A Proof of Mode Collapse

The probability density function of the c-th component of mixture priors could be defined by its natural parameters η_c according to the definition of exponential family:

$$p_\eta(z|c) = \exp(<\eta_c, \phi(z)> - A(\eta_c)),$$

in which $\phi = [\phi_1, \phi_2, ..., \phi_D]$ is a vector of functions of z called sufficient statistics. For example, the sufficient statistics of normal distribution is $[z, z^2]$. For each sufficient statistic ϕ_d, there is a corresponding natural parameter η_d. $<\eta, \phi(z)>$ is the inner-product of vector η and ϕ. $A(\eta_c)$ is the log-partition function, which is a function of natural parameters η_c.

Considering the $\mathcal{R}_z$ term in ELBO:

$$\begin{aligned}&\mathbb{E}_{q_\phi(c|x)} \mathbb{E}_{q_\phi(z|x)} \log \frac{p_\eta(z|c)}{q_\phi(z|x)} = \\ &\qquad \mathbb{E}_{q_\phi(z|x)}[\mathbb{E}_{q_\phi(c|x)} \log p_\eta(z|c)] - \mathbb{E}_{q_\phi(z|x)} \log q_\phi(z|x).\end{aligned} \tag{1}$$

In terms of the probability density function of exponential family, $\mathbb{E}_{q_\phi(c|x)} \log p_\eta(z|c)$ could be re-written as

$$\begin{aligned}&\mathbb{E}_{q_\phi(c|x)} \log \exp(<\eta_c, \phi(x)> - A(\eta_c)) \\ =&< \mathbb{E}_{q_\phi(c|x)} \eta_c, \phi(x)> - \mathbb{E}_{q_\phi(c|x)} A(\eta_c)\end{aligned} \tag{2}$$

Define the weighted expectation of natural parameters $\mathbb{E}_{q_\phi(c|x)} \eta_c$ as $\overline{\eta}$. Then,

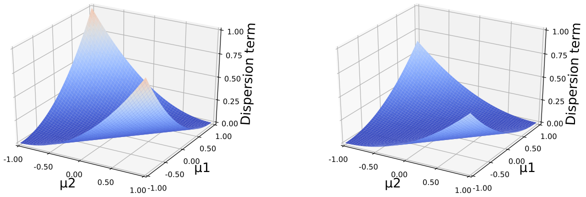
$$\begin{aligned}&\mathbb{E}_{q_\phi(c|x)} \log p_\eta(z|c) \\ =&< \overline{\eta}, \phi(x)> - A(\overline{\eta}) - (\mathbb{E}_{q_\phi(c|x)} A(\eta_c) - A(\overline{\eta})) \\ =& \log \hat{p}_{\overline{\eta}}(z|x) - (\mathbb{E}_{q_\phi(c|x)} A(\eta_c) - A(\mathbb{E}_{q_\phi(c|x)} \eta_c)),\end{aligned} \tag{3}$$

in which $\hat{p}_{\overline{\eta}}(z|x) = \exp(<\overline{\eta}, \phi(x)> - A(\overline{\eta}))$. $\hat{p}_{\overline{\eta}}(z|x)$ is a distribution with the same sufficient statistics as priors but different parameters $\overline{\eta}$. Noted that the feasible domain of exponential family is a convex set, thus $\overline{\eta}$ is a feasible natural parameter as well.

As a result, the $\mathcal{R}_z$ could be re-written as an "average" KL divergence term of z and a dispersion term with respect to priors parameters and $q_\phi(c|x)$,

$$\begin{aligned}&\mathbb{E}_{q_\phi(z|x)} \sum_c q_\phi(c|x) \log \frac{p_\eta(z|c)}{q_\phi(z|x)} = \\ &\qquad \underbrace{- \text{KL}(q_\phi(z|x)||\hat{p}_\eta(z|x))}_{\text{Average } \mathcal{R}_z} \\ &\qquad - \underbrace{(\mathbb{E}_{q_\phi(c|x)} A(\eta_c) - A(\mathbb{E}_{q_\phi(c|x)} \eta_c))}_{\text{Dispersion term } \mathcal{L}_d}.\end{aligned} \tag{4}$$

We further analyze the relationship between dispersion term $\mathcal{L}_d$ and the degree of dispersion of prior parameters.



(a) $q(c = 1|x) = 0.5$ (b) $q(c = 1|x) = 0.2$

Figure 1: Dispersion term $\mathcal{L}_d$ of GMVAE with respect to the parameters of two prior components under different posterior $q(c|x)$.

Firstly, we define the a weighted variance of natural parameters as the trace of variance matrix of natural parameters:

$$\begin{aligned}\text{Var}_{q(c|x)}\eta_c &= \text{Tr}\Big(\mathbb{E}_{q(c|x)}[(\eta_c - \mathbb{E}_{q(c|x)} \eta_c)^T(\eta_c - \mathbb{E}_{q(c|x)} \eta_c)]\Big), \\ &= \sum_d \mathbb{E}_{q(c|x)}(\eta_c^d)^2 - (\mathbb{E}_{q(c|x)} \eta_c^d)^2.\end{aligned} \tag{5}$$

in which η_c^d is the d-th dimension of η_c. The weighted variance term reflects the discrepancy between parameters of priors under distribution $q(c|x)$. If we calculate the gradient of this variance term with respect to η_c:

$$\nabla_{\eta_c} \text{Var}_{q(c|x)}\eta_c = 2q(c|x)(\eta_c - (E_{q(c|x)}\eta_c)). \tag{6}$$

In the meantime, we calculate the gradient of dispersion term $\mathcal{L}_d$ with respect to η_c:

$$\nabla_{\eta_c} \mathcal{L}_d = q(c|x)(\nabla A|_{\eta_c} - \nabla A|_{E_{q(c|x)}\eta_c}) \tag{7}$$

Noted that the log-partition function A is convex. It has $(\nabla A|x - \nabla A|y)(x - y) \geq 0$. As a result, we have

$$(\nabla_{\eta_c} \mathcal{L}_d)^T \cdot (\nabla_{\eta_c} \text{Var}_{q(c|x)}\eta_c) \geq 0. \tag{8}$$

When A is strictly convex, the dispersion term is larger than zero and inner product of gradients in Eq. 8 is larger than zero as well. If the gradient descent methods are used for optimizing the ELBO, it will implicitly minimizing the weighted variance of prior parameters when mining the $\mathcal{L}_d$.

It should be noticed that the variance of prior parameters are weighted by posterior distribution $q_\phi(c|x)$. The dispersion terms of two-Gaussian mixture priors under different posterior distribution are shown in Fig 1. Under an unbalanced posterior distribution $q_\phi(c|x)$, the dispersion term is smaller in general, although it still takes the minimum value $(= 0)$ when parameters of all components are equal. If x corresponds to a certain c, which means the $q_\phi(c|x)$ has a one-hot probability mass function, both the dispersion term and variance term come to zero. Therefore, the mutual information is helpful to alleviate the mode-collapse as well because it implicitly enhances the correspondence between x and c.

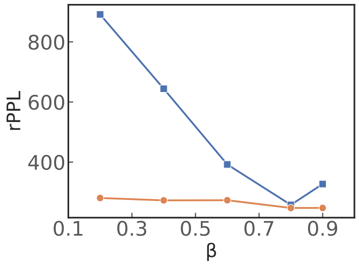


Figure 2: Reverse perplexity (rPPL) of DGMVAE with different β, on the validation set of PTB. Results of DGMVAE and DGMVAE + $\mathcal{L}_{\text{mi}}$ are displayed by the blue lines and orange lines, respectively.

B Hyper-Parameters

Bidirectional GRU encoder and one-layer GRU decoder with 512 hidden units are adopted. The size of word embedding is set as 512. Sentence longer than 40 will be cut off. Vocabulary size is set to 10,000. Latent variable z is concatenated with input word embedding at each decoding step. Adam optimizer is adopted with learning rate of 0.001. Batch size is set to 30. During training of all VAE models, we sample the latent variable z from posterior $q_\phi(z|x)$ for 20 times, and average their the reconstruction loss. All results were obtained by repeating the experiment three times and taking an average.

We also illustrate how β will affect the model performance in Fig. 2; and we generally find that larger β can get better results.

C More Cases

More examples of actions discovered by DGMVAE is shown in Tab. 1. More examples on responses generated by DGMVAE are shown in Tab. 2. In Tab. 2, an example without context is given to show the ability to begin a dialog in different topics (weather, navigation and scheduling). For DGMVAE, we can sample different continuous latent variables from one component. As shown in Tab. 3, diverse responses with the same actions could be generated.

Action Name	Request-location
Utterances	Which location do you want the weather for? Which location should I look up information about? Which city are you asking about?
Action Name	Inform-time/appointment
Utterances	Your next dinner event is with your father on Friday. Your father will be attending your yoga activity on the 2nd with you. Your doctor's appointment is Monday at 1 pm.
Action Name	General-thanks
Utterances	Thanks. Thanks a lot. Perfect. Thanks.

Table 1: Example actions discovered by DGMVAE on SMD. The action name is annotated by experts.

Context	*User:* What's the temperature going to be this week? *Sys:* What city are you wanting to know the temperature for?
Predict	(1-1-0, inform-address) Cleveland. (4-0-2, request-weather) Will it rain in Redwood City today?
Context	None
Predict	(1-4-0, request-route/address) find me a nearby coffee shop (1-0-4, request-weather) what's the weather going to be like today and tomorrow (1-2-4, command-reminder) remind me about meeting later

Table 2: Dialog cases on SMD, which are generated by sampling different c from policy network. The label of sampled c are listed in parentheses with the annotated action name. When context is None, it means to predict the beginning a dialog.

Context	*User:* What is the highest temperature in Brentwood over the next two days?
Action Name	(2-3-0) inform-weather
Predict	It is currently foggy in Brentwood on Tuesday. It will be between 70 - 40F and turn - 40F on Saturday.
Context	*User:* I need gas.
Action Name	(2-4-2) inform-route/address
Predict	There is a Chevron 3 miles from you. There is a Safeway.
Context	*User:* schedule meeting
Action Name	(3-0-4) request-time
Predict	What day and time should I set your meeting for? What time should I set the alarm?

Table 3: Dialog cases on SMD, which are generated by sampling different z from the same actions (i.e., the c).