 2. Although non-comparable KL values makes it difficult to fairly compare models, we still observe IGP VAEs can activate more dimensions, and have better reconstruction quality with a rather smaller KL-divergence than DGP VAEs (e.g., IGP VAEs ($\beta = 0.4$) and $\beta = 0.8$)
We attribute this success to having a more representative sample efficiency of both V AEs, by using 0.1%, 1% and 10% of training data of DBpedia to them with $C = 15$ and do classification on the test set as before. Accuracy scores are reported in Figure 3 (bottom). Using VAE with IGP in a few-shot setting can lead to a better performance on this classification task. For instance, the mean accuracy gap at 0.1% is quite significant being above 7 points, and VAE gets the gap down to 4 points at 100% (still significant).

**Reserved Label Information.** We conduct the agreement experiment (Bosc and Vincent 2020), extrinsically examining the amount of label information preserved during the reconstruction phase. We first train a regular text classifier with a 512D LSTM and two hidden MLP layers with 128 neurons and ReLU activation function each on top of the training set of DBpedia and Yahoo using the same training set as VAEs. We then calculate the macro-averaged F1-scores of reconstructed test sets and divide them with the the macro-averaged F1-score of the original test set to represent the proportion of label information reserved in reconstruction. The results are provided in Table 3.
We quantitatively evaluate the quality of generated sentences via forward perplexity and reverse perplexity experiments (Zhao et al. 2018). We first generate 100K sentences using standard Gaussian distribution. Then, we use a LSTM layer with 512 dimensions and a 200-dimension word embedding layer to form a common language model. We train the language model on the original training set and calculate standard forward perplexity on generated sentences of models. The reverse perplexity is calculated on the test set based on the language model trained on generated sentences. All number of training epochs is set to 10 and results are provided in Table 3. We also include the standard perplexity on test set.

In most cases, IGP VAEs have both lower forward and reverse perplexity than DGP VAEs, which indicates their better generation quality. For the forward perplexity, both IGP and DGP VAEs have a better quality than AE. This validates the superiority of VAE over AE in generation. However, for the reverse perplexity, VAEs are not necessarily better than AE and those with lower C value usually have significantly larger perplexity. The explanation might be that the limited capacity indicated by lower C prevent VAEs to learn richer features for modelling sentences in the latent space.

### 4.3 Generation

We quantitatively evaluate the quality of generated sentences via forward perplexity and reverse perplexity experiments (Zhao et al. 2018). We first generate 100K sentences using standard Gaussian distribution. Then, we use a LSTM layer with 512 dimensions and a 200-dimension word embedding layer to form a common language model. We train the language model on the original training set and calculate standard forward perplexity on generated sentences of models. The reverse perplexity is calculated on the test set based on the language model trained on generated sentences. All number of training epochs is set to 10 and results are provided in Table 3. We also include the standard perplexity on test set.

In most cases, IGP VAEs have both lower forward and reverse perplexity than DGP VAEs, which indicates their better generation quality. For the forward perplexity, both IGP and DGP VAEs have a better quality than AE. This validates the superiority of VAE over AE in generation. However, for the reverse perplexity, VAEs are not necessarily better than AE and those with lower C value usually have significantly larger perplexity. The explanation might be that the limited capacity indicated by lower C prevent VAEs to learn richer features for modelling sentences in the latent space.

**Robustness.** We further apply word dropout on sentences by randomly deleting 30% of words in a sentence, and repeating the classification experiment. The classification accuracy on the representation of polluted sentences are reported in Table 4. IGP VAEs outperform DGP VAEs in all cases with up-to 9% absolute gain. Moreover, with C ≥ 15, IGP VAEs are more robust to polluted sentences than AE. This further suggests the superiority of IGP VAEs on classification tasks.

