3rd Symposium on Advances in Approximate Bayesian Inference, 2020 1–10

Posterior Collapse and Latent Variable Non-identifiability

Anonymous Authors
Anonymous Institution

Abstract

Variational autoencoders (VAEs) model high-dimensional data by positing low-dimensional latent variables that are mapped through a flexible, implicit distribution parametrized by a neural network. Unfortunately, VAEs often suffer from posterior collapse: the posterior of the latent variables is equal to its uninformative prior, which renders the latent variables useless in producing meaningful representations. In this paper, we consider posterior collapse as a problem of identifiability. We prove that posterior collapses if and only if the latent variable is non-identifiable in the generative model. This result implies that posterior collapse is not a phenomenon specific to the use of neural networks or variational inference. Rather, it can occur in classical probabilistic models (e.g. Gaussian mixture models) even with exact inference, which we also demonstrate. Based on these insights, we propose a class of identifiable VAEs, which is as flexible as classical VAE while identifiable. This model class resolves the latent variable non-identifiability by leveraging the existence and uniqueness of monotone transport maps and parameterizing them with input convex neural networks. Across four datasets, we show that the identifiable VAEs mitigates posterior collapse.

1. Introduction

Variational autoencoders (VAEs) are a class of powerful generative model for high-dimensional data (Diederik et al., 2014; Rezende et al., 2014). Its key idea is to combine the inference principles of probabilistic modeling and the flexibility of neural networks (Johnson et al., 2016). In a VAE, each datapoint is independently generated by a low-dimensional latent variable drawn from a prior distribution, and then mapped through a flexible implicit distribution parametrized by a neural network. A VAE then employs variational inference to infer the posterior of these per-data-point latent variables. For each latent variable, it posits a variational approximating family (e.g. a Gaussian distribution) whose parameters are a neural network mapping of the corresponding datapoint. It then approximates the exact posterior by finding the member within this family that is closest to the exact posterior in Kullback-Leibler (KL) divergence.

Unfortunately, VAEs often suffer from posterior collapse, a phenomenon where the posterior of the latent variables is equal to their (pre-specified) prior (Bowman et al., 2016; Hoffman & Johnson, 2016; Sønderby et al., 2016; Kingma et al., 2016; Chen et al., 2016; Zhao et al., 2018; Yeung et al., 2017; Alemi et al., 2017; Lucas et al., 2019; Fu et al., 2019; Asperti, 2019; Li et al., 2019; Seybold et al., 2019; Dai et al., 2019; Zhao et al., 2020; Havrylov & Titov, 2020; Dieng et al., 2018; He et al., 2019; Kim et al., 2018; Razavi et al., 2019; Shu, 2016; Tomczak & Welling, 2017). This phenomenon is also know as latent variable collapse, KL vanishing, and over-pruning. Posterior collapse renders the VAEs—a latent variable model—useless in producing meaningful representations. The reason is that the per-data-point latent variables can no longer convey meaningful information about their corresponding data-points; all of them have the exact same posterior. Posterior collapse is commonly

observed in the VAEs whose generative model is highly flexible (Dai et al., 2019; Dieng et al., 2018; Bowman et al., 2016; Sønderby et al., 2016; Kingma et al., 2016; Chen et al., 2016; Zhao et al., 2018; Yeung et al., 2017).

Why does posterior collapse occur? Is it because the VAEs involve flexible neural networks? Is it because the VAEs performs variational inference? Can we avoid posterior collapse? In this paper, we answer these questions by considering posterior collapse as a problem of latent variable identifiability.

By appealing to the recent results in Bayesian non-identifiability (San Martın & González, 2010; Raue et al., 2013, 2009; Xie & Carlin, 2006; Poirier, 1998), we show that posterior collapse occurs if and only if the latent variable is non-identifiable in the generative model. Here is the intuition. Consider a dataset X and a latent variable Z. Loosely, a latent variable is non-identifiable (Poirier, 1998) when the likelihood of the generative model $p(x\,|\,z)$ does not depend on this latent variable Z, i.e. $p(x\,|\,z) \propto f(x)$ for some function f. Then, by the Bayes rule, its (exact) posterior is proportional to the product of the prior and the likelihood

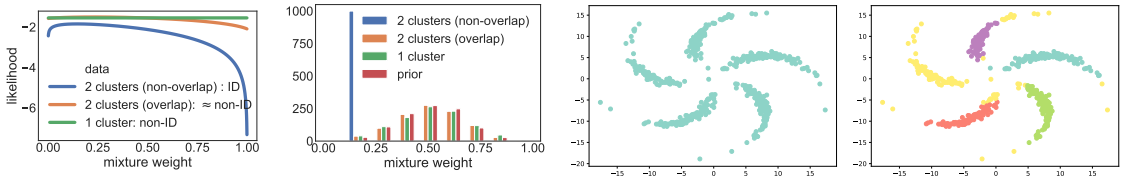
$$p(z\,|\,x) \propto p(z) \cdot p(x\,|\,z) \propto p(z) \cdot f(x) \propto p(z).$$

Therefore, the posterior must be equal to the prior because the likelihood is not dependent on Z.

This connection between posterior collapse and non-identifiability implies that posterior collapse is not a phenomenon specific to the use of neural networks or variational inference. Rather, it can occur in classical probabilistic models fitted with exact inference methods. As an example, the latent variables in Gaussian mixture model can be non-identifiable and suffer from posterior collapse when the model has more components than the true data generating model. Figure 1 illustrates such a Gaussian mixture model suffering from posterior collapse even with (near-)exact inference. Moreover, we show that VAEs suffer from posterior collapse for the same reason: the per-data-point latent variable is non-identifiable in VAEs.

Connecting posterior collapse to non-identifiability results in a natural solution to mitigating posterior collapse in VAEs: one must make the generative model identifiable. We propose a class of identifiable VAEs, which is as flexible as classical VAEs while also being identifiable. This model class resolves the latent variable non-identifiability by leveraging monotone transport maps (Peyré et al., 2019; McCann et al., 1995) and parameterizing them with input convex neural networks (Amos et al., 2017; Makkuva et al., 2019). Across four datasets, we show that the identifiable VAEs mitigates posterior collapse; see Section 1 for an example. (The supplementary material includes software that reproduces the studies.)

Contributions. We consider posterior collapse as a problem of latent variable non-identifiability. We show that posteriors collapse if and only if the latent variable in non-identifiable. It implies that posterior collapse is neither specific to flexible VAEs nor because of its variational approximation. Rather, it pertains to the structure of the generative model. We then propose a class of flexible and identifiable VAEs and demonstrate that it mitigates posterior collapse across four datasets. Appendix A discusses how these contributions situate among the related works.



