Posterior Collapse in Variational Gradient Origin Networks

1st Peter Clapham
School of Computing
University of Kent
Canterbury, England
pgc8@kent.ac.uk

2nd Marek Grzes
School of Computing
University of Kent
Canterbury, England
m.grzes@kent.ac.uk

***Abstract*—Posterior collapse is a phenomenon that occurs when the posterior distribution degenerates to the prior, leading to a decline in the quality of latent encodings and generative models. While it is known to occur in Variational Autoencoders (VAEs), it is unknown whether it occurs in Variational Gradient Origin Networks (VGONs). The goal of this paper is to compare the posterior collapse of Variational Gradient Origin Networks and Variational Autoencoders. By checking the latent encodings of VGONs against the key posterior collapse metrics, our experiments reveal that VGONs do exhibit posterior collapse both in the decline of the Kullback-Leibler divergence (KLD) and the collapse of individual variables. Furthermore, the results show that VGONs and VAEs have a similar polarized regime, suggesting that the cause of posterior collapse is not specific to the architecture of the model used to find an encoding. These findings support the claim made in previous research that posterior collapse is a general issue that affects a wide range of latent variable models.**

***Index Terms*—Representation Learning, Variational Autoencoders, Gradient Origin Networks, Posterior Collapse, Polarized Regime**

I. Introduction

The Variational Gradient Origin Network (VGON) [1] is a generative model that involves learning a latent representation and subsequently decoding it. In this way, it performs a similar function to a Variational Autoencoder (VAE) [2]. However, instead of training an encoder network, it uses one-shot learning to obtain a representation. Initially, a representation of zeros is passed into the VGON network, but then a single step of gradient descent predicts the lowest loss representation for a given input data item.

Since this model does not have a dedicated neural network for encoding, there may be some differences in the representations. It has been suggested that this may have some positive impact on the latent representation, as it may not experience posterior collapse [1].

Posterior collapse is when the latent representation (the posterior) degenerates to the prior. Since the prior is Gaussian noise, this makes the distribution of latent features useless. While it may be beneficial for some latent features to collapse, to form a polarized regime [3], it is a hugely important issue if they all collapse.

Many publications have investigated the cause of posterior collapse. Some papers have suggested that the design of the network may be the cause of collapse [4]. On the other hand, other papers have shown that the design of the network is not the cause. Instead, these argue that the loss function, the Evidence Lower Bound (ELBO) [5], is to blame for posterior collapse [6], [7]. Others argue that both are factors [8].

VGONs use a slightly different architecture to VAEs, particularly where encoding is involved. On the other hand, while the loss function is slightly different to that of a traditional VAE, it is still based on the ELBO. This different setup for the task of autoencoding provides a fresh perspective on the problem of posterior collapse. With a similar loss function to VAEs but different encoding strategy, VGONs are of particular interest to posterior collapse research.

Finally, the polarized regime [3] is of great interest as it can be seen as partial posterior collapse. We know that VAEs learn a polarized regime, but that is not yet known for VGONs.

The contributions of this paper are as follows:

- We verify whether posterior collapse is exhibited by VGONs when they are over-regularised
- We compare the magnitude of its collapse to VAEs
- We compare the similarity of encodings obtained by VGONs and VAEs to investigate the polarized regime

The explorations included in this paper show that posterior collapse is, contrary to prior assumptions, exhibited by VGONs. They also show that it is experienced very similarly to VAEs with only minor differences. These similarities crucially include the prevalent polarized regime.

II. Preliminaries

A. Variational Autoencoders

Variational Autoencoders (VAEs) [2] are autoencoders that produce a latent representation by compressing data down to a probability distribution, rather than a deterministic vector. This model aims to find a distribution for the encoder $q_\phi(z|x) = \mathcal{N}(\mu, \sigma^2)$, and a data distribution $p_\theta(x|z)$.

Specifically, the latent encoding z is obtained by sampling from a distribution parametrised by the encoder's outputs, μ and σ (see Fig. 1). If we were to sample directly from this distribution to get z, that computation would be non-deterministic. If we were to backpropagate through it, the

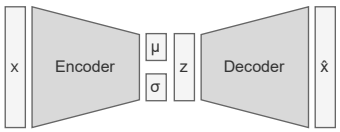


Fig. 1: VAE Architecture.

gradients would be random and hard to train with. For this reason, the reparamaterization trick is used:

$$z = \mu(x) + \epsilon\sigma(x), \tag{1}$$

where z is the sampled latent representation, μ and σ are functions of the input x, and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.