### 4.4 Image Modality

To examine the effectiveness of IGP, we run experiments on Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017). Since KL-collapse is less of a concern in image domain (due to the absence of powerful auto-regressive decoding), we consider vanilla VAE with DGP and IGP posteriors, AE and IWAE (Burda, Grosse, and Salakhutdinov 2016) as our baselines. We follow the encoder-decoder architecture as in (Burda, Grosse, and Salakhutdinov 2016), and set the latent dimension to 32. We train VAEs for 50 epochs with 128 batch size.

---

**Table 3:** Robustness and Proportion of reserved label information on test sets.

| Model | DBpedia | Yahoo Question |
|-------|---------|----------------|
| AE    | 91.85±0.06 | 64.77±0.59 |
| C = 5, DGP | 53.79±3.83 | 24.33±0.72 |
| C = 5, IGP | 46.62±0.51 | 28.75±0.48 |
| C = 15, DGP | 80.45±1.71 | 35.89±0.96 |
| C = 15, IGP | 84.24±0.74 | 54.42±0.21 |
| C = 50, DGP | 89.40±0.37 | 51.09±0.33 |
| C = 50, IGP | 93.48±0.37 | 64.88±0.37 |

**Table 4:** Forward and Reverse perplexities.

| Model | Forward | Reverse |
|-------|---------|---------|
| Real  | 418.9±24.3 | 95.9±7.3 |
| AE    | 46.9±0.89 | 85.3±2.07 |
| C = 5, DGP | 8.37±0.013 | 879.3±19.5 |
| C = 5, IGP | 8.37±0.11 | 879.3±19.5 |
| C = 15, DGP | 15.06±0.48 | 879.3±19.5 |
| C = 15, IGP | 14.08±0.07 | 879.3±19.5 |
| C = 50, DGP | 51.25±0.07 | 879.3±19.5 |
| C = 50, IGP | 52.20±0.81 | 879.3±19.5 |

---

4 See Supplementary Material for experiments on MNIST digits (LeCun, Cortes, and Burges 2010).
the carnegie library in unk washington is a building from 1911. it was listed on the national register of historic places in 1982.

the carnegie library in unk washington is a unk (unk ft) high school in the unk district of unk in the province of unk in the unk province of armenia.

the carnegie library in unk washington was built in 1909. it was listed on the national register of historic places in unk was designed by architect john unk.

st. marys catholic high school is a private roman catholic high school in phoenix arizona. it is located in the roman catholic diocese of phoenix.

st. marys catholic high school is a unk-unk school in unk unk county new jersey united states. the school is part of the unk independent school district.

st. marys catholic high school is a private roman catholic high school in unk california. it is located in the roman catholic diocese of unk.

- Table 5: Word imputation experiment on DBpedia test set.

| Model  | Rec. | KL   | AU | CLS   |
|--------|------|------|----|-------|
| AE     | 211.10 | 13.33 | 32 | 81.89% |
| VAE    | 227.14 | 13.35 | 8  | 77.83% |
| DGP    | 230.47 | 11.87 | 13 | 77.57% |
| IWAE (k = 1) | 246.58 | 10.04 | 13 | 77.78% |
| IWAE (k = 5) | 246.22 | 14.75 | 32 | 82.37% |
| IWAE (k = 50) | 304.62 | 8.82  | 32 | 80.01% |

- Table 6: Reconstruction losses, KL, unit activation, and classification accuracy (CLS) of IGP and DGP models on Fashion-MNIST dataset. All results are calculated on test set.

and Adam (Kingma and Ba 2015) optimizer with learning rate 0.001, and trained with 5 different random starts. We select the model with smallest ELBO for VAE and IWAE, and lowest reconstruction loss for AE and report losses in Table 6. Using IGP usually leads to a higher KL value and ELBO, but its effectiveness on activating all dimensions of the latent space is well-preserved on image domain.

Classification Experiment. We use the latent code of images to do classification similar to text classification setting. Results are included in Table 6. IGP VAEs still outperform DGP VAEs and AE, and achieves the best classification accuracy, whereas DGP VAEs cannot have a comparable accuracy with AE. This observation is consistent with our earlier observation for text.

