LOCALGAN: MODELING LOCAL DISTRIBUTIONS FOR ADVERSARIAL RESPONSE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper presents a new methodology for modeling the local semantic distribution of responses to a given query in the human-conversation corpus, and on this basis, explores a specified adversarial learning mechanism for training Neural Response Generation (NRG) models to build conversational agents. The proposed mechanism aims to address the training instability problem and improve the quality of generated results of Generative Adversarial Nets (GAN) in their utilizations in the response generation scenario. Our investigation begins with the thorough discussions upon the objective function brought by general GAN architectures to NRG models, and the training instability problem is proved to be ascribed to the special local distributions of conversational corpora. Consequently, an energy function is employed to estimate the status of a local area restricted by the query and its responses in the semantic space, and the mathematical approximation of this energy-based distribution is finally found. Building on this foundation, a local distribution oriented objective is proposed and combined with the original objective, working as a hybrid loss for the adversarial training of response generation models, named as LocalGAN. Our experimental results demonstrate that the reasonable local distribution modeling of the query-response corpus is of great importance to adversarial NRG, and our proposed LocalGAN is promising for improving both the training stability and the quality of generated results.

1 INTRODUCTION

End-to-End generative conversational agents (a.k.a., generative Chat-bots) are believed to be practicable on the basis of the Sequence-to-Sequence (Seq2Seq) architecture (Sutskever et al., 2014) trained with large amounts of human-generated conversation sessions (Shang et al., 2015a; Sordoni et al., 2015), and this task is named as Neural Response Generation (NRG). Similar to the Neural Machine Translation (NMT) approaches (Bahdanau et al., 2014; Wu et al., 2016), the deep Seq2Seq models are expected to directly generate appropriate and meaningful responses according to the input query. Compared to the success of NMT systems, the application progress of NRG models is not satisfying at present due to the "safe response" problem (Li et al., 2016). That is, most of the generated responses are boring and meaningless, which blocks the continuation of conversations. Indeed, eliminating "safe responses" is the essential task of NRG models. Thus, various methods have been considered to address this problem (Li et al., 2016; Xu et al., 2017; Li et al., 2017; Xing et al., 2017; Pandey et al., 2018; Zhang et al., 2018a; Du et al., 2018).

More recently, Generative Adversarial Nets (GAN) (Goodfellow et al., 2014) have been introduced to eliminate "safe responses" (Li et al., 2017; Xu et al., 2017). Basically, this methodology is reasonable since the GAN framework involves an adversarial discriminator that helps NRG models leap out of the shortsighted state of minimizing the empirical risk on word distribution, by providing feedback on real samples from the model generated ones. Despite of the improvement on the diversity, the adversarial training process of GAN based response generation models is generally unstable and sensitive to the training strategy (Yu et al., 2017).

The unstable convergence problem is largely ascribed to the complicated data distribution in practical scenarios (Arora et al., 2017; Arora & Zhang, 2017). For the response generation oriented GAN models, in particular, the data distribution appears to be much more complicated. Fundamentally, an essential characteristic of conversation data is that, for each given query, there always exists a group of semantically-diverse responses, rather than the semantically-unified ones. Furthermore, the response groups of two different queries tend to keep great divergences in the semantic space. In this scenario, the discriminator needs to consider the distributions of the generated result in the semantic space, rather than simply examining whether one single sample comes from the generator or the original dataset, so as to make the generator sense the distribution of conversation dataset in the adversarial training procedure.

This paper aims at presenting a specific adversarial training schema for neural response generation. Beginning with the investigation on the reason for the unsatisfying performance of GANs on the NRG task, we find the upper-bound of the current GAN learning strategies taking query-response pairs as the independent training samples. On this basis, we claim that the training schema, including the actual adversarial strategy and the overall loss function, should be re-defined to agree with the distribution of NRG training samples in the semantic space, rather than roughly adopting the GAN framework designed for generating images.

Consequently, we describe the distributional state of the given query and the corresponding responses with the free energy defined on the basis of Deep Boltzmann Machines (DBM) (Salakhutdinov & Hinton, 2009). In this way, we can quantify the formation process of generating the response set with the topic restriction of the given query. From the perspective of free energy, this paper proposes a new cost function to measure the expansion degree of the responses in the local area of the real-valued semantic space. Cooperating with the traditional implicit density discriminating loss of GAN, the proposed cost actually provides an explicit density approximation for the local distribution of each response cluster. Thus, the adversarial learning procedure can be expected to be more stable with better response generation results obtained[1].