Likelihood function | Posterior histogram | Non-ID GMVAE | ID GMVAE

Figure 1: When a latent variable is non-identifiable (non-ID) in a model, its likelihood function is a constant function and its posterior is equal to the prior, i.e. its posterior collapses. (a)-(b) Consider a Gaussian mixture model with two clusters $x \sim \alpha \cdot \mathcal{N}(\mu_1, \sigma_1^2) + (1-\alpha) \cdot \mathcal{N}(\mu_2, \sigma_2^2)$, treating the mixture weight α as the latent variable and others as parameters. Fit the model to datasets generated respectively by one Gaussian cluster (α non-identifiable), two overlapping Gaussian clusters (α nearly non-identifiable), and two non-overlapping Gaussian clusters (α identifiable). Under optimal parameters, the likelihood function $p(x \mid \alpha)$ is (close to) a constant when the latent variable α is (close to) non-identifiable; its posterior is also (close to) the prior. Otherwise, the likelihood function is non-constant and the posterior is peaked. (c)-(d) Fit a Gaussian mixture VAE (Shu, 2016; Kingma et al., 2014; Dilokthanakul et al., 2016) with five clusters to the pinwheel dataset (Johnson et al., 2016). Each data-point is colored by the category predicted by its corresponding latent variable. The categorical latent variables are non-identifiable (non-ID) in the vanilla Gaussian mixture VAE (GMVAE). Therefore, their posteriors collapse; they predict that all data-points belong to the same category. In Section 2.2, we propose an identifiable variant of the GMVAE. The posteriors of its latents do not collapse and produce a meaningful categorization of the data points.

2. Variational autoencoders, posterior collapse, and non-identifiability

Consider a dataset of n independent data-points $\boldsymbol{X} = (X_1, \ldots, X_n)$; each data-point is m-dimensional. Positing n latent variables $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$, a variational autoencoder (VAE) assumes that each data-point X_i is generated by a K-dimensional latent variable Z_i:

$$Z_i \sim p(z_i), \qquad X_i \mid Z_i \sim p(x_i \mid z_i; \theta), \tag{1}$$

where θ are parameters of the likelihood function $p(x_i \mid z_i; \theta)$. Equation (1) constitutes the generative model of a VAE. It also encompasses classical probabilistic models like Gaussian mixture model (GMM) and probabilistic principle component analysis (PPCA).

To perform inference, a VAE attempts to optimize the parameters θ by maximizing the log marginal likelihood, and then infer the posterior of the latent variables $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ at the maximum likelihood (ML) parameters θ^*. Formally, the ML parameters solve $\theta^* = \arg\max_\theta \log p(\boldsymbol{x}; \theta) = \arg\max_\theta \sum_{i=1}^n \log \int p(x_i \mid z_i; \theta) p(z_i) \mathrm{d}z_i$. Then the exact posterior of the latent variables are computed at the ML parameters, $p(\boldsymbol{z} \mid \boldsymbol{x}; \theta^*) = \prod_{i=1}^n p(z_i \mid x_i; \theta^*)$.

However, the integral $\int p(x_i \mid z_i; \theta) p(z_i) \mathrm{d}z_i$ in computing the ML parameters is often intractable in practice. Therefore, VAEs approximate this integral with variational inference. In the ideal case where the variational approximation is exact, the usual VAE objective (Diederik et al., 2014) coincides with maximizing log marginal likelihood. Moreover, the approximate posterior of the latents $q(\boldsymbol{z} \mid \boldsymbol{x})$ also coincides with the exact posterior $p(\boldsymbol{z} \mid \boldsymbol{x}; \theta^*)$. (We prove this fact in Appendix B.) We focus on this ideal case below, abstracting away computational considerations.

2.1. Posterior collapse and latent variable non-identifiability

We now discuss posterior collapse and its connections to latent variable non-identifiability.

Posterior collapse is a phenomenon where the posterior of the latents in a VAE is equal to its uninformative prior

$$p(\boldsymbol{z} \mid \boldsymbol{x}; \theta^*) = p(\boldsymbol{z}). \tag{2}$$

To be clear, most of the literature defines this phenomenon with respect to the approximate posterior, but we focus on the ideal case where the approximation is exact. This phenomenon is commonly observed in VAEs with a flexible likelihood, i.e. a flexible neural network of f_θ (Bowman et al., 2016; Sønderby et al., 2016; Kingma et al., 2016; Chen et al., 2016; Zhao et al., 2018; Yeung et al., 2017).

When posterior collapse occurs, it prevents the latent variable from providing meaningful low-dimensional representations of the high-dimensional data-points. For example, if Z_i is categorical, posterior collapse implies that its posterior is equal to $\prod_{i=1}^{n} \text{Categorical}(z_i; 1/K)$. Treating the latent variable Z_i's as a summarization of the corresponding data-point X_i, it says all data-points have an equal chance to belong to each of the K categories. Such a representation does not distinguish among different data-points and does not summarize the data-points in a meaningful way. In this way, the VAE only performs density estimation, which defeats the purpose of density estimation specifically through latent variable modeling.

Why does posterior collapse occur? Below we provide an explanation of posterior collapse via a connection to latent variable non-identifiability. We will show that posterior collapse (Equation (2)) is equivalent to the latent variable $\boldsymbol{z}$ being non-identifiable in the model. We first define latent variable non-identifiability.

Definition 1 (Latent variable non-identifiability) *Consider a likelihood $p(\boldsymbol{x} \mid \boldsymbol{z}; \theta)$ where $\boldsymbol{X}$ is the dataset, $\boldsymbol{Z}$ is the latent variables, and θ is the parameter to be optimized. The latent variable $\boldsymbol{Z}$ is non-identifiable if there exists a $\theta^* \in \Theta$ such that, for a given dataset $\boldsymbol{x} = \{x_1, \ldots, x_n\}$,*

$$p(\boldsymbol{x} \mid \boldsymbol{z} = \boldsymbol{z}^1; \theta^*) = p(\boldsymbol{x} \mid \boldsymbol{z} = \boldsymbol{z}^2; \theta^*) \qquad \forall \boldsymbol{z}^1, \boldsymbol{z}^2 \in \mathcal{Z}, \tag{3}$$

where $\log p(\boldsymbol{x} \mid \theta^*) = \max_{\theta \in \Theta} \log p(\boldsymbol{x} \mid \theta)$.

Loosely, this definition says that the latent variable $\boldsymbol{Z}$ is non-identifiable when the likelihood function $p(\boldsymbol{x} \mid \boldsymbol{z}; \theta^*)$ does not depend on the latent, i.e. constant in $\boldsymbol{z}$. Equation (3) implies that the conditional likelihood given $\boldsymbol{Z}$ must be equal to the marginal, $p(\boldsymbol{x} \mid \boldsymbol{z} = \boldsymbol{z}^1; \theta^*) = p(\boldsymbol{x}; \theta^*) \quad \forall \boldsymbol{z}^1 \in \mathcal{Z}$.

We note that Theorem 1 only requires Equation (3) be true for a given realization of the dataset $\boldsymbol{X} = \boldsymbol{x}$. Thus we should expect that a latent variable may be identifiable or not depending on the combination of the dataset $\boldsymbol{x}$ and the model $p(\boldsymbol{x} \mid \boldsymbol{z}; \theta)$. A latent variable may be identifiable in a model given one dataset but not another. Moreover, Equation (3) only consider the value of the parameters θ at its ML value θ^*, i.e. when the parameter θ^* maximizes the log marginal likelihood $\log p(\boldsymbol{x} \mid \theta)$. The behavior of $p(\boldsymbol{x} \mid \boldsymbol{z}; \theta)$ at other values of θ is unconstrained. Finally, Theorem 1 is closely related to the definition of $\boldsymbol{Z}$ being conditionally non-identifiable (or conditionally uninformative) given θ^* (San Martın & González, 2010; Raue et al., 2013, 2009; Xie & Carlin, 2006; Poirier, 1998).

Next we show that posterior collapse is equivalent to latent variable non-identifiability.

Theorem 2 (Posterior collapse ⇔ non-identifiability) *The posterior of the latent variable $\boldsymbol{z}$ collapses with positive probability if and only if the latent variable $\boldsymbol{Z}$ is non-identifiable.*

Proof First suppose the ML parameters θ^* are unique. Then, by the Bayes rule, posterior collapse (Equation (2)) implies

$$p(\boldsymbol{z}) = p(\boldsymbol{z}\,|\,\boldsymbol{x};\theta^*) \propto p(\boldsymbol{z}) \cdot p(\boldsymbol{x}\,|\,\boldsymbol{z};\theta^*). \tag{4}$$

Equation (4) is equivalent to $p(\boldsymbol{x}\,|\,\boldsymbol{z};\theta^*)$ being a constant function in $\boldsymbol{z}$, i.e. the latent variable $\boldsymbol{z}$ being non-identifiable as in Theorem 1. To see it intuitively, note that $p(\boldsymbol{z}) \propto p(\boldsymbol{z})p(\boldsymbol{x}\,|\,\boldsymbol{z};\theta^*)$ holds when $p(\boldsymbol{x}\,|\,\boldsymbol{z};\theta^*)$ is constant in $\boldsymbol{z}$. Moreover, if $p(\boldsymbol{x}\,|\,\boldsymbol{z};\theta^*)$ non-trivially depend on $\boldsymbol{z}$, then $p(\boldsymbol{z})$ must be different from $p(\boldsymbol{z})p(\boldsymbol{x}\,|\,\boldsymbol{z};\theta^*)$ as a function of $\boldsymbol{z}$. Therefore, posterior collapse occurs if and only if the latent variable is non-identifiable.

Next, suppose the ML parameters θ^* are not unique; there are multiple values of θ^* that maximize $\log p(\boldsymbol{x}\,|\,\theta)$. Then one of them must satisfy Equation (4) and be equivalent to posterior collapse, due to latent variable $\boldsymbol{Z}$ being non-identifiable. Because there is a positive probability for any of these ML parameters to be reached in maximizing log marginal likelihood, there is a positive probability of posterior collapse. ■

The proof of Theorem 2 may seem straightforward. But this simple argument shows that it is essential to understand posterior collapse from the standpoint of the model and the data, rather than inference or optimization. Below we give examples to illustrate this equivalence between posterior collapse and latent variable non-identifiability.

Example 3 (Gaussian mixture VAE (GMVAE)) *Consider a* GMVAE*,*

$$Z_i \sim \text{Categorical}(1/K), \qquad X_i\,|\,Z_i \sim \mathcal{N}(f_\theta(\mathcal{N}(\mu_{z_i}, \Sigma_{z_i})), \sigma^2 \cdot I_m),$$

where the parameters are $\theta = (f_\theta, \{\mu_k, \Sigma_k\}_{k=1}^K)$; μ_k's are d-dimensional and Σ_k are $d \times d$-dimensional. Fit the model to a dataset drawn from the same GMVAE*.*

Suppose the neural network function f_θ is fully flexible. One optimizer of the log marginal likelihood is $\theta^* = (f_\theta^*, \{\mu_k^*, \Sigma_k^*\}_{k=1}^K) = (f_\theta^*, \{0, I_d\}_{k=1}^K)$, where

$$\mathcal{N}(f_\theta^*(\mathcal{N}(0, I_d)), \sigma^2 \cdot I_m) = \int \mathcal{N}(f_\theta(\mathcal{N}(\mu_{z_i}, \Sigma_{z_i})), \sigma^2 \cdot I_m) \cdot p(z_i)\,\mathrm{d}z_i.$$

That is, the ML *f^* produces the same distribution as the original* GMVAE *by mapping from a single standard Gaussian cluster, as opposed to the original mixture of K clusters. Under this* ML *parameter θ^*, the latent variable Z_i is non-identifiable because the K clusters are the same, similar to Example 1.1. Hence the posterior of the latent variable Z_i also collapses; Section 1 illustrates a fit of this (non-identifiable)* GMVAE *to the pinwheel data (Johnson et al., 2016).*

Together with two other examples of GMM and PPCA in Appendix C, it illustrates different ways that a latent variable can be non-identifiable in a model and suffer from posterior collapse. They illustrate that even exact inference methods can not prevent posterior collapse in non-identifiable models (Sections 1 and 2.2 and theorem 6). Therefore, posterior collapse is an intrinsic problem of the model and the data, rather than specific to the use of neural networks or variational inference in

VAEs, or any inference algorithms. The equivalence between posterior collapse and latent variable non-identifiability in Theorem 2 also implies that, to mitigate posterior collapse, we should try to resolve latent variable non-identifiability.

2.2. Identifiable VAEs via monotone transport maps

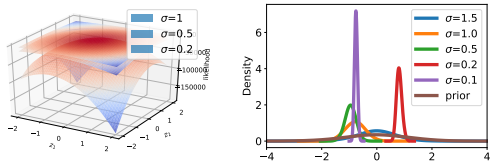
We develop a class of identifiable VAEs to mitigate posterior collapse. To resolve latent variable non-identifiability in VAEs, we propose to use monotone transport maps (Ball, 2004). For example, a transport map that maps everything to the right is monotone, i.e. its mapping preserves the ordering of the data points. A monotone transport map between two distributions is, under weak conditions, guaranteed to exist and be unique (McCann et al., 1995). This property will enable us to resolve non-identifiability in VAEs and mitigate posterior collapse.

Definition 4 (Identifiable VAEs via monotone transport maps) *An identifiable* VAE *via monotone transport maps generates an m-dimensional data-point X_i from the following process:*

$$Z_i \sim p(z_i), X_i \,|\, Z_i ; \theta \sim \text{expfam}(h \circ g_\theta(\text{expfam}(z_i^\top \beta_\theta ; \gamma_\theta)) ; \lambda), \tag{5}$$

where Z_i is a K-dimensional latent variable. The parameters of the model are $\theta = (g_\theta, \beta_\theta, \gamma_\theta)$, where β_θ is $K \times m$-dimensional matrix whose first column must be positive and strictly increasing, i.e. $0 < \beta_\theta[0,0] < \beta_\theta[1,0] < \ldots < \beta_\theta[K,0]$, and $g_\theta : \mathcal{Z}^m \to \mathcal{Z}^m$ is a monotone transport map. The function $h(\cdot)$ is a one-to-one link function for the exponential family, e.g. the sigmoid function.

This class of generative model emulates many existing VAEs, including exponential family mixtures or vanilla VAEs. How does this model guarantee identifiability? It is for two reasons: the positive and strictly increasing requirement on β_θ and the monotone transport map requirement on g_θ. The first requirement on β_θ prevents Z_i from being non-identifiable due to indistinguishable latent clusters or zero latent dimensions (cf. Examples 1 and 2). The second requirement on g_θ guarantees that g_θ must be the unique monotone transport map from p_1 to p_2, due to the key result of McCann et al. (1995), which shows that monotone transport maps between probability distributions must be unique under weak conditions. Appendix D further discusses the theoretical and practical aspects of the identifiable VAEs via monotone transport maps. Appendix E demonstrates its empirical performance.



Likelihood function Posteriors

Figure 2: As the noise level increases in PPCA, the latent variable becomes closer to non-identifiable and more susceptible to posterior collapse. Its likelihood surface becomes flatter and its posterior becomes closer to the prior.

3. Discussion

We study the posterior collapse phenomenon from the perspective of latent variable non-identifiability. We show that posterior collapses if and only if the latent variable is non-identifiable in a probabilistic model. It shows that posterior collapse is not specific to the use of neural networks or particular inference algorithms in VAEs. Rather, it is an intrinsic issue of the model and the dataset.

References

Alemi, A. A., Poole, B., et al. (2017). Fixing a broken elbo. *arXiv preprint arXiv:1711.00464*.

Amos, B., Xu, L., & Kolter, J. Z. (2017). Input convex neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (pp. 146–155).: JMLR. org.

Asperti, A. (2019). Variational autoencoders and the variable collapse phenomenon. *Sensors & Transducers*, 234(6), 1–8.

Ball, K. (2004). An elementary introduction to monotone transportation. In *Geometric aspects of functional analysis* (pp. 41–52). Springer.

Bowman, S., Vilnis, L., et al. (2016). Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning* (pp. 10–21).

Chen, X., Kingma, D. P., et al. (2016). Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*.

Dai, B., Wang, Z., & Wipf, D. (2019). The usual suspects? reassessing blame for vae posterior collapse. *arXiv preprint arXiv:1912.10702*.

Diederik, P. K., Welling, M., et al. (2014). Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, volume 1.

Dieng, A. B., Kim, Y., Rush, A. M., & Blei, D. M. (2018). Avoiding latent variable collapse with generative skip models. *arXiv preprint arXiv:1807.04863*.

Dilokthanakul, N., Mediano, P. A., et al. (2016). Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*.

Fu, H., Li, C., et al. (2019). Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*.

Havrylov, S. & Titov, I. (2020). Preventing posterior collapse with levenshtein variational autoencoder. *arXiv preprint arXiv:2004.14758*.

He, J., Spokoyny, D., Neubig, G., & Berg-Kirkpatrick, T. (2019). Lagging inference networks and posterior collapse in variational autoencoders. *arXiv preprint arXiv:1901.05534*.

Hoffman, M. D. & Johnson, M. J. (2016). Elbo surgery: yet another way to carve up the variational evidence lower bound.

Johnson, M. J., Duvenaud, D. K., Wiltschko, A., Adams, R. P., & Datta, S. R. (2016). Composing graphical models with neural networks for structured representations and fast inference. In *Advances in neural information processing systems* (pp. 2946–2954).

Kim, Y., Wiseman, S., Miller, A., Sontag, D., & Rush, A. (2018). Semi-amortized variational autoencoders. In *International Conference on Machine Learning* (pp. 2678–2687).

Kingma, D. P., Mohamed, S., Rezende, D. J., & Welling, M. (2014). Semi-supervised learning with deep generative models. In *Advances in neural information processing systems* (pp. 3581–3589).

Kingma, D. P., Salimans, T., et al. (2016). Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems* (pp. 4743–4751).

Li, B., He, J., Neubig, G., Berg-Kirkpatrick, T., & Yang, Y. (2019). A surprisingly effective fix for deep latent variable modeling of text. *arXiv preprint arXiv:1909.00868*.

Lucas, J., Tucker, G., Grosse, R. B., & Norouzi, M. (2019). Don't blame the elbo! a linear vae perspective on posterior collapse. In *Advances in Neural Information Processing Systems* (pp. 9403–9413).

Makkuva, A. V., Taghvaei, A., Oh, S., & Lee, J. D. (2019). Optimal transport mapping via input convex neural networks. *arXiv preprint arXiv:1908.10962*.

McCann, R. J. et al. (1995). Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80(2), 309–324.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6), 355–607.

Poirier, D. J. (1998). Revising beliefs in nonidentified models. *Econometric Theory*, 14(4), 483–509.

Raue, A., Kreutz, C., et al. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15), 1923–1929.

Raue, A., Kreutz, C., Theis, F. J., & Timmer, J. (2013). Joining forces of bayesian and frequentist methodology: a study for inference in the presence of non-identifiability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984), 20110544.

Razavi, A., Oord, A. v. d., Poole, B., & Vinyals, O. (2019). Preventing posterior collapse with delta-vaes. *arXiv preprint arXiv:1901.03416*.

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*.

San Martın, E. & González, J. (2010). Bayesian identifiability: Contributions to an inconclusive debate. *Chilean Journal of Statistics*, 1(2), 69–91.

Seybold, B., Fertig, E., Alemi, A., & Fischer, I. (2019). Dueling decoders: Regularizing variational autoencoder latent spaces. *arXiv preprint arXiv:1905.07478*.

Shu, R. (2016). Gaussian mixture VAE: Lessons in variational inference, generative models, and deep nets.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., & Winther, O. (2016). How to train deep variational autoencoders and probabilistic ladder networks. In *33rd International Conference on Machine Learning (ICML 2016)*.