Generative Experiment. We use samples from prior to generate images in Figure 5. As expected, AE cannot generate a valid image with clear boundary of object shape from random points, but both DGP and IGP models can generate images of various objects with different shapes. Intuitively, some images generated by DGP model have more emphasis on darker pixels and hence have a sharper boundary. In contrast, IGP model generates images with a lighter choice of pixels. However, there is no distinguishable difference between images generated by DGP and IGP VAE in terms of recognising objects.

5 Analysis

In this section we provide various theoretical and empirical analyses from different perspectives.

5.1 Is This a Capacity Utilization Problem?
The encoder-decoder of VAEs can be seen as the sender-receiver in a communication network, where an input message \( x \), is encoded into a compressed code \( z \), and the receiver aims to decode \( z \) back into \( x \) (Alemi et al. 2018). The capacity of encoder in such a network can be measured in terms of the mutual information between \( x, z \) bounded by:

\[
H - D \leq I(x, z) \leq R,
\]
Table 7: Comparison of models trained fully under IGP or DGP vs hybrid models trained first via IGP, then initialised using this Base model and switched to DGP.

| Config. | Rec. | KL  | AU  | BLEU-2/4 |
|---------|------|-----|-----|----------|
| C = 5, DGP | 24.71 | 5.08 | 4 | 19.97/7.49 |
| C = 5, IGP | 24.39 | 5.12 | 8 | 26.89/12.04 |
| Base, IGP | 28.18 | 5.04 | 8 | 18.29/5.47 |
| Base, C = 5, DGP | 24.53 | 5.11 | 6 | 24.48/10.60 |

Table 8: Reports \( ||\mu||^2_2, \log \det(\text{Cov}[q_\phi(z)]) \).

| CBT | DBpedia | Yahoo |
|-----|---------|-------|
| DGP | [0.11, −0.41] | [0.11, −0.63] | [0.10, −0.51] |
| IGP | [0.11, −1.48] | [0.12, −6.22] | [0.08, −4.86] |

5.3 Posterior Shape

To understand the impact of isotropy on the aggregated posterior, \( q_\phi(z) = \sum_{x \sim q(x)} q_\phi(z|x) \), we obtain unbiased samples of \( z \) by sampling an \( x \) from data and then \( z \sim q_\phi(z|x), \) and measure the log determinant of covariance of the samples \((\log \det(\text{Cov}[q_\phi(z)]))\) as well as the mean of the samples to measure \( ||\mu||^2_2 \). Table 8 reports these for \( C = 15 \). We observe that \( \log \det(\text{Cov}[q_\phi(z)]) \) is significantly lower for IGP across indicating a sharper approximate posteriors. Additionally, the roughly matching \( ||\mu||^2_2 \), in the presence of higher AU for IGP (see Figure 2) suggests while both models were forced to transmit the same amount of information via a fixed channel capacity \( C \), the IGP model has a more even distribution of information across different axes of the representation space, whereas DGP utilised only a small subspace. This, as observed earlier, translates into better discrimination in classification tasks.

6 Conclusion

In this work, we proposed a neglected variant of VAEs based on Isotropic Gaussian Posteriors (IGP) which addresses some of the sub-optimal properties of vanilla VAEs in representation space utilisation. We compare IGP VAEs with DGP VAEs, AEs, and IWAE on a series of quantitative and qualitative experiments including classification and generation.

Our proposed modification consistently improved upon several established criteria from reconstruction quality to downstream task performance on various datasets. On the text classification tasks, IGP VAEs could outperform AEs, whereas DGP VAEs struggle to match AEs on performance. On the text generation tasks, IGP VAEs generally have better generation quality than DGP VAEs. These successes translate also into the image domain, specifically for downstream classification problem. Our empirical findings suggest IGP VAE is a good universal choice for both classification and generation tasks, as well as both text and image modalities.