2 The limitation of General GAN in the NRG Scenario

According to (Goodfellow et al., 2014), the standard GAN framework contains a generator G and a discriminator D, which are trained by an iterative adversarial learning procedure based on the following objective function:

$$J^{(D)} = \mathbb{E}_{x \sim p_d}\left[\log D(x)\right] + \mathbb{E}_{z \sim p_z}\left[\log(1 - D(G(z)))\right] \tag{1}$$

$$J^{(G)} = \mathbb{E}_{z \sim p_z}\left[\log D(G(z))\right] \tag{2}$$

where p_z denotes the prior on input noise variables, and p_d is the true data distribution.

It should be noted that the GAN tries to learn the manifold of a given dataset (Khayatkhoei et al., 2018; Kumar et al., 2017), and the discriminator D actually provides a metric for judging whether the results generated according to z fits the expected manifold or not. In the NRG scenario, a naive Seq2Seq model without the guidance signal from D can not capture the data manifold of the real query-response corpus, which is one of the major facts the safe-response problem can be ascribed to. Assuming that there exists an oracle discriminator with the ability of distinguishing the generated fake samples from the ground-truth ones, by mapping each query-response pair (q, r)[2] to a confidence score s, we can ideally define the capability of a GAN-based NRG framework as follows:

$$\mathcal{C} = -KL((q, r, s), (q, r, \tilde{s})) \tag{3}$$

where $\tilde{s}$ indicates the confidence score given by any practically existing discriminator of GAN. Consequently, to improve the capability of GAN-NRG, it is wise to construct more powerful discriminators for more reasonable $J^{(G)}$.

Now let's pay attention to the actual change of NRG models brought by GAN. In the generative conversation agent scenario, $G(z)$ is corresponding to a generated response $\tilde{r}$ to a given query q. Thus, the objective of the generator in the GAN based NRG model can be SIMPLY formulated as:

$$J^{(G)} = \mathbb{E}_{(q,\tilde{r}) \sim p_g}\left[\log D(q, \tilde{r})\right] \tag{4}$$

[1] The code of our proposed model LocalGAN can be found in `https://github.com/Kramgasse49/local_gan_4generation`

[2] Here q and r represent the vectorized query and its response. We take the simple embedding-averaging based method to transform texts into vectors. Besides, due to our adopted text vectorizing method, the pre-training phase of the Deep Boltzmann Machine takes some specified trick detailed in Appendix A.

The training of the generator is actually the procedure to maximize $J^{(G)}$ toward $\mathbb{E}_{(q,r)\sim p_d}\left[\log D(q,r)\right]$, so as to generate realistic responses according to given queries. Thus, in the context of adversarial learning, $J^{(G)}$ should satisfy the following inequation:

$$J^{(G)} \leqslant \mathbb{E}_{(q,r)\sim p_d}\left[\log D(q,r)\right] \tag{5}$$

However, it is well known that a conversational dataset should not be simply taken as a collection $\{(q,r)\}$ composed of independent query-response pairs. Instead, to each given query q, there exists a finite set of corresponding responses $R_q = \{r_i\}$. In this case, it is of great necessity to consider the whole training dataset as a collection of R_q, which takes the form of a number of clusters with their own local distributions in the semantic space. And we can rewrite the joint distribution p_d in (5) as $p(q)p(r|q)$ and assume every corresponding response to a query follows equal-probability distribution, which means that $p(r|q) = \frac{1}{|R_q|}$. Thus, in the real NRG scenario, on the basis of Equation 4 and 5, $J^{(G)}$ follows the inequation as below:

$$\begin{aligned} J^{(G)} &\leqslant \sum_{q,R_q} \mathbb{E}_{(q,r)\sim q,R_q}\left[\log D(q,r)\right] \\ &= \sum_{q}\sum_{r\in R_q} p(q)\frac{1}{|R_q|}\left[\log D(q,r)\right] \leqslant \sum_{q} p(q)\log\left[\frac{1}{|R_q|}\sum_{r\in R_q} D(q,r)\right] \end{aligned} \tag{6}$$

where $p(q)$ denotes the probability of the query q, R_q is defined above and $\tilde{R}_q$ represents the set of generated responses of query q.