Variational inference is used to train such models, where the loss is the ELBO:

$$\mathcal{L}^{ELBO} = E_{q_\phi(z|x)}\big[\log(p_\theta(x|z))\big] - D_{KL}(q_\phi(z|x)||p(z)), \tag{2}$$

where $p(z)$ is the prior distribution of the latent representation, given as $p(z) = \mathcal{N}(0, \mathbf{I})$. In the first term of this equation, we have the error of the decoder given an encoding. In the second term, we have the Kullback-Leibler divergence (KLD) of the posterior distribution (z) and the prior. We can multiply the second term in ELBO by a hyper-parameter, β, to increase or decrease the importance of the KLD term. Such a VAE with the hyper-parameter β is known as a β-VAE [9] (see equation 3). Research has shown that when the KLD term is weighed higher, the representation is more likely to form a polarized regime [3]. In these cases, all active units will be disentangled [10], [11].

$$\mathcal{L}^{\beta\text{VAE}} = E_{q_\phi(z|x)}\big[\log(p_\theta(x|z))\big] - \beta D_{KL}(q_\phi(z|x)||p(z)). \tag{3}$$

B. Posterior Collapse

Posterior collapse occurs when the latent distribution $q(z|x)$ degenerates to the prior distribution $p(z)$. This is experienced in β-VAEs when the value of β in the ELBO is too high, causing it to over-regularise [12]. In the context of the polarized regime, it occurs when all variables become passive. Recently, [7] showed that posterior collapse occurs due to over-regularisation of the mean representation rather than the variance representation.

Initially, tracking the KL divergence between the latent distribution and the prior distribution can tell us if collapse has taken place [12]. However, this has been shown to not always determine posterior collapse [13]. For example, in [13], they trained a model with a fixed variance σ_z^2 but learnable mean μ_z. Equation 4 is the closed-form calculation for the KL divergence between the representation z and the normal prior. Even if the mean collapses to zero, if we fix σ_z^2 to a sufficiently high value, then that component of the ELBO may be higher than any D_{KL} threshold we use to define collapse. This shows that such a model will have a meaningful lower bound on its KL divergence.

$$D_{KL} = -0.5 \cdot \sum_z \big(1 + \log(\sigma_z^2) - \mu_z^2 - \sigma_z^2\big) \tag{4}$$

Hence, it was shown that these models can still experience posterior collapse even with a high KL divergence. The model will have collapsed by other metrics, but, since the variance is fixed, there may still be high KL divergence.

In practical settings, this shouldn't be the case as we won't be fixing the variance. That way, the variance can (and will) collapse to the prior's variance. However, an interesting observation is that this lends some evidence to the claim that the mean representation (μ) is what leads to posterior collapse rather than the variance representation (σ), as in [7].

A disadvantage to only tracking the KLD of the full posterior is that it does not explain what individual units are doing. In recent work, VAEs have been shown to learn a polarized regime [14], [3]. Here, individual units within the latent distributions adhere to strict modes, active and passive. Active variables are variables which are used by the decoder in reconstruction. Passive variables, on the other hand, are variables which are not used by the decoder in reconstruction. These variables will typically have zero mean and unit variance, in correspondence with the prior.

Given these two modalities, it is no longer as useful to measure the average KL divergence given the full representation. Instead, we care about what each unit is doing.

[6] amended the KL divergence measure to now detect if an individual unit has collapsed within the representation:

$$\mathbb{P}_{x\sim d}\big[D_{KL}(q(z_i|x)||p(z_i)) < \epsilon\big] \geq 1 - \delta. \tag{5}$$

If the KL divergence of a unit i falls below a threshold ϵ for a sufficiently high portion δ of the dataset, it will be labelled as collapsed. That way, we can track the percentage that a representation has collapsed as a function of its individual units.

[15] explored the polarized regimes further and split the variables into three categories: either they are passive, active, or mixed. Passive variables are those that are collapsed, and they are similar to the prior. Active variables are those that are not collapsed, and their distributions should be tight ($\sigma(Var) \approx 0$) since they need to convey as much information as possible with high precision. Mixed variables are those that flip between active and passive depending on the data examples. By tracking the numbers of active / mixed / passive variables, we can track the progression of posterior collapse. Once all the variables are passive, a model can be seen as having totally collapsed to the prior.

C. Variational Gradient Origin Network

As in VAEs, the goal of VGONs [1] is to transform data into a probabilistic encoding, and then to decode that back into a data distribution. Although the objectives are the same, VGON's approach to encoding is novel. Rather than using a neural network that is trained to encode the data, it uses just

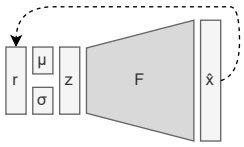


Fig. 2: VGON Architecture.

one network. In this sense, the VGON's single network has a dual purpose. It is both to produce the encoding and to decode it.

The encoding method proposed by [1] starts with a vector of zeros (r_0) which is used as an initial estimate of the representation r. Once passed through the network (denoted by $F(\cdot)$), we can find the loss. With this, the gradient of the loss with respect to the initial representation estimate is used as a new representation:

$$r = -\nabla_{r_0} \mathcal{L}^{ELBO}\big(x, F(r_0)\big). \tag{6}$$

This can be iterated many times, but it is argued in [1] that only one step of gradient descent is required to obtain an accurate latent distribution.