Tomczak, J. M. & Welling, M. (2017). Vae with a vampprior. *arXiv preprint arXiv:1705.07120*.

Xie, Y. & Carlin, B. P. (2006). Measures of bayesian learning and identifiability in hierarchical models. *Journal of Statistical Planning and Inference*, 136(10), 3458–3477.

Yeung, S., Kannan, A., Dauphin, Y., & Fei-Fei, L. (2017). Tackling over-pruning in variational autoencoders. *arXiv preprint arXiv:1706.03643*.

Zhao, T., Lee, K., & Eskenazi, M. (2018). Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1098–1107).

Zhao, Y., Yu, P., Mahapatra, S., Su, Q., & Chen, C. (2020). Discretized bottleneck in vae: Posterior-collapse-free sequence-to-sequence learning. *arXiv preprint arXiv:2004.10603*.

Appendix A. Related work

Our work draws on two themes around generative models.

The first is a body of work on algorithms that mitigate posterior collapse in VAEs. Many works have proposed methods to avoid posterior collapse in the VAE training. These methods often focus on the optimization perspective of VAEs and modify the optimization objective or the optimization algorithm in variational inference (Bowman et al., 2016; Hoffman & Johnson, 2016; Sønderby et al., 2016; Kingma et al., 2016; Chen et al., 2016; Zhao et al., 2018; Yeung et al., 2017; Alemi et al., 2017; Fu et al., 2019; Asperti, 2019; Li et al., 2019; Seybold et al., 2019; Zhao et al., 2020; Havrylov & Titov, 2020; Dieng et al., 2018; He et al., 2019; Kim et al., 2018; Razavi et al., 2019; Shu, 2016; Tomczak & Welling, 2017). More recently, a few works try to provide explanations for the posterior collapse phenomenon. For example, (Dai et al., 2019) shows that posterior collapse can be partially attributed to the local optima in training VAEs with deep neural networks; (Lucas et al., 2019) shows that posterior collapse is not specific to the variational inference training objective; absent variational approximation, the log marginal likelihood of PPCA has bad local optima that can lead to posterior collapse. Related to these works, this paper try to both provide explanations and propose solutions to posterior collapse. However, in contrast to the common optimization perspective, this paper focuses on the modeling perspective. It studies the connection between posterior collapse and the non-identifiability of latent variables in probabilistic models, including both the classical ones like GMM and the modern ones like VAEs. This modeling perspective characterizes global optimal solutions, abstracting away optimization challenges like local optima and saddle points.

The second theme is the latent variable identifiability in probabilistic models. Identifiability of latent variables has long been studied in the statistics literature, including (San Martın & González, 2010; Raue et al., 2013, 2009; Xie & Carlin, 2006; Poirier, 1998). More recently, Betancourt (2017) studies the effect of latent variable identifiability in GMM on Bayesian computation. Khemakhem et al. (2019) studies the non-identifiability of deep latent variable models and its effect on disentanglement. Moreover, it proposes to resolve non-identifiability by appealing to external data. Related to these works, this work demonstrates posterior collapse as one additional aspect in which the classical concept of identifiability can play a key role in modern probabilistic modeling. It also opens the door to many new solutions to posterior collapse via existing techniques of resolving non-identifiability.

Appendix B. VAEs approximately maximize marginal likelihood

To perform inference, a VAE attempts to optimize the parameters θ by maximizing the log marginal likelihood, and then infer the posterior of the latent variables $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ at the ML parameters θ^*. Formally, the ML parameters solve

$$\theta^* = \arg\max_{\theta} \log p(\boldsymbol{x}; \theta) = \arg\max_{\theta} \sum_{i=1}^{n} \log \int p(x_i \mid z_i; \theta) p(z_i) \mathrm{d}z_i. \tag{6}$$

Then the exact posterior of the latent variables are computed at the ML parameters,

$$p(\boldsymbol{z} \mid \boldsymbol{x}; \theta^*) = \prod_{i=1}^{n} p(z_i \mid x_i; \theta^*) = \prod_{i=1}^{n} \frac{p(z_i) p(x_i \mid z_i; \theta^*)}{\int p(x_i \mid z_i; \theta^*) p(z_i) \mathrm{d}z_i}. \tag{7}$$

The integral $\int p(x_i \mid z_i; \theta) p(z_i) \mathrm{d}z_i$ in Equations (6) and (7) is often intractable. So a VAE employs variational approximations to approximate this integral. It first posits an approximating family $\mathcal{Q}^n$,

$$\mathcal{Q}^n = \{q_\phi(\{z_i\}_{i=1}^n) = \prod_{i=1}^n q_\phi(z_i \mid x_i) : \phi \in \Phi\}, \tag{8}$$

where $q_\phi(\cdot)$ is a neural network mapping with known structure and parameters ϕ. For example, $q_\phi(z_i \mid x_i) = \mathcal{N}(z_i; g_\phi(x_i), I_K)$. It then approximates the log marginal $\log p(x_i; \theta)$ with a lower bound

$$\log p(x_i; \theta) \geq \log p(x_i; \theta) - \min_{q_{\phi(\theta)} \in \mathcal{Q}^n} \mathrm{KL}(q_{\phi(\theta)}(z_i \mid x_i) || p(z_i \mid x_i; \theta)), \tag{9}$$

where the right side of the inequality is the evidence lower bound (ELBO), which is easier to maximize (Blei et al., 2017). The parameters $\phi(\theta)$ indicates that the parameters of the approximate distribution ϕ can take different values with different θ. This inequality becomes equality when the approximating family $\mathcal{Q}^n$ contains all distributions. In this case, the approximate $q_\phi(z_i)$ also coincide with the exact posterior $p(z_i \mid x_i; \theta)$.

Finally, the VAE approximately optimizes the parameters

$$\theta^* = \arg\max_\theta \left[\log p(x_i; \theta) - \min_{q_{\phi(\theta)} \in \mathcal{Q}^n} \mathrm{KL}(q_{\phi(\theta)}(z_i \mid x_i) || p(z_i \mid x_i; \theta)) \right]. \tag{10}$$

This nested optimization can be equivalently unpacked into a single-level optimization (Murphy & Van der Vaart, 2000)

$$q_\phi^*, \theta^* = \arg\max_{q_\phi \in \mathcal{Q}^n, \theta} \left[\log p(x_i; \theta) - \mathrm{KL}(q_{\phi(\theta)}(z_i \mid x_i) || p(z_i \mid x_i; \theta)) \right] \tag{11}$$

$$= \arg\max_{q_\phi \in \mathcal{Q}^n, \theta} \sum_{i=1}^n \mathbb{E}_{q_\phi(z_i)} \left[\log p(x_i, z_i; \theta) - \log q_\phi(z_i) \right], \tag{12}$$

where Equation (12) is the usual VAE objective (Diederik et al., 2014). It is equivalent to Equation (6).

Appendix C. Examples of posterior collapse ⇔ latent variable non-identifiability

Example 5 (Gaussian mixture model (GMM)) *Consider a Gaussian mixture model with two clusters,*

$$p(\alpha) = \mathrm{Beta}(\alpha; 5, 5), \qquad p(x_i \mid \alpha; \theta) = \alpha \cdot \mathcal{N}(x_i; \mu_1, \sigma_1^2) + (1 - \alpha) \cdot \mathcal{N}(x_i; \mu_2, \sigma_2^2).$$

Here α is the latent variable and $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2)$ are the parameters of the model. Fit this model to three datasets, each with 10^5 samples.

1. *Samples from a single Gaussian distribution, $X_i \sim \mathcal{N}(-1, 1)$.*

   The latent variable α is non-identifiable in this case. The reason is that one set of ML *parameters is $\theta^* = (\mu_1^*, \mu_2^*, \sigma_1^*, \sigma_2^*) = (-1, -1, 1, 1)$, i.e. setting both of the two mixture components equal*

*to the true data generating Gaussian distribution. Under this θ^*, the latent variable α is non-identifiable and its likelihood function $p(\{x_i\}_{i=1}^n \mid \alpha; \theta^*)$ is constant in α because the two mixture components are equal; Section 1 illustrates this fact. Moreover, the posterior of α collapses, $p(\alpha \mid \{x_i\}_{i=1}^n; \theta^*) = p(\alpha)$. Section 1 illustrates this fact: The Hamiltonian Monte Carlo (*HMC*) samples of the α posterior closely match those drawn from the prior. (Exact inference is intractable in this case, so we use* HMC *as a close approximation to exact inference.) This example demonstrates the connection between non-identifiability and posterior collapse; it also shows that posterior collapse is not specific to variational inference but is an issue of the model and the data.*

2. *Samples from a mixture of two overlapping clusters, $X_i \sim 0.15 \cdot \mathcal{N}(-0.5, 1) + 0.85 \cdot \mathcal{N}(0.5, 1)$.*

   The latent variable α is identifiable in this case. However, it is nearly non-identifiable. While the two data generating clusters are different, they are very similar to each other because they overlap. Therefore, the likelihood function $p(x_i \mid \alpha; \theta^*)$ is slowly varying under ML *parameters $\theta^* = (\mu_1^*, \mu_2^*, \sigma_1^*, \sigma_2^*) = (-0.5, 0.5, 1, 1)$; see Section 1. Consequently, the posterior of α remains very close to the prior; see Section 1.*

3. *Samples from a mixture of two non-overlapping clusters, $X_i \sim 0.15 \cdot \mathcal{N}(-10, 1) + 0.85 \cdot \mathcal{N}(10, 1)$.*

   The latent variable α is identifiable in this case. The two data generating clusters are substantially different, so the likelihood function varies across $\alpha \in [0, 1]$ under the ML *parameters (Section 1). The posterior of α is also peaked (Section 1) and differs much from the prior.*

Example 6 (Probabilistic principle component analysis (PPCA)) *Consider a* PPCA *with two latent dimensions,*

$$p(z_i) = \mathcal{N}(z_i; 0, I_2), \qquad p(x_i \mid z_i; \theta) = \mathcal{N}(x_i; z_i^\top w, \sigma^2 \cdot I_5),$$

where z_i's are the latent variables of interest and others $\theta = (w, \sigma^2)$ are parameters of the model. Fit this model to two datasets, each with 500 samples.

1. *Samples from a one-dimensional* PPCA, *$X_i \sim \mathcal{N}(x_i; \mathcal{N}(0, I_1) \cdot \bar{w}_1, \bar{\sigma}_1 \cdot I_5)$.*

   The latent variables Z_i's are not (fully) identifiable in this case. The reason is that one set of ML *parameters is $\theta^* = (w^*, \sigma^*) = ([\mathbf{0}, \bar{w}_1], \bar{\sigma}_1)$, i.e. setting one latent dimension as zero and the other equal to the true data generating direction. Under this θ^*, the likelihood function is constant in the first dimension of the latent variable, i.e. Z_{i1}; see Theorem 6. The posterior of Z_{i1} thus collapses, matching the prior, while the posterior of Z_{i2} stays peaked (Theorem 6).*

2. *Samples from a two-dimensional* PPCA, *$X_i \sim \mathcal{N}(x_i; \mathcal{N}(0, I_2) \cdot \bar{w}_2, \bar{\sigma}_2 \cdot I_5)$.*

   The latent variables Z_i's are identifiable in this case. The likelihood function varies against both Z_{i1} and Z_{i2}; the posteriors of both Z_{i1} and Z_{i2} are peaked (Theorem 6).

   Next, consider a variant of PPCA with fixed noise σ^2. Here the only parameter to be optimized is w. When the noise σ^2 is set to a large value, the latent variable Z_i may become nearly non-identifiable. The reason is that the likelihood function $p(x_i \mid z_i)$ becomes slower-varying as σ^2 increases. For example, Section 2.2 shows that the likelihood surface becomes flatter as σ^2 increases. Accordingly, Section 2.2 shows that the posterior becomes closer to the prior as σ^2 increases. When $\sigma = 1.5$, the posterior closely matches the prior (i.e. collapses).

Likelihood (1D PPCA) Posterior (1D PPCA) Likelihood (2D PPCA) Posterior (2D PPCA)

Figure 3: Fitting PPCAS with more latent dimensions than enough leads to non-identifiable local latent variables and collapsed posteriors. (a)-(b) Fit a two-dimensional PPCA to data drawn from a one-dimensional PPCA. The likelihood surface is constant in one dimension of the latent variable, and the corresponding posterior collapses. (c)-(d) Fit a two-dimensional PPCA to data from a two-dimensional PPCA does not suffer from posterior collapse; its likelihood surface varies in all dimensions.

Appendix D. Theoretical and practical details of identifiable VAEs

The identifiable VAE with monotone transport maps emulates many existing VAEs. Letting Z_i be categorical (one-hot) vectors, the distribution $\mathbf{expfam}(z_i^\top \beta_\theta\,;\gamma_\theta)$ is an exponential family mixture. The identifiable VAE then maps this mixture model through a flexible function g_θ. When Z_i is real-valued, it mimics classical VAEs by mapping an exponential family PCA through flexible functions.

How does the VAE model with monotone transport maps guarantee identifiability? It is identifiable for two reasons: the positive and strictly increasing requirement on β_θ and the monotone transport map requirement on g_θ. The first requirement on β_θ guarantees that each dimension of β must be distinct and positive. It prevents Z_i from being non-identifiable due to indistinguishable latent clusters or zero latent dimensions (cf. Examples 1 and 2). Therefore, the distribution $\mathbf{expfam}(z_i^\top \beta_\theta\,;\gamma_\theta)$ is guaranteed to be different under different values of Z_i.

The second requirement on g_θ requires that it must be monotone. Loosely, monotone functions are functions such that their mapping preserves (or reverse) the order of the data points. To resolve non-identifiability, we build off the key result of McCann et al. (1995), which shows that monotone transport maps between probability distributions must be unique under weak conditions. It implies that if the g_θ function can successfully transport distributions p_1 to p_2 (both must be non-degenerate), then g_θ must be the unique monotone transport map. Returning to the VAEs, the monotonicity of g_θ preserves the distinctiveness of $\mathbf{expfam}(z_i^\top \beta_\theta\,;\gamma_\theta)$ under different values of z_i. Together with h being one-to-one, it implies that $h \circ g_\theta(\mathbf{expfam}(z_i^\top \beta_\theta\,;\gamma_\theta))$ must take different values under different values of z_i, Therefore, the likelihood $p(x_i \,|\, z_i\,;\theta)$ must be non-constant, and the latent variables Z_i's must be identifiable and, by Theorem 2, do not suffer from posterior collapse.

A reader might ask: Do these requirements constrain the flexibility of the identifiable VAEs? It turns out that the requirements on β_θ and g_θ only resolves non-identifiability but do not limit the representation power of the identifiable VAE. Any full-rank β_θ can be row-permuted and re-centered to satisfy $0 < \beta_\theta[0,0] < \beta_\theta[1,0] < \ldots < \beta_\theta[K,0]$. Moreover, constraining g_θ to be monotone does not limit the capacity of the generative model because monotone transport maps (almost) always exists (Theorem 6 of (McCann et al., 1995)).

The following theorem summarizes these results.

Theorem 7 *The latent variable Z_i is identifiable in the identifiable* VAEs *via monotone transport maps. Moreover, it has the same capacity as the* VAE *absent all parameter constraints.*

Next, we attend to practical aspects of identifiable VAEs. The first aspect is how to parametrize the monotone transport map g_θ? We parametrize the monotone g_θ as the gradient of input convex neural networks (ICNNs) (Amos et al., 2017; Makkuva et al., 2019), which can approximate any convex function on a compact domain in sup norm (Theorem 1 of Chen et al. (2018).) The rationale is that the gradient of any convex functions must be a monotone transport map (McCann et al., 1995). More concretely, we require that the function $g_\theta(x) : \mathscr{X}^m \to \mathscr{X}^m$ must be an L-layer feed forward neural network such that for $l = 1, \ldots, L-1$,

$$z_{l+1} = \sigma_l(W_l z_l + A_l x + b_l), \qquad g_\theta(x) = \frac{\partial z_L}{\partial x}, \tag{13}$$

where the last layer z_L must be a scalar, $\{W_l\}$ are non-negative with $W_0 = \mathbf{0}$, and $\{\sigma_l\}$ are convex and non-decreasing. A common choice of σ_0 is the square of leaky RELU, $\sigma_0(x) = (\max(\alpha \cdot x, x))^2$ with $\alpha = 0.2$; the remaining σ_l's are set be a leaky RELU, $\sigma_l(x) = \max(\alpha \cdot x, x)$.

The second practical aspect is how to perform inference on the identifiable VAEs. As the identifiable VAEs differ from the canonical ones only in its parameter constraints, the canonical amortized inference algorithms of VAEs directly apply here. Thus, to maximize log marginal likelihood, we posit an auxiliary variable $u_i = \text{expfam}(z_i^\top \beta_\theta ; \gamma_\theta)$ and maximize the ELBO of the log marginal likelihood,

$$\max_{q \in \mathscr{Q}} L(q, \theta) = \max_{q \in \mathscr{Q}} \mathbb{E}_{q(u_i, z_i)}[\log p(x_i, u_i, z_i ; \theta) - \log q(u_i, z_i)], \tag{14}$$

The variational family is $\mathscr{Q} = \{q(u_i, z_i) : q(u_i, z_i) = q_{\phi_u}(u_i \,|\, z_i ; x_i) \cdot q_{\phi_z}(z_i ; x_i)\}$, where q_{ϕ_u} and q_{ϕ_z} are neural network mappings. These steps follow closely from the amortized inference in VAEs.

	ELBO NON-ID	**ELBO** ID	% **AU** NON-ID	% **AU** ID
Pinwheel (Johnson et al., 2016)	-6	-6	0.2	**1.0**
MNIST (LeCun et al., 2010)	-108	**-96**	0.1	**1.0**
Fashion MNIST (Xiao et al., 2017)	-259	**-243**	0.1	**1.0**
Omniglot (Lake et al., 2015)	-862	**-824**	0.02	**1.0**

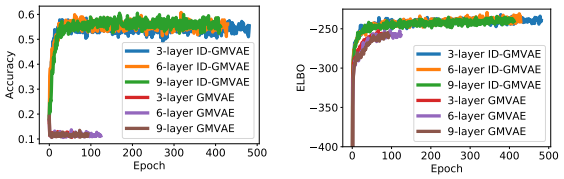
Table 1: The identifiable GMVAEs do not suffer from posterior collapse and achieves better model fit than its classical counterpart in a 9-layer generative model. % AU indicates the proportion of the units being active. ELBO indicates the fit of the model. (Higher is better.)

Appendix E. Empirical studies

We study the identifiable GMVAEs and the classical non-identifiable GMVAEs on four datasets: pinwheel (Johnson et al., 2016), MNIST (LeCun et al., 2010), Fashion MNIST (Xiao et al., 2017), and Omniglot (Lake et al., 2015). We find that the identifiable VAEs do not suffer from posterior collapse as the generative model becomes more flexible. They also achieve higher ELBOs than the GMVAEs, suggesting better fits to the data. (Throughout the empirical studies, we use two-layer [512,512] RealNVPs (Dinh et al., 2016) as approximating families to maximally tease out the approximation effect of variational inference.)

Evaluation metrics. To evaluate posterior collapse, we follow (Burda et al., 2015) to compute the number of active units (AU), $\text{AU} = \sum_{d=1}^{D} \mathbb{1}\{\text{Cov}_{p(\boldsymbol{x})}(\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\boldsymbol{Z}_d]) \geq \epsilon\}$, where $\boldsymbol{Z}_d$ is the dth dimension of the latent variable $\boldsymbol{Z}$ and the threshold ϵ is chosen to be 0.01. We also evaluate the predictive accuracy of the categorical latents against ground truth labels to quantify their informativeness.

Results. Section 1 illustrate a fit of the GMVAE and the identifiable GMVAE to the pinwheel data (Johnson et al., 2016). The identifiable GMVAE produces categorical latents faithful to the clustering structure, but the posterior of the GMVAE latents collapse, attributing all data-points to the same latent cluster. Figure 4 examines the identifiable GMVAEs as we increase the flexibility of the generative model. Appendix E shows that the categorical latents of the identifiable GMVAEs are substantially more predictive of the true labels than its classical counterparts. Moreover, its performance does not degrade as the generative model becomes more flexible. Appendix E shows that the identifiable GMVAEs consistently achieve higher ELBOs. Table 1 compares the identifiable GMVAE and the GMVAE in a 9-layer generative model. The identifiable GMVAE does not suffer from posterior collapse and achieves higher ELBOs.



Accuracy wrt true labels ELBO

Figure 4: Fashion MNIST: The identifiable GMVAEs produces posteriors that are substantially more informative than GMVAE. It also achieves higher ELBO and its performance does not degrade as the generative model becomes more flexible.

Appendix F. Experiment details

Experimental details. For all experiments, we use the Adam optimizer with learning rate 0.0001. All hidden layers of the neural networks have 512 units. For continuous latent variables We use two-layer RealNVP ((Dinh et al., 2016)) as an approximating family to tease out the effect of variational inference.

In Table 1, we choose the number of units as the number of categories in ground truth labels. For continuous latents, it is chosen as 200.

References

Alemi, A. A., Poole, B., et al. (2017). Fixing a broken elbo. *arXiv preprint arXiv:1711.00464*.

Amos, B., Xu, L., & Kolter, J. Z. (2017). Input convex neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (pp. 146–155).: JMLR. org.

Asperti, A. (2019). Variational autoencoders and the variable collapse phenomenon. *Sensors & Transducers*, 234(6), 1–8.

Betancourt, M. (2017). Identifying bayesian mixture models.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518), 859–877.

Bowman, S., Vilnis, L., et al. (2016). Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning* (pp. 10–21).

Burda, Y., Grosse, R., & Salakhutdinov, R. (2015). Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*.

Chen, X., Kingma, D. P., et al. (2016). Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*.

Chen, Y., Shi, Y., & Zhang, B. (2018). Optimal control via neural networks: A convex approach. *arXiv preprint arXiv:1805.11835*.

Dai, B., Wang, Z., & Wipf, D. (2019). The usual suspects? reassessing blame for vae posterior collapse. *arXiv preprint arXiv:1912.10702*.

Diederik, P. K., Welling, M., et al. (2014). Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, volume 1.

Dieng, A. B., Kim, Y., Rush, A. M., & Blei, D. M. (2018). Avoiding latent variable collapse with generative skip models. *arXiv preprint arXiv:1807.04863*.

Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2016). Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.

Fu, H., Li, C., et al. (2019). Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*.

Havrylov, S. & Titov, I. (2020). Preventing posterior collapse with levenshtein variational autoencoder. *arXiv preprint arXiv:2004.14758*.

He, J., Spokoyny, D., Neubig, G., & Berg-Kirkpatrick, T. (2019). Lagging inference networks and posterior collapse in variational autoencoders. *arXiv preprint arXiv:1901.05534*.

Hoffman, M. D. & Johnson, M. J. (2016). Elbo surgery: yet another way to carve up the variational evidence lower bound.

Johnson, M. J., Duvenaud, D. K., Wiltschko, A., Adams, R. P., & Datta, S. R. (2016). Composing graphical models with neural networks for structured representations and fast inference. In *Advances in neural information processing systems* (pp. 2946–2954).

Khemakhem, I., Kingma, D. P., & Hyvärinen, A. (2019). Variational autoencoders and nonlinear ica: A unifying framework. *arXiv preprint arXiv:1907.04809*.

Kim, Y., Wiseman, S., Miller, A., Sontag, D., & Rush, A. (2018). Semi-amortized variational autoencoders. In *International Conference on Machine Learning* (pp. 2678–2687).

Kingma, D. P., Salimans, T., et al. (2016). Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems* (pp. 4743–4751).

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266), 1332–1338.

LeCun, Y., Cortes, C., & 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., & Yang, Y. (2019). A surprisingly effective fix for deep latent variable modeling of text. *arXiv preprint arXiv:1909.00868*.

Lucas, J., Tucker, G., Grosse, R. B., & Norouzi, M. (2019). Don't blame the elbo! a linear vae perspective on posterior collapse. In *Advances in Neural Information Processing Systems* (pp. 9403–9413).

Makkuva, A. V., Taghvaei, A., Oh, S., & Lee, J. D. (2019). Optimal transport mapping via input convex neural networks. *arXiv preprint arXiv:1908.10962*.

McCann, R. J. et al. (1995). Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80(2), 309–324.

Murphy, S. A. & Van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95(450), 449–465.

Poirier, D. J. (1998). Revising beliefs in nonidentified models. *Econometric Theory*, 14(4), 483–509.

Raue, A., Kreutz, C., et al. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15), 1923–1929.

Raue, A., Kreutz, C., Theis, F. J., & Timmer, J. (2013). Joining forces of bayesian and frequentist methodology: a study for inference in the presence of non-identifiability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984), 20110544.

Razavi, A., Oord, A. v. d., Poole, B., & Vinyals, O. (2019). Preventing posterior collapse with delta-vaes. *arXiv preprint arXiv:1901.03416*.

San Martın, E. & González, J. (2010). Bayesian identifiability: Contributions to an inconclusive debate. *Chilean Journal of Statistics*, 1(2), 69–91.

Seybold, B., Fertig, E., Alemi, A., & Fischer, I. (2019). Dueling decoders: Regularizing variational autoencoder latent spaces. *arXiv preprint arXiv:1905.07478*.

Shu, R. (2016). Gaussian mixture VAE: Lessons in variational inference, generative models, and deep nets.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., & Winther, O. (2016). How to train deep variational autoencoders and probabilistic ladder networks. In *33rd International Conference on Machine Learning (ICML 2016)*.

Tomczak, J. M. & Welling, M. (2017). Vae with a vampprior. *arXiv preprint arXiv:1705.07120*.

Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747.

Xie, Y. & Carlin, B. P. (2006). Measures of bayesian learning and identifiability in hierarchical models. *Journal of Statistical Planning and Inference*, 136(10), 3458–3477.

Yeung, S., Kannan, A., Dauphin, Y., & Fei-Fei, L. (2017). Tackling over-pruning in variational autoencoders. *arXiv preprint arXiv:1706.03643*.

Zhao, T., Lee, K., & Eskenazi, M. (2018). Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1098–1107).

Zhao, Y., Yu, P., Mahapatra, S., Su, Q., & Chen, C. (2020). Discretized bottleneck in vae: Posterior-collapse-free sequence-to-sequence learning. *arXiv preprint arXiv:2004.10603*.