Our ongoing work suggests the representation utilisation achieved by IGP has the potential to be exploited towards desired representational properties such as disentanglement. We are investigating such potentials as our future direction.

References

Alemi, A.; Poole, B.; Fischer, I.; Dillon, J.; Saurous, R. A.; and Murphy, K. 2018. Fixing a broken ELBO. In International Conference on Machine Learning, 159–168. PMLR.

Bosc, T.; and Vincent, P. 2020. Do sequence-to-sequence VAEs learn global features of sentences? In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 4296–4318. Online: Association for Computational Linguistics.

where \( R \) is \( \text{KL}(q_\phi(z|x)||p(z)) \), and \( D \) is \( -E_{q_\phi(z|x)}[\log p_\theta(x|z)] \), and \( H \) is the empirical data entropy (constant). In this framework, increasing the KL term can control the maximum capacity of the encoder channel. Tying variances of a diagonal Gaussian distribution can yield a higher KL-divergence requirement as well as learn a more representative mean vector for the latent space, which could be simply proved by the following theorem.

**Theorem 1.** Assume that diagonal Gaussian Distribution \( p(x) = \mathcal{N}(\mu; \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)) \), is a probability density function for \( n \) random variables \( X_1, X_2, \ldots, X_n \), and \( q(y) = \mathcal{N}(\mu; \text{diag}(\sigma_1^2 I_n)) \), where \( \sigma = \min(\sigma_1, \sigma_2, \cdots, \sigma_n) \), is another probability density function for \( n \) random variables \( Y_1, Y_2, \ldots, Y_n \). Then, we have:

\[
\text{KL}(p(x)||\mathcal{N}(0; I)) \leq \text{KL}(q(y)||\mathcal{N}(0; I)),
\]

and for any \( \varepsilon \),

\[
\Pr\left( \frac{1}{n} \sum_{i=1}^{n} |X_i - \mu_i| \leq \varepsilon \right) \leq \Pr\left( \frac{1}{n} \sum_{i=1}^{n} |Y_i - \mu_i| \leq \varepsilon \right).
\]

**Proof.** See Supplementary Material. \( \square \)

Hence, the direct use of isotropic Gaussian posterior can lead to a more optimal utilization of the latent space.

5.2 Is This an Optimization Problem?

In the text domain, training VAEs faces the problem of posterior collapse. Popular strategies of avoiding posterior collapse in text such as \( \beta \)-VAE or the usage of \( C \) usually make the ELBO looser. It has been argued for that ELBO tightness might hurt the performance [Rainforth et al. 2018]. Our empirical observation verifies the similar findings as tying the variances of a diagonal Gaussian posterior relaxes the tightness on ELBO and still leads to a better global optimum.

To highlight the optimisation advantage of IGP over DGP, we make an empirical observation on the SNLI dataset using batch size of 100 to first train a base VAE with IGP for 5 epochs with \( C \) gradually increased to 5 (we refer to this as base model). Then using its weights, we untie variances to initialize a DGP model and further train it with \( C = 5 \) from the base initialization. The results are provided in Table 7.

We observe that using IGP as an initializer for vanilla VAE indeed improves performance in terms of reconstruction but does not make the DGP VAE to have the same AU under the same level of KL-divergence (compare rows \( C = 5 \), IGP with Base, \( C = 5 \), DGP). This indicates that the change of posterior might make the optimization of VAEs easier.
Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 632–642. Lisbon, Portugal: Association for Computational Linguistics.

Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A. M.; Józefowicz, R.; and Bengio, S. 2016. Generating Sentences from a Continuous Space. In Proceedings of the 20th SIGMEL Conference on Computational Natural Language Learning, CoNLL 2016, 10–21. Berlin, Germany: ACL.

Boyd, S.; and Vandenberghe, L. 2004. Convex optimization. Cambridge University Press.

Burda, Y.; Grosse, R. B.; and Salakhutdinov, R. 2016. Importance Weighted Autoencoders. In 4th International Conference on Learning Representations, ICLR 2016, Conference Track Proceedings. San Juan, Puerto Rico.