The vector r obtained by gradient descent gives a linear transformation of the input x into μ and σ, which are the parameters of the normal distribution representing the latent features. When this latent distribution is sampled, we obtain the sample representation z. Like VAEs, VGONs utilize the reparameterization trick. The sampled representation can be passed into the neural network to produce a final reconstruction of the original data $\hat{x}$. With this final reconstruction, we can perform normal backpropagation to update the weights of the network. Figure 2 visualises this model.

In essence, therefore, a VGON is very similar to a VAE. The network, F, acts as its decoder, while the gradient acts as its encoder.

III. Experiments

The VAE model consisted of a 4-layer convolutional and a 4-layer transposed convolutional neural network for the encoder and decoder respectively, plus a linear transformation from the encoder to the mean and variance representaiton. The VGON model consisted of a 4-layer transposed convolutional neural network plus a linear transformation from r_0 to the mean and variance representations. These models were trained for 200 epochs.

We trained a number of networks on the following datasets:

- Mnist [16]
- smallNorb [17]
- dSprites [18]

Mnist was used as a benchmark dataset. It is relatively simple, but with overlapping classes. smallNorb is a far more challenging dataset with a variety of ground-truth factors. dSprites was designed with disentanglement in mind. In dSprites, there are a small set of independent and deterministic ground truth factors which should ideally be captured by a disentangled representation.

For each dataset, we trained models for several values of the hyper-parameter β (0.5, 1.0, 5.0, 10.0, 20.0, 25.0, and 30.0) and ten different seeds. β was chosen to vary as it is the regularisation strength. We expect greater posterior collapse as β increases. Our values for β were particularly high, as we wanted to see what happens during total collapse.

While it has its limitations, we tracked the KL divergence of the representations. As we do not fix the variance, we expect to see coherence between the KLD metric and others. We additionally use the definitions of collapsed variables used in [15]. We opted to not include the KL divergence collapse metric from [6], viewing it as a more computationally expensive equivalent to the metrics in [15]. A mutual information metric [13] was also omitted from this study for computational reasons.

As we are examining the collapse of each unit in the polarized regime, it is worth looking at the distribution of each unit's two representations (mean and variance) for the full dataset. The purpose of showing these distributions is to show the shape of the three types of variable present in the polarized regime and verify that VGONs converge to similar representations to VAEs.

If we find that VGONs do collapse, and all the metrics are aligned, that indicates a clear similarity between VGON representations and VAE representations. So, to more closely examine the representations, we use linear Centered Kernel Alignment (CKA) [19]. This has been used to compare representations within neural networks. The results of this experiment should show if VGONs are, indeed, arriving at similar representations.

IV. Results

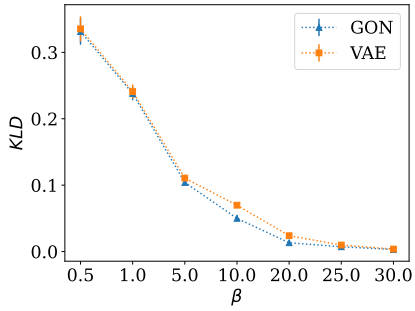
A. KL-Divergence

Firstly, we report the change of KL divergence when β changes. Figure 3 clearly shows that VGONs are experiencing KL collapse as KLD diminishes with higher β. This is to be expected. They also show, though, that VAEs and VGONs come to very similar values of KL divergence.

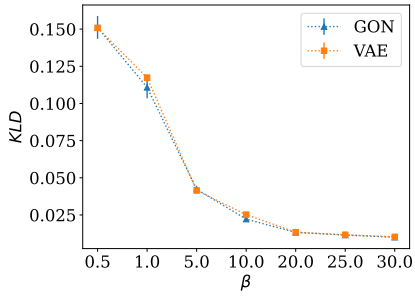
B. Polarized Regime

Figures 4 and 5 show the state of each variable given by [15]'s metrics. Variables are transitioning from active variables at low values of β to passive variables at high values of β. This is a strong indication of posterior collapse. The overall collapse of the models is the same. This can be identified by comparing the number of passive variables at a given value of β. The extent to which they collapse, however, varies between datasets. For instance, comparing 5c and 4c, we observe a slight difference at intermediate values of β. Additionally, in figure 4b the VGON trained on MNIST has far fewer passive variables than VAEs (figure 5b) until very large values of β where the models reach total collapse.

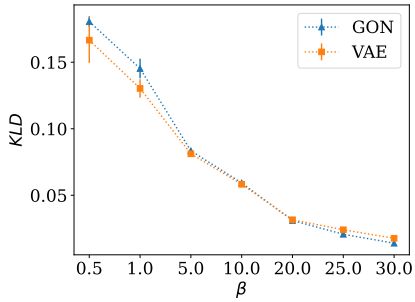
Crucially, there is no meaningful difference between the overall pattern of collapse between VAEs and VGONs. The number of active, mixed and passive variables is very similar throughout.



(a) Mnist



(b) smallNorb



(c) dSprites

Fig. 3: Change of KLD

C. Representation Histograms

Figures 6 to 9 display the distributions of individual units of the representation z. All figures are results from the Mnist dataset with $\beta = 5.0$ and the same seed with the exception of the mixed variables which were selected from the smallNorb dataset. We used smallNorb there for reasons explained below. Type 1 plots in figures 9a and 9c are from $\beta = 25.0$ and type 2 are from $\beta = 5.0$.

In figure 6, we can clearly see the representations for active variables. In line with [15], we observe the mean representation taking many values across the dataset, while the variance representation is incredibly tight around zero.

Passive variables, by contrast, have a very tight mean representation (figures 7a and 7b) around zero. The variance representation (figure 7c and 7d), also, is tight around one. This closely matches the prior, as the unit has collapsed.

Mixed variables have more diverse histograms (note that β of 25.0 induces more collapse). The mean representation (figure 8) appears to have reasonably high variance of the mean, a contrast to passive variables. However, the variance representation (figure 9) indicates two types of behaviour across the three datasets. Type 1 is the most common with $\beta = 25.0$, occurring for all the identified mixed variables in Mnist and dSprites, and in type 1 smallNorb variables shown in figures 9a and 9c. In this instance, the variance representation is picked at some intermediate point between 0.0 and 1.0. This behaviour indicates that with this high value of $\beta = 25.0$, the polarized regime may be violated. Type 1 variables do not transition from active to passive through a mixture (which would make them mixed as a result of polarized regime), but instead are progressing more directly towards the passive state, not adhering to the polarized regime. This explanation agrees with results in [8].

Type 2, on the other hand, does comply with our intuition for mixed variables [15] in polarized regimes [3]. This type is expressed strongly in the smallNorb dataset (type 2 in figures 9b and 9d with $\beta = 5.0$), and shows a clear mixture of two Gaussian distributions with one mode at 0.0 (these would be active) and another mode at 1.0 (these would be passive).

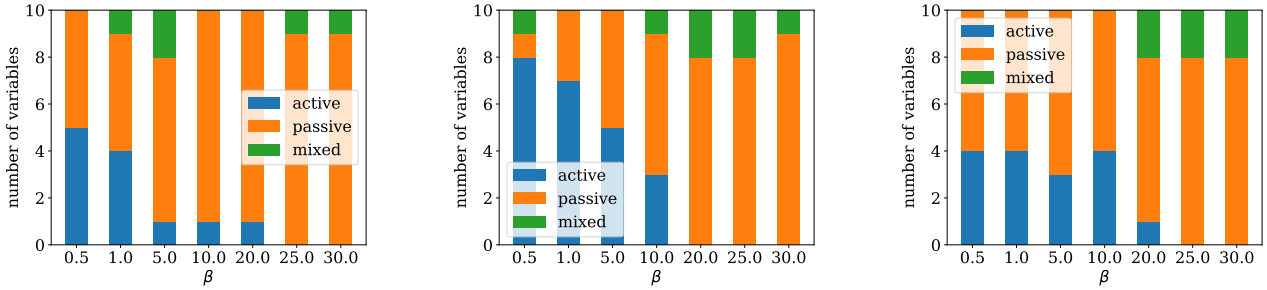
Overall, mixed variables with high β warrant further investigation with respect to polarized regime. For example, [14] say that polarized regime can be violated by 'bad local optima'. Thus, we could hypothesize that when β is excessively large (both in a VAE and VGON), type 1 behaviour may lead to a better local optimum than that of type 2 behaviour with a mixture distribution, even though the latter provides informative encoding for the data examples that are in the active part of the mixture [15]. Type 1 behaviour essentially shows that the polarized regime may not hold in the over-regularised case. A formal theoretical proof of this violation of polarized regime would be desirable.

Overall, the histograms reported in this section show that VGONs and VAEs are very closely aligned for each of the three key variable types. This indicates that they both experience a polarized regime.

D. Similarity of Representations

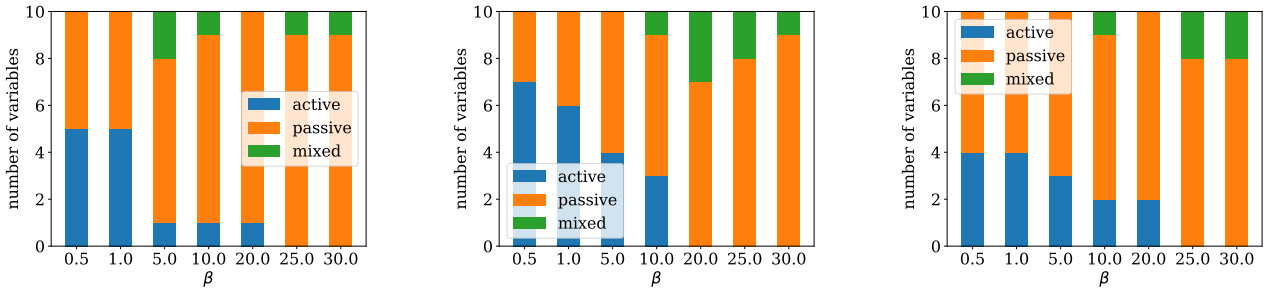
This experiment is intended to provide more explicit comparisons between the two networks, as very few differences between VAEs and VGONs have been found up to this point. We produced a series of similarity heatmaps that show the linear CKA between VGON and VAE representations at various values of β. For instance, in figure 10a the point (10.0, 5.0) shows the similarity between the mean representation of a VAE trained with $\beta = 10.0$ and the mean representation of a VGON trained with $\beta = 5.0$ for the Mnist dataset. In all heatmaps, darker colours express smaller similarity between the activations of the compared layers.

In the similarity heatmap for the mean representation, figure 10, one can observe the impact of β is consistent in each dataset. Along the columns and rows is the change of similarity for each model at β, with the other model constant. Along the diagonals, is the change of similarity of the two models with the same values of β.



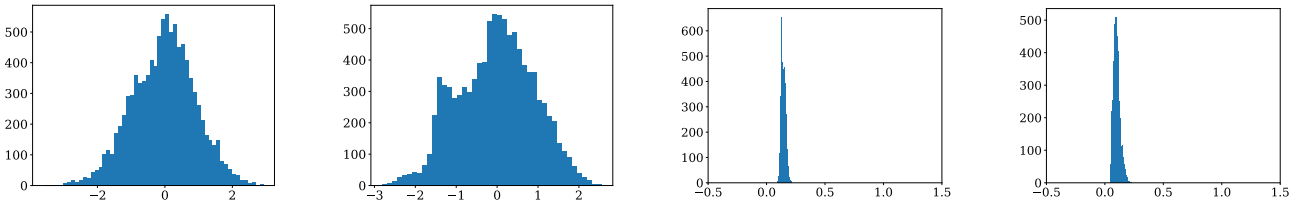
(a) smallNorb (b) Mnist (c) dSprites

Fig. 4: Collapsed variables VGON



(a) smallnorb (b) Mnist (c) dSprites

Fig. 5: Collapsed variables VAE



(a) VGON (b) VAE (c) VGON (d) VAE

Fig. 6: Typical active variable mean (left) and variance (right) representations with $\beta = 5.0$

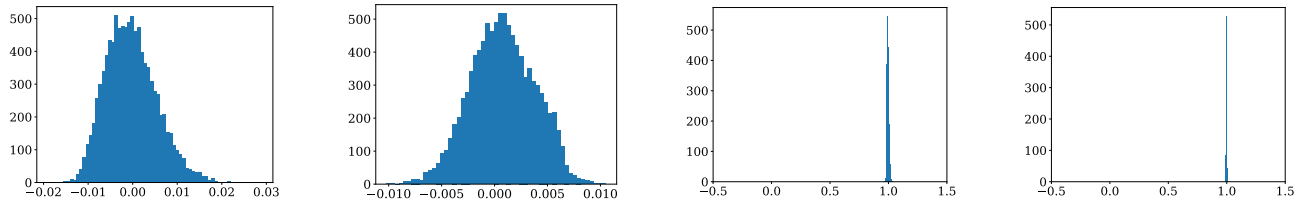
Excluding the dSprites dataset (figure 10b), each model's mean representation gradually becomes more dissimilar as the two values of β become further apart. In the bottom-right corner of figure 10c, we can see what happens when a model's posterior has fully collapsed, where the similarity of collapsed representations is high for a range of βs. When one model's β is equal to 20.0, and the other model's β is lower, there is considerable dissimilarity, which is the evidence of posterior collapse.

Meanwhile, looking along the diagonals, it can be seen that the model's representations are becoming more similar. This is consistent with our intuition of posterior collapse. When both models have collapsed, at about $\beta = 20.0$ in figure 10c, they have converged to the same (prior) distribution. The diagonal is still not explicit on the dSprites dataset (figure 10b). Here, the point where the two models are the most similar, apart from the trivial point at (30.0, 30.0), is at (1.0, 20.0). This could indicate that VAEs collapsed at a lower value of β. Overall, the mean representations for the two models are considerably similar to each other.

The variance representation (figure 11) paints a much different picture. In figures 11a and 11b, similarity is small for all values of β. In figure 11c, there is a pattern along the y-axis, indicating that the VGON isn't changing much compared to the VAE. At $\beta = 20.0$ there is a significant jump in the VAE's variance representation. Overall, there is very little relationship between the two models' variance representations.

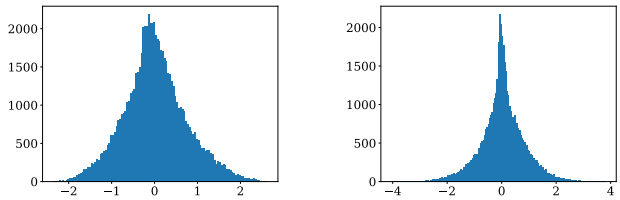
Additionally, we computed the similarity of each model's latent representations to itself (figures 12 – 14) in order to observe how they change as β is increased. The proceeding experiments give a possible explanation for the poor similarities in the variance representations (figure 11) but strong similarities in the mean representations (figure 10).

The mean representations behave as expected as β changes. The rate of change is mostly constant in the case of Mnist in figure 12a. However, as can be seen in figure 13a, once a



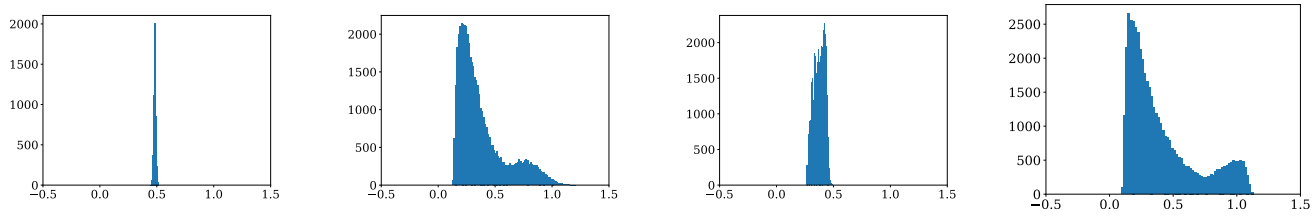
(a) VGON (b) VAE (c) VGON (d) VAE

Fig. 7: Typical passive variable mean (left) and variance (right) representations with $\beta = 5.0$.



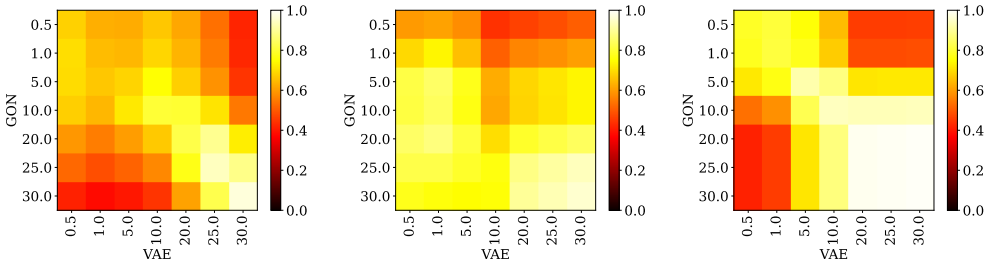
(a) VGON (b) VAE

Fig. 8: Typical mixed variable mean representation



(a) VGON type 1 (b) VGON type 2 (c) VAE type 1 (d) VAE type 2

Fig. 9: Typical mixed variable type 1 (left) and type 2 (right) variance representations



(a) Mnist (b) dSprites (c) smallNorb

Fig. 10: Mean representation similarity heatmaps for each dataset

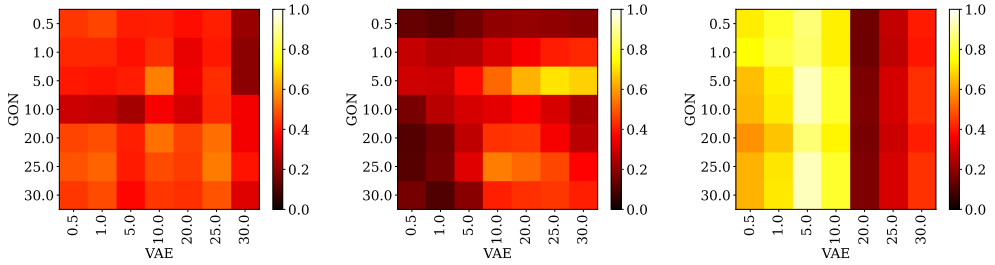
model has collapsed, its mean representation is collapsed and won't change much (bottom-right corner in figure 13a).

To contrast the rate of change seen by the mean representation, the variance representation changes in a slightly more extreme fashion. At $\beta = 10.0$ in figure 12c we see that suddenly the representation is very different. Following this, there is much less change.

The peculiar pattern in figure 11c can be seen as a reflection of the graphs in figures 13c and 13d. VGON's variance representation barely changes at all, so the only pattern expressed is from the change in the VAE's variance representation.

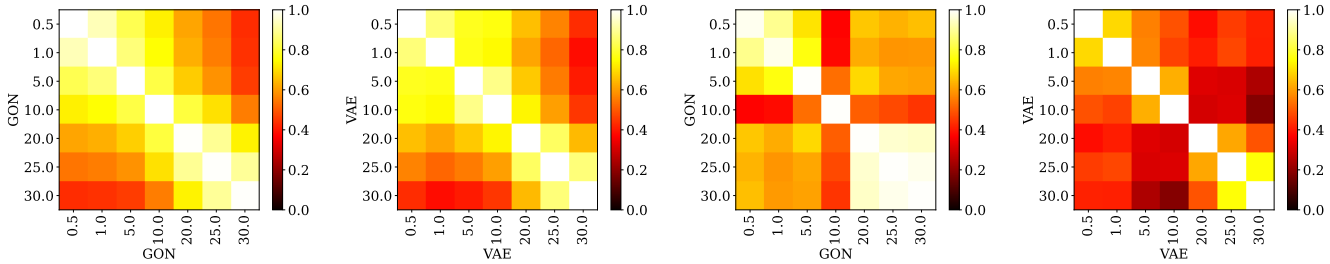
The variance representations of VGONs against themselves have values on the diagonal different from zero, which ordinarily should not be the case in CKA. However, as mentioned before, VGONs require one step of sampling to arrive at their representation. This results in the representation being non-deterministic, which explains why there can be a non-zero value along the diagonal.

Overall, our exploration of similarities between representations learned by VAEs and VGONs show that both models experience posterior collapse since the posteriors collapse to the prior and are very similar between those models. The



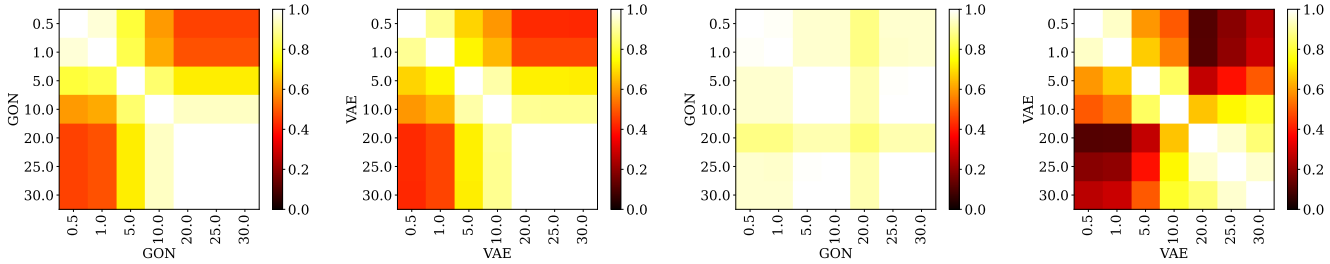
(a) Mnist (b) dSprites (c) smallNorb

Fig. 11: Variance representation similarity heatmaps for each dataset



(a) VGON (b) VAE (c) VGON (d) VAE

Fig. 12: Mean (left) and variance (right) representation self-similarity heatmaps on Mnist



(a) VGON (b) VAE (c) VGON (d) VAE

Fig. 13: Mean (left) and variance (right) representation self-similarity heatmaps on smallNorb

rate of collapse as a function of β is consistent in the mean representation, but the variance representations are starkly different.

There is one effect that is consistent across all datasets and both VGONs & VAEs. In the mean representation, we observe a gradual change up to a point where there is no more change. In the variance representation, there is a point where it suddenly changes, and subsequently stays the same.

To be clear, a value that has changed a lot will have a different CKA score. Reading along an axis, if the CKA score is suddenly low then it has changed. If the CKA score is high, it has not changed.

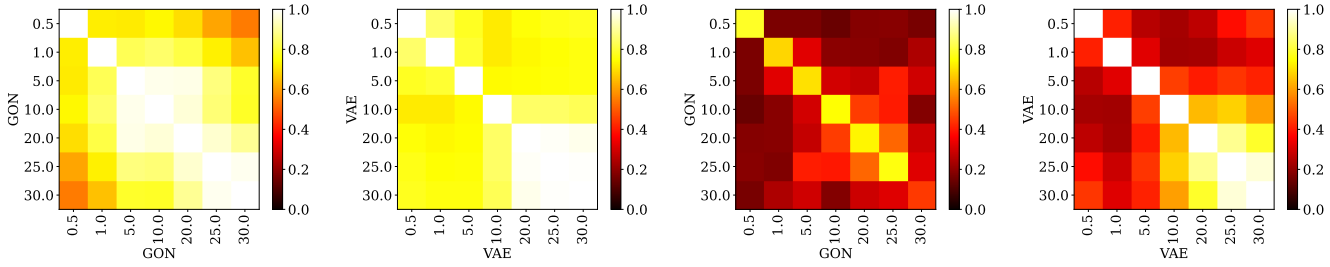
This pattern can be broadly described as the 'point of total collapse'. This occurs when the KLD component of the loss has become so large that most (or all) would-be active variables are not sufficiently expressive to justify their cost to the KLD. Hence, they are collapsed into passive variables. This results in a sharp change in the representation, as represented in the linear CKA. Once already collapsed, there is little more that can change between the representations, leading to the region of little change.

This 'point of total collapse' can be observed, also, in the KLD graphs and stacked graph. In the KLD graphs, when the KLD changes steeply this is when the overall representation is suddenly collapsing. In the stacked graph, the number of passive units increases and active units decreases at roughly this point.

While we observe and begin to define the 'point of total collapse', it is not within the scope of this work to fully explore this.

V. Conclusion

It is clear from our experiments that VGONs exhibit posterior collapse when β is sufficiently high. Not just in the decline in KLD, which is intuitive given the greater regularisation strength, but also in the collapse of individual



(a) VGON (b) VAE (c) VGON (d) VAE

Fig. 14: Mean (left) and variance (right) representation self-similarity heatmaps on dSprites

variables. As such, VGONs have also been shown to learn a polarized regime with moderate β. We have also shown that extremely high β used in VAEs and VGONs may destroy the polarized regime, since the variables transition from active to passive without the bi-modal mixture distribution. These are all important similarities to VAEs, especially given that VGONs and VAEs do not have the same mechanism for finding an encoding. This paper lays the foundation for future work identifying a theoretical link between the two models that could cause these shared behaviours.

VI. Acknowledgement

The authors would like to thank Lisa Bonheme for her contributions to the paper, from polarized regime guidance to CKA.

References

[1] S. Bond-Taylor and C. G. Willcocks, "Gradient Origin Networks," *arXiv preprint arXiv:2007.02798*, 2020.

[2] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *CoRR*, vol. abs/1312.6114, 2014.

[3] M. Rolinek, D. Zietlow, and G. Martius, "Variational Autoencoders Pursue PCA Directions (by Accident)," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12 398–12 407, 2018.

[4] J. He, D. Spokoyny, G. Neubig, and T. Berg-Kirkpatrick, "Lagging Inference Networks and Posterior Collapse in Variational Autoencoders," *ArXiv*, vol. abs/1901.05534, 2019.

[5] M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, pp. 183–233, 1999. [Online]. Available: https://api.semanticscholar.org/CorpusID:2073260

[6] J. Lucas, G. Tucker, R. Grosse, and M. Norouzi, "Understanding posterior collapse in generative latent variable models," in *DGS@ICLR*, 2019.

[7] Z. Wang and L. Ziyin, "Posterior Collapse of a Linear Latent Variable Model," *ArXiv*, vol. abs/2205.04009, 2022.

[8] B. Dai, Z. Wang, and D. Wipf, "The Usual Suspects? Reassessing Blame for VAE Posterior Collapse," in *ICML*, 2020.

[9] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," in *ICLR*, 2017.

[10] Y. Bengio, A. C. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1798–1828, 2013.

[11] F. Locatello, S. Bauer, M. Lucic, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations," *ArXiv*, vol. abs/1811.12359, 2019.

[12] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Józefowicz, and S. Bengio, "Generating Sentences from a Continuous Space," in *CoNLL*, 2016.

[13] Y. Takida, W.-H. Liao, T. Uesaka, S. Takahashi, and Y. Mitsufuji, "Preventing Posterior Collapse Induced by Oversmoothing in Gaussian VAE," *ArXiv*, vol. abs/2102.08663, 2021.

[14] B. Dai, Y. Wang, J. A. D. Aston, G. Hua, and D. P. Wipf, "Connections with Robust PCA and the Role of Emergent Sparsity in Variational Autoencoder Models," *J. Mach. Learn. Res.*, vol. 19, pp. 41:1–41:42, 2018.

[15] L. Bonheme and M. Grzes, "Be More Active! Understanding the Differences between Mean and Sampled Representations of Variational Autoencoders," *ArXiv*, vol. abs/2109.12679, 2021.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, pp. 2278–2324, 1998.

[17] Y. LeCun, F. J. Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 2, pp. II–104 Vol.2, 2004.

[18] L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner, "dSprites: Disentanglement testing Sprites dataset," https://github.com/deepmind/dsprites-dataset/, 2017.

[19] S. Kornblith, M. Norouzi, H. Lee, and G. E. Hinton, "Similarity of Neural Network Representations Revisited," *ArXiv*, vol. abs/1905.00414, 2019.