According to Equation 6, the upper bound of $J^{(G)}$ is obtained, in which the $\log\left[\frac{1}{|R_q|}\sum_{r\in R_q} D(q,r)\right]$ part is the essence. Apparently, the expression $\frac{1}{|R_q|}\sum_{r\in R_q} D(q,r)$ indicates the mean value of the confidence scores given by the discriminator to each member of the response set R_q to a given query q. Moreover, it should be noted that current studies tend to utilize semantic relevance oriented models to build the discriminators of GAN (Xu et al., 2017; Li et al., 2017). Consequently, $D(q,r)$ can be actually considered as the spatial relationship of q and r in the semantic space. Thus, $\frac{1}{|R_q|}\sum_{r\in R_q} D(q,r)$ stands for the spatial center of all the responses within R_q. That is, the optimization process of adversarial learning upon conversational datasets will make the generated responses approach to the center of each local distribution of R_q corresponding to each given dependent query.

The practical value of this change lies in that, intuitively, the GAN architecture forces the generation to pay attention to the local distributions of the individual response clusters, rather than taking the (q,r)-pairs as an entirety. According to the thorough studies on the safe responses of NRG models (Li et al., 2016; Xu et al., 2017; Zhang et al., 2018a; Pandey et al., 2018), it can be inferred that the general Seq2Seq will fall into the divergence state of generating the patterns with the maximum probabilities taking account of the entire dataset, ignoring the individual-difference of each query. By introducing the implicit loss focusing on the response clusters, GAN makes the divergence of the generator much closer to the 'local patterns' rather than the general patterns, and thus the higher diversity can be expected and observed (Xu et al., 2017; Li et al., 2017).

The problem turns to: Is the upper bound in Equation 6 powerful enough? Apparently, there exists an obvious gap between the 'local patterns' and the vivid and interesting generated responses. The upper bound only focuses on the mean of each response set, but the local distribution (or the actual 'shape') of each cluster has not been taken into account. This situation does not change when the cost function of adversarial training is defined by Wasserstein GANs (WGAN) (Arjovsky et al., 2017) with 1-Lipschitz function f:

$$J_W^{(G)} = \mathbb{E}_{(q,\tilde{r})\sim p_g}\left[f(q,\tilde{r})\right] \tag{7}$$

Intuitively, it is of paramount importance to estimate both the 'location' and the 'shape' of the response set R_q in the semantic space (indeed, $\frac{1}{|R_q|}\sum_{q,r} D(q,r)$ is only relative with 'location'), so as to determine the optimization objective of adversarial training. Consequently, we have two critical problems to be discussed and addressed in the following sections:

- How to describe the state of the response set R_q with a given query q in the semantic space?
- Taking account of the reasonable state modeling of (q, R_q), what is the loss function for adversarial training to generate responses?

3 Modeling the State of the Local Distribution for Responses

As mentioned above, the semantic one-to-many relationship between queries and responses makes it necessary to model the local distribution of the response cluster R_q corresponding to each query in the space, and it is paramount to turn to the fitting of each local distribution in the adversarial learning procedure, rather than considering each (q, r)-pair as an independent sample. Basically, this issue equals to the task of reasonably modeling the state of (q, R_q) in the semantic space, by considering each (q, R_q) as a systematic entirety and assigning the state of the entirety with probabilistic distribution. The additional major challenge of this task is, indeed, we have to infer the state of a local area in the semantic space from a limited samples, since it is impossible to sample all the possible responses to a query from the given corpus, regardless of the corpus size.

3.1 Representing Local Distributions with Query-Response oriented Free Energy

In this part, we typically take an energy based statistical model, the Average Free Energy (Hinton & Zemel, 1994; Friston et al., 2006; Ngiam et al., 2011; Friston et al., 2012), to describe the state of the local distribution of (q, R_q) in the semantic space, for the reasons that: a) energy based models are considered as a promising avenue towards learning explicit generative models (LeCun et al., 2006; Le Roux & Bengio, 2008), by representing data distributions without any prior assumptions; and b) energy based models can be trained in the unsupervised way, and the energy functions of such models have the potential to estimate the state of generative models (Zhao et al., 2016).

At first, the free energy of a given query-response pair (q, r) can be defined as following:

$$F(q, r) = -\log\left[\sum_H \exp(-E(q, r, H))\right] \tag{8}$$

where $E(q, r, H)$ stands for the energy function defined according to the relationship of the query q and its response r via the hidden variable H.