Burgess, C. P.; Higgins, I.; Pal, A.; Matthey, L.; Watters, N.; Desjardins, G.; and Lerchner, A. 2018. Understanding disentangling in β-VAE. CoRR, abs/1804.03599.

Choi, J.; Kim, T.; and Lee, S. 2019. A Cross-Sentence Latent Variable Model for Semi-Supervised Text Sequence Matching. In Korhonen, A.; Traum, D. R.; and Márquez, L., eds., ACL, 4747–4761.

Dieng, A. B.; Kim, Y.; Rush, A. M.; and Blei, D. M. 2019. Avoiding Latent Variable Collapse with Generative Skip Models. In The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, volume 89 of Proceedings of Machine Learning Research, 2397–2405. Naha, Okinawa, Japan: PMLR.

Fu, H.; Li, C.; Liu, X.; Gao, J.; Celikyilmaz, A.; and Carin, L. 2019. Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 240–250. Minneapolis, Minnesota: Association for Computational Linguistics.

Havrylov, S.; and Titov, I. 2020. Preventing Posterior Collapse with Levenshtein Variational Autoencoder. CoRR, abs/2004.14758.

He, J.; Spokony, D.; Neubig, G.; and Berg-Kirkpatrick, T. 2019. Lagging Inference Networks and Posterior Collapse in Variational Autoencoders. In 7th International Conference on Learning Representations, ICLR 2019. New Orleans, LA, USA.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In 5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings. Toulon, France.

Hill, F.; Bordes, A.; Chopra, S.; and Weston, J. 2016. The Goldilocks Principle: Reading Children’s Books with Explicit Memory Representations. In 4th International Conference on Learning Representations, ICLR 2016. San Juan, Puerto Rico.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In 3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings. San Diego, CA, USA.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In 2nd International Conference on Learning Representations, ICLR 2014, Conference Track Proceedings. Banff, AB, Canada.

LeCun, Y.; Cortes, C.; and Burges, C. 2010. MNIST handwritten digit database. ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist, 2.

Li, B.; He, J.; Neubig, G.; Berg-Kirkpatrick, T.; and Yang, Y. 2019. A Surprisingly Effective Fix for Deep Latent Variable Modeling of Text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3603–3614. Hong Kong, China: Association for Computational Linguistics.

Li, C.; Gao, X.; Li, Y.; Peng, B.; Li, X.; Zhang, Y.; and Gao, J. 2020. Optimus: Organizing Sentences via Pre-trained Modeling of a Latent Space. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 4678–4699. Online: Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Pelsmaeker, T.; and Aziz, W. 2020. Effective Estimation of Deep Generative Language Models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 7220–7236. Online: Association for Computational Linguistics.

Prokhorov, V.; Shareghi, E.; Li, Y.; Pilehvar, M. T.; and Collier, N. 2019. On the Importance of the Kullback-Leibler Divergence Term in Variational Autoencoders for Text Generation. In Proceedings of the 3rd Workshop on Neural Generation and Translation@EMNLP-IJCNLP 2019, 118–127. Hong Kong: Association for Computational Linguistics.

Rainforth, T.; Kosiorek, A.; Le, T. A.; Maddison, C.; Igl, M.; Wood, F.; and Teh, Y. W. 2018. Tighter Variational Bounds are Not Necessarily Better. In Dy, J.; and Krause, A., eds., Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, 4277–4285. PMLR.

Razavi, A.; van den Oord, A.; Poole, B.; and Vinyals, O. 2019. Preventing Posterior Collapse with delta-VAEs. In 7th International Conference on Learning Representations, ICLR 2019. New Orleans, LA, USA.
van der Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605.

Wei, L.; and Deng, Z.-H. 2017. A Variational Autoencoding Approach for Inducing Cross-lingual Word Embeddings. In *IJCAI*, 4165–4171.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *CoRR*, abs/1708.07747.

Xu, W.; Sun, H.; Deng, C.; and Tan, Y. 2017. Variational Autoencoder for Semi-Supervised Text Classification. In Singh, S. P.; and Markovitch, S., eds., *AAAI*, 3358–3364.

Yang, Z.; Hu, Z.; Salakhutdinov, R.; and Berg-Kirkpatrick, T. 2017. Improved Variational Autoencoders for Text Modeling using Dilated Convolutions. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, 3881–3890. Sydney, NSW, Australia: PMLR.

Yin, P.; Zhou, C.; He, J.; and Neubig, G. 2018. StructVAE: Tree-structured Latent Variable Models for Semi-supervised Semantic Parsing. In Gurevych, I.; and Miyao, Y., eds., *ACL*, 754–765.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-Level Convolutional Networks for Text Classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS ’15*, 649–657. Cambridge, MA, USA: MIT Press.

Zhao, J.; Kim, Y.; Zhang, K.; Rush, A.; and LeCun, Y. 2018. Adversarially Regularized Autoencoders. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 5902–5911. PMLR.

Zhu, Y.; Shareghi, E.; Li, Y.; Reichart, R.; and Korhonen, A. 2021. Combining Deep Generative Models and Multi-lingual Pretraining for Semi-supervised Document Classification. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 894–908. Online: Association for Computational Linguistics.
A Proof of Theorem 1

Proof. The first thing can be simply proved by the monotonic decreasing property of KL divergence in the Gaussian case. For the second thing, according to assumptions, we have:

\[
Pr \left( \bigcap_{i=1}^{n} |X_i - \mu_i| < \varepsilon \right) = \prod_{i=1}^{n} \int_{-\varepsilon}^{\varepsilon} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{x^2}{2\sigma_i^2}} dx,
\]

and

\[
Pr \left( \bigcap_{i=1}^{n} |Y_i - \mu'_i| < \varepsilon \right) = \prod_{i=1}^{n} \int_{-\varepsilon}^{\varepsilon} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{y^2}{2\sigma_i^2}} dy_i
\]

\[= \left[ \int_{-\varepsilon}^{\varepsilon} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2\sigma^2}} dy \right]^n.\]

Construct auxiliary function:

\[f(\sigma) = \int_{-\varepsilon}^{\varepsilon} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx = 2 \int_{0}^{\varepsilon} \frac{1}{\sqrt{\pi}} e^{-x^2} dx.\]

We have \(f(\sigma) > 0\). Calculating the first derivative of \(f(\sigma)\), we have:

\[f'(\sigma) = -\frac{2\varepsilon}{\sqrt{2\pi}\sigma^2} e^{-\frac{x^2}{2\sigma^2}} < 0.\]

Therefore, \(f(\sigma)\) is a strictly decreasing function and \(f(\sigma_i) \leq f(\sigma)\). We have:

\[Pr \left( \bigcap_{i=1}^{n} |X_i - \mu_i| < \varepsilon \right) = \prod_{i=1}^{n} f(\sigma_i)
\]

\[\leq
\]

\[\left[ f(\sigma) \right]^n = Pr \left( \bigcap_{i=1}^{n} |Y_i - \mu'_i| < \varepsilon \right)\]

\[\square\]

B Full Results

We report detailed reconstruction loss, KL-divergence, active units and results of BLEU and ROUGE scores on reconstructed test set in Table 9.

C Results on Mixture Priors

In addition to standard Isotropic Gaussian prior, we also test how well both posteriors work with Mixture of Gaussian (MoG) prior (Pelsmaeker and Aziz 2020). Since the regular MoG prior incorporates several diagonal Gaussian distributions, we also employ Mixture of Isotropic Gaussian (MoIG) to see the influence of introducing isotropy. We set the number of components of priors to 5 and train VAEs on CBT. Table 10 reports reconstruction loss, BLEU-2 and active units.

As observed using either MoG or MoIG as prior degrades VAEs performance for both posteriors. The combination of MoIG prior and Isotropic posterior has a clear advantage of MoG due to consistency between prior and posterior distributional families. In all cases IGP outperforms its vanilla VAE counterpart (DGP), indicating the better compatibility of IGP with other priors.

D Full Agreement and Classification Results

The results of agreement experiment based on reconstructed test set and classification task based on the representations from VAEs trained on classification datasets are shown in Table 11 for three runs.

E Results on MNIST

We show the losses of models trained on MNIST, and their classification accuracy in Table 12. We also provide generation examples of AE, DGP VAE and IGP VAE in Figure 6.
Table 9: Results are calculated on the test set. We report mean value and standard deviation across 3 runs. Rec, AU, and PPL denote reconstruction loss, number of Active Units and estimated perplexity, respectively. DGP, and IGP denote diagonal Gaussian posteriors and isotropic Gaussian posteriors, respectively. $C$ is the target KL value.

|       | Rec.         | KL    | AU    | BLEU-2/4 | ROUGE-2/4 |
|-------|--------------|-------|-------|----------|-----------|
| CBT   |              |       |       |          |           |
| AE    | 16.94±0.11  | -     | 32.0±0.0 | 39.71±0.14/30.81±0.12 | 43.51±0.11/37.01±0.09 |
| $C = 5$, DGP | 60.04±0.04  | 5.14±0.03 | 4.3±0.5 | 10.92±0.24/2.17±0.07 | 5.04±0.24/0.32±0.02 |
| $C = 5$, IGP | 60.50±0.02  | 5.09±0.03 | 32.0±0.00 | 15.45±0.18/4.00±0.09 | 8.74±0.10/0.78±0.03 |
| $C = 15$, DGP | 52.11±0.11  | 15.11±0.04 | 11.0±1.6 | 14.76±0.04/4.81±0.01 | 8.79±0.10/2.13±0.02 |
| $C = 15$, IGP | 52.38±0.07  | 15.24±0.05 | 32.0±0.00 | 26.84±0.30/11.73±0.15 | 20.54±0.21/6.39±0.10 |
| $C = 50$, DGP | 32.50±0.20  | 50.11±0.10 | 31.0±0.00 | 26.94±0.34/16.53±0.20 | 25.39±0.39/16.44±0.32 |
| $C = 50$, IGP | 31.97±0.08  | 30.33±0.07 | 32.0±0.00 | 33.27±0.20/22.27±0.27 | 33.38±0.30/23.65±0.22 |
| DBpedia |              |       |       |          |           |
| AE    | 66.32±0.11  | -     | 32.0±0.00 | 40.96±0.25/27.57±0.17 | 35.87±0.19/23.78±0.07 |
| $C = 5$, DGP | 100.69±0.08  | 5.09±0.01 | 4.0±0.00 | 23.47±0.79/14.00±0.47 | 20.85±0.09/10.19±0.34 |
| $C = 5$, IGP | 101.73±0.31  | 5.04±0.01 | 32.0±0.00 | 25.09±0.55/14.83±0.22 | 22.29±0.17/10.47±0.10 |
| $C = 15$, DGP | 94.16±0.19  | 15.00±0.04 | 8.7±0.90 | 35.35±0.49/22.37±0.31 | 30.54±0.43/17.41±0.16 |
| $C = 15$, IGP | 95.52±0.08  | 15.08±0.05 | 32.0±0.00 | 37.23±0.27/24.47±0.11 | 34.19±0.12/19.32±0.06 |
| $C = 50$, DGP | 80.65±0.53  | 50.02±0.04 | 31.7±0.5 | 40.54±0.21/26.95±0.19 | 35.19±0.25/22.13±0.24 |
| $C = 50$, IGP | 80.58±0.04  | 50.15±0.04 | 32.0±0.00 | 44.79±0.30/30.72±0.17 | 39.91±0.12/25.40±0.08 |
| Yahoo Question |              |       |       |          |           |
| AE    | 17.64±0.28  | -     | 32.0±0.00 | 42.88±0.51/32.86±0.58 | 41.63±0.58/31.67±0.67 |
| $C = 5$, DGP | 50.58±0.06  | 5.14±0.01 | 5.7±0.5 | 17.07±0.71/6.04±0.25 | 10.99±0.41/1.50±0.06 |
| $C = 5$, IGP | 51.24±0.01  | 5.06±0.03 | 32.0±0.00 | 20.91±0.03/8.07±0.03 | 14.48±0.13/2.21±0.02 |
| $C = 15$, DGP | 43.00±0.12  | 15.06±0.04 | 9.3±1.2 | 22.62±0.37/10.81±0.21 | 16.04±0.32/4.76±0.08 |
| $C = 15$, IGP | 44.43±0.05  | 15.20±0.12 | 32.0±0.00 | 29.76±0.06/14.99±0.08 | 23.11±0.17/6.94±0.12 |
| $C = 50$, DGP | 28.29±0.40  | 50.00±0.19 | 31.3±0.9 | 31.78±0.73/20.47±0.70 | 27.14±0.85/15.07±0.77 |
| $C = 50$, IGP | 26.18±0.19  | 50.19±0.08 | 32.0±0.00 | 39.68±0.20/27.49±0.31 | 35.73±0.40/22.57±0.55 |
| SNLI   |              |       |       |          |           |
| AE    | 24.70±0.08  | 5.10±0.02 | 4.3±0.5 | 21.84±1.57/8.74±0.90 | 15.89±1.47/2.52±0.37 |
| $C = 5$, DGP | 24.54±0.18  | 5.12±0.01 | 8.0±0.0 | 26.59±0.82/12.00±0.97 | 20.51±0.86/4.41±0.89 |
| $C = 15$, DGP | 19.25±0.42  | 15.04±0.04 | 7.0±0.0 | 31.34±0.24/18.99±0.52 | 28.07±0.39/12.79±0.77 |
| $C = 15$, IGP | 18.58±0.18  | 15.09±0.01 | 8.0±0.0 | 35.56±0.45/22.49±0.44 | 32.75±0.63/16.88±0.60 |
| $C = 50$, DGP | 20.35±0.63  | 49.94±0.04 | 8.0±0.0 | 26.16±1.11/14.15±1.08 | 21.75±1.25/8.14±1.06 |
| $C = 50$, IGP | 18.43±0.71  | 50.00±0.07 | 8.0±0.0 | 30.78±1.60/18.81±1.69 | 27.50±1.77/12.95±1.72 |

Table 10: Results are calculated on the test set. We report mean value and standard deviation across 3 runs. IG, MoG, and MoIГ denote isotropic Gaussian prior, Mixture of Gaussian prior, and Mixture of isotropic Gaussian prior, respectively.
Figure 6: **Left:** AE; **Middle:** DGP, VAE; **Right:** IGP, VAE ($k = 50$).
| Model | Rec. | KL | AU | CLF |
|-------|------|----|----|-----|
| AE    | 58.86| -  | 32 | 90.82% |
| VAE   | 78.64| 21.98| 15 | 91.52% |
| IWAES (k = 1) | 79.29| 21.65| 15 | 91.49% |
| IWAES (k = 5) | 81.62| 20.42| 18 | 91.74% |
| IWAES (k = 50) | 91.03| 17.74| 21 | 91.91% |
| DGP   |      |    |    |     |
| IWAES (k = 1) | 78.93| 29.09| 32 | 88.82% |
| IWAES (k = 5) | 79.51| 28.64| 32 | 88.87% |
| IWAES (k = 50) | 86.61| 24.42| 32 | 89.81% |
| IWAES (k = 50) | 104.14| 19.47| 32 | 90.71% |

Table 12: Comparison of IGP and DGP models on MNIST datasets.