We employ the Deep Boltzmann Machine (DBM) (Salakhutdinov & Larochelle, 2010; Smolensky, 1986; Hinton & Salakhutdinov, 2006) to implement $E(q, r, H)$, as illustrated by Figure 1. The reason for this choice lies in that, from the view of conversational agents, the meaningful query and the corresponding response are generally considered to maintain strong semantic relevance. Thus, the query and response can be mutually transformed into each other, which is supported by the considerable amount of studies on response generation (Shang et al., 2015b; Shao et al., 2017; Zhang et al., 2018a; Baheti et al., 2018) and question generation (Du et al., 2017; Zhao et al., 2018; Sun et al., 2018). Without loss of generality, the pairwise semantic relationship of the query q and the corresponding response r can be modeled by a two-layer DBM. The bottom layer is actually an abstract version of Seq2Seq models, in which a response r can be generated based on a hidden variable h_q, and h_q depends on the given query q theoretically. In the top layer, h_q conditionally depends on a hyper hidden variable h. Figure 1 illustrates the Deep Boltzmann Machine for modeling the semantic relationship of a query and its responses.

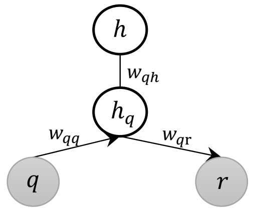

Figure 1: The DBM for modeling the semantic relationship of Query-Response pairs.

Following (Salakhutdinov & Hinton, 2009), on the basis of the DBM in Figure 1, the energy of the state $\{(q, r), H\}$ is defined as:

$$E(q, r, H) = E(q, r, h_q, h) = -h_q^T W_{qr} r - h_q^T W_{qq} q - h^T W_{qh} h_q \tag{9}$$

where q denotes the query, r stands for the response and $H = \{h_q, h\}$ represents the hidden units. W_{qr}, W_{qq} and W_{qh} stand for the weights on the corresponding connections of the query, response and the hidden variables respectively in the graph model shown by Figure 1.

Consequently, we can define the average free energy of the query q and its response set R_q as follows:

$$F(q, R_q) = \frac{1}{|R_q|} \sum_{r_i \in R_q} F(q, r_i) \tag{10}$$

For better conducting the following discussion, we further define the energy difference between response r_i and r_j as:

Definition 3.1 (**Scaled Energy Difference**).

$$\Delta_{q,r_i,r_j} = \frac{F(q, r_i) - F(q, r_j)}{F(q, R_q)} \tag{11}$$

Meanwhile, it is necessary to assign a spatial intuition to R_q in the semantic space by defining:

Definition 3.2 (**Response Cluster**). In the semantic space, the meaningful responses to the given query q lie in a restricted region (e.g., a hyper sphere), which can be named as the **Response Cluster**, in which $F(q, r)$ can be taken as the distance from a response r to the cluster center.

3.2 Estimation of $F(q, R_q)$

Basically, the DBM in Figure 1 provides the definition of the energy function $E(q, r, H)$ of the free energy given in Equation 9. Thus, the local distribution state of the responses R_q with the given query q can be mathematically described by $F(q, R_q)$ based on Equation 8 - 10. In practice, however, this procedure is not operable yet because the computation of $F(q, R_q)$ requires all the response in R_q, and it is impractical to perform the exhaustive enumeration over all the possible responses of the given query, regardless of the amount of the training query-response pairs. Consequently, it is highly necessary to approximate $F(q, R_q)$ under some reasonable assumptions.

According to the response cluster defined in 3.2, by taking the response r as a random variable, we can assume that the response follows multivariate normal distribution with mean $\mathbf{r_c}$ and covariance matrix Σ. It is obvious that the sample mean $\frac{1}{|R_q|} \sum_i r_i$ is an estimator for the population mean r_c. First, we propose the following two lemmas[3]:

Lemma 1. *If the possible responses are sufficient, $F(q, R_q)$ approximates $\mathrm{E}[F(q, \mathbf{r})]$.*

Lemma 2. *If the expected Euclidean distance between random variable $\mathbf{r}$ and constant variable $\mathbf{r_c}$ is small enough, $F(q, \mathbf{r_c})$ can be taken as the approximation of $\mathrm{E}[F(q, \mathbf{r})]$.*

Based on the lemmas we have:

Theorem 1. *Assume that $\mathbf{r_c}$ can be estimated with $\hat{\mathbf{r}}_\mathbf{c}$ properly and the expected Euclidean distance between $\mathbf{r}$ and $\mathbf{r_c}$ is small enough, we can take $F(q, \hat{\mathbf{r}}_\mathbf{c})$ as the approximation of $F(q, R_q)$.*

proof of theorem 1.

$$\begin{aligned} |F(q, R_q) - F(q, \hat{\mathbf{r}}_\mathbf{c})| &= |F(q, R_q) - \mathrm{E}[F(q, \mathbf{r})] + \mathrm{E}[F(q, \mathbf{r})] - F(q, \mathbf{r_c}) + F(q, \mathbf{r_c}) - F(q, \hat{\mathbf{r}}_\mathbf{c})| \\ &\leq |F(q, R_q) - \mathrm{E}[F(q, \mathbf{r})]| + |\mathrm{E}[F(q, \mathbf{r})] - F(q, \mathbf{r_c})| + |F(q, \mathbf{r_c}) - F(q, \hat{\mathbf{r}}_\mathbf{c})| \end{aligned} \tag{12}$$

First, the implicit assumption that $|R_q|$ is large enough can be met easily. Following lemma 1, we have $|F(q, R) - \mathrm{E}[F(q, \mathbf{r})]| \leq \epsilon$.

Second, the assumption that the expected Euclidean distance between $\mathbf{r}$ and $\mathbf{r_c}$ is small enough can yield $|\mathrm{E}[F(q, \mathbf{r})] - F(q, \mathbf{r_c})| \leq \epsilon$ according to lemma 2.

Finally, since the function $F(q, \mathbf{r})$ is the composition and combination of simple continuous functions, $F(q, \mathbf{r})$ is continuous as well. Thus, according to the assumption that r_c can be estimated with $\hat{r}_c$ properly, we have $|F(q, \mathbf{r_c}) - F(q, \hat{\mathbf{r}}_\mathbf{c})| \leq \epsilon$.

[3] The proof of Lemma 1 and Lemma 2 is given in Appendix B.

To sum up, under these assumptions,

$$|F(q, R_q) - F(q, \hat{\mathbf{r}}_{\mathbf{c}})| \leq 3\epsilon \tag{13}$$

Therefore, $F(q, \hat{\mathbf{r}}_{\mathbf{c}})$ is proved to be the approximation of $F(q, R_q)$.

□

4 THE HYBRID LOSS OF ADVERSARIAL RESPONSE GENERATION

As discussed in Section 2, the ability of the response generator in the general GAN architecture is limited to learning the dense distribution around $\frac{1}{|R_q|}\sum_{r \in R_q} D(q, r)$ (see Equation 6), which is composed of the most frequent patterns in the semantic space. By contrast, it is difficult for the general architecture to sense the remaining sparse space containing high-quality diverse responses. Therefore, reasonably describing the local distribution of the responses to a given query is highly necessary. According to the analysis in Section 3, the average free energy can be taken to model the state of the local area of the responses to a query, and such energy can be reasonably approximated via the DBM defined on the query-response pairs. On the basis of the previous sections, this section will finally propose the new hybrid loss function to force the generator to produce responses with better diversity through the more stable adversarial training process.

4.1 THE RADIAL DISTRIBUTION FUNCTION OF THE RESPONSE

The analysis in Section 3 have shown that the local distribution state of the responses R_q to the given query q can be modeled by the average free energy $F(q, R_q)$. On this basis, it is possible to propose the description of the spatial state of a single response r in the semantic space, and consequently, we can give a new adversarial loss indicating the cost of simulating the local distribution of R_q.

According to Definition 3.1 and 3.2, in each response cluster R_q, the distance from a response r to the cluster center r_c is actually equivalent to the scaled energy difference between them, that is,

$$\Delta_{q,r,r_c} = \frac{F(q, r) - F(q, r_c)}{F(q, R_q)} \tag{14}$$

Meanwhile, on the basis of Theorem 1, $F(q, R_q)$ can be approximated by $F(q, \hat{r}_c)$, and $\hat{r}_c$ is modeled from training data, and thus we have:

$$\Delta_{q,r,r_c} \approx \frac{F(q, r) - F(q, \hat{r}_c)}{F(q, \hat{r}_c)} = \alpha_{(q,r)} \tag{15}$$

Here we approximate Δ_{q,r,r_c} with $\alpha_{(q,r)}$, and formally call $\alpha_{(q,r)}$ as the Radial Distribution Function (RDF), indicating the relative cost ratio to $F(q, r_c)$ for obtaining r from a given q (also the distinctiveness of r, actually).

4.2 THE HYBRID OBJECTIVE FUNCTION

Based on the previous discussions, for the adversarial response generation methodology, the essence is to reasonably describe the state of the local distribution of the response cluster given by Definition 3.2, and further more, to take this important element into account in the final optimization progress.

Especially, in Subsection 4.1, we have defined the Radial Distribution Function in Equation 15 to quantify the distinctiveness of a response, the very basis of which is the description of the local state $F(q, R_q)$ in the semantic space. Thus, we can further build a mechanism to quantify the difference between the generated response and the golden response as follows:

$$\delta\alpha = \alpha_{(q,r)} - \alpha_{(q,\tilde{r})} \tag{16}$$

where $\tilde{r}$ is the generated response given by the generator and r comes from the original data. If $\delta\alpha$ moves toward zero, $\alpha_{(q,\tilde{r})}$ would be close to $\alpha_{(q,r)}$ sharing the same $F(q, r_c)$.

Consequently, a new expectation comes out. That is, the generator needs to provide results that can minimize $\delta\alpha$, so as to fit the local distribution of the existing responses to a given query. Thus, a hybrid objective of the generator can be finally defined as:

$$\min J^{G} = -\mathbb{E}\left[\log D(q, r)\right] + ReLU(\delta\alpha) \tag{17}$$

It should be noted that a hinge loss, conducted by the $ReLU$ function $ReLU(\delta\alpha) = \max(0,\ \delta\alpha)$, is especially introduced to reform $\delta\alpha$. The primary reason of this operation is that the $ReLU$ function has positive output only if $\delta\alpha \geqslant 0$, according to the definition of $ReLU(\delta\alpha)$. Apparently, $\delta\alpha < 0$ indicates that the generated response $\tilde{r}$ is too far from the center of the response cluster in the semantic space, so that its relevance may be highly questionable. Meanwhile, minimizing a negative variable is against the optimization direction. After the ReLU transformation, there remains valid loss only when $\delta\alpha \geqslant 0$, and thus both the diversity and the relevance of generated results are taken into account.

4.3 The Phase-wise Optimization

According to the analysis in Section 2, the trivial adversarial training directed by $-\mathbb{E}\left[\log D(q, r)\right]$ can only determine the form of general responses to a given query. From the spatial perspective in the semantic space, the original adversarial objective is helpful to roughly locate the response cluster to be generated. However, the local distribution can not be captured by this procedure.

By contrast, according to the discussions above, the proposed hybrid objective actually provides a way to force the generated responses, originally gathering around the general form, to expand into the expected local shape described by the golden truth, by conducting a phase-wise optimizing operation. This mechanism can be detailed in an intuitive way:

Foundation: Once the DBM in Figure 1 is well-trained with the query-response corpus, the semantic center r_c of a Response Cluster can be determined by the given query q.

Phase-1: In the early stage of the adversarial training, a generated response $\tilde{r}$ is not semantically relevant to the query q. Thus, it can be inferred that $\tilde{r}$ is radially farther from the cluster center r_c than the golden response r. In this situation, according to Equation 15 and Equation 16, we can claim that $\delta\alpha \leq 0$. In this phase, the hyper objective goes back to the general adversarial objective due to the ReLU function. Thus, the model is trying to force the generated samples to approach the center of each cluster, ignoring local distributions.

Phase-2: During the adversarial training in *Phase-1*, the generated result $\tilde{r}$ will go approaching to the cluster center r_c, which means $\alpha_{(q,\tilde{r})} \to 0$. It should be noted that, for any meaningful existing training sample r, $\alpha_{(q,r)} > 0$. Therefore, at some point, it turns to $\delta\alpha > 0$ and the right part of the hybrid objective in Equation 17 takes effect. Consequently, for each given query, the distribution of the generated results will expand to fit the local distribution of the golden samples.

5 Experiments

5.1 Experimental Setups

Datasets. Our experiments are conducted on two main stream open-access conversation corpora: The Opensubtitles corpus and the Sina Weibo corpus. The OpenSubtitles dataset contains 5,200,000 movie dialogues, where we extract query-response pairs following (Xu et al., 2018; Li et al., 2016). The Sina Weibo Corpus (Shang et al., 2015a) contains 2,500,000 single-turn Chinese dialogues, in which the length of the query and response ranges from 4 to 30. We sample 100,000, and 2,000 unique query-response pairs as validation and testing dataset respectively from both of the corpora[4].

Baselines. For meaningful comparison, we introduce the following models as the baselines:

(1) ***Seq2Seq***: a sequence-to-sequence model trained with maximum likelihood estimation (MLE);

(2) ***Seq2Seq-MMI***: the NRG model with a Maximum Mutual Information (MMI) criterion (Li et al., 2016);

[4] Both the English and the Chinese datasets used in our experiments are uploaded to `https://www.dropbox.com/sh/k8i079gd2111lsb/AACLLt1NAzile543Da8Qs9tFa?dl=0`.

Table 1: Performances of LocalGAN and Baselines on the Opensubtitles and Weibo Datasets.

	Opensubtitle				Weibo			
Model	Dist-1	Dist-2	Ent4	Rel.	Dist-1	Dist-2	Ent4	Rel.
Seq2Seq	0.025	0.081	5.650	1.090	0.055	0.153	6.400	0.315
Seq2Seq-MMI	0.027	0.086	5.698	1.067	0.059	0.172	6.860	0.309
Adver-REGS	0.0296	0.098	5.701	1.113	0.061	0.181	7.658	0.320
GAN-AEL	0.030	0.100	5.733	1.106	0.062	0.183	7.765	0.318
AIM	0.0292	0.095	5.783	1.120	0.064	0.189	7.833	0.321
DAIM	0.031	0.103	5.873	1.098	0.067	0.195	8.042	0.316
LocalGAN	**0.036**	**0.110**	**6.073**	**1.132**	**0.071**	**0.212**	**8.561**	**0.327**

(3) ***Adver-REGS***: the NRG model trained using adversarial framework, in which the policy gradient was employed to transfer the reward of the discriminator to the generator (Yu et al., 2017; Li et al., 2017);

(4) ***GAN-AEL***: an adversarial framework with an approximate embedding layer for connecting the generator with the discriminator directly (Xu et al., 2017).

(5) ***AIM / DAIM***: the adversarial training strategy allowing distributional matching of synthetic and real responses and explicitly optimizing a variational lower bound on pairwise mutual information between the query and response, so as to improve the informativeness and diversity of generated responses (Zhang et al., 2018b)[5].

Evaluation Metrics. To evaluate diversity of generated results, we adopt three widely-applied metrics: Distinct-1 (**Dist-1**), Distinct-2 (**Dist-2**), and Entropy (**Ent4**) (Li et al., 2016; Zhang et al., 2018b; Jost, 2006). In addition, the relevance (**Rel.**) is measured by summing three embedding-based similarities (greedy, average, extreme) (Liu et al., 2016) upon the ground-truth and generated responses.

Training Details. The details of the model training are given as follows: The vocabulary size of both datasets is 40,000. The embedding layer of OpenSubtitles and Sina Weibo is initialized using 200-dimensional Glove vectors (Pennington et al., 2014) and 300-dimensional Weibo vectors (Li et al., 2018) respectively. All the models are first pre-trained by MLE, and then the models including Adver-REGS, GAN-AEL, AIM, DAIM and LocalGAN are trained with adversarial learning. The discriminator of Adver-REGS and GAN-AEL are based on CNN following (Yu et al., 2017; Xu et al., 2017), in which the filter sizes are set to (1,2,3,4) and the filter number is 128, while that of LocalGAN adopts DBM with (2×embedding_size, 128, 128) to represent the semantic of queries and responses. The hidden size of the generator is set to 256 and 512 in GAN-based models and Seq2Seq respectively. To guarantee the performance consistency of AIM and DAIM, we adopt the recommended parameter settings given by Zhang et al. (2018b). The experiments are conducted on the Tesla K80 GPU.

5.2 Results & Analysis

Table 1 lists quantitative results on the diversity and relevance of generated responses on both datasets. As shown by the results, compared to Seq2Seq and Seq2Seq-MMI, the GAN-based methods give better results on the diversity oriented metrics, including Dist-1, Dist-2 and Ent4. This observation indicates that adversarial learning does provide the meaningful guidance to NRG models to avoid some of the safe-responses.

It can be observed that LocalGAN outperforms the baselines with adversarial learning architecture (Adver-REGS, GAN-AEL, AIM and DAIM) on both the diversity metrics and the relevance metrics. Generally, a notable improvement on diversity may lead to some negative influence on relevance, and thus promoting the diversity of generated response while maintaining their relevance is essentially desired for any methodologies, which has been achieved by our LocalGAN. The performances of

[5] We have taken the codes of AIM and DAIM from `https://github.com/dreasysnail/converse_GAN` implemented by the authors of this work for comparisons.

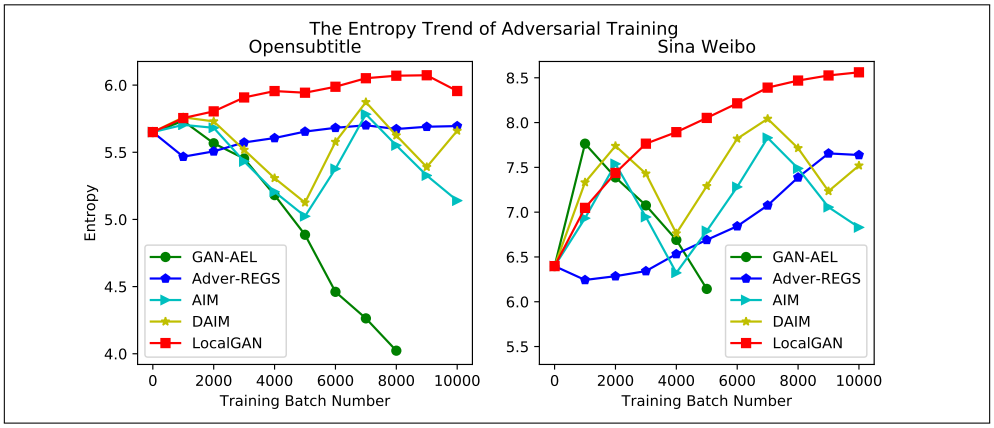

Figure 2: The Entropy Trend of LocalGAN, GAN-AEL, Adver-REGS, AIM and DAIM in the adversarial learning process.

LocalGAN can be attributed to the fact that LocalGAN has taken the local distribution of responses to a given query into account. By adopting the hybrid objective function, the proposed adversarial model gets to capture the spatial characteristics of response clusters, and the generation process is consequently forced to fit the semantic distributions of response clusters.

The training stability is a tough issue to be addressed for adversarial learning (Yu et al., 2017), and as discussed in the previous sections, one of the motivations of our LocalGAN is to make adversarial learning more stable. To valid this aspect, we track the changing of the Entropy (Ent4) of results given by GAN-AEL, Adver-REGS, AIM, DAIM and LocalGAN, as shown in Figure 2.

It can be observed that the training of LocalGAN and Adver-REGS is stable. By contrast, there exist obvious fluctuations on the curves of AIM and DAIM, and GAN-AEL rapidly gets out of control after 1000 batch, which makes it rather difficult to grasp the models with the best status. This group of results indicates the necessity of introducing additional restrictions into adversarial learning processes. For this purpose, Adver-REGS introduces a teacher-forcing loss (Li et al., 2017), while AIM and DAIM have taken the informativeness oriented constraints to partially control the stability (Zhang et al., 2018b). However, GAN-AEL only takes the Wasserstein distance as the objective (Xu et al., 2017), and thus the entropy goes down rapidly. Compared to the Adver-REGS, our LocalGAN achieves better diversity with even a more smooth entropy curve. The training of LocalGAN benefits from the the phase-wise optimization driven by the hybrid loss, and its stability also indicates the meaningfulness of modeling and utilizing local distributions of responses.

6 Conclusions

This paper has given the theoretical proof of the upper bound of the adversarial training leveraged models on the Seq2Seq-based neural response generation task. The proof indicates that, due to the local distribution nature of query-response corpora, the GAN based NRG models will converge to the states mostly generating specialized patterns corresponding to given queries. To address this issue, we proposed to model the local distribution of queries and their response in the semantic space by adopting energy-based function, and found the approximation of this function. According to this approximated distribution representation, a new loss function describing the local expansion cost in the fitting of response distribution is presented and finally combined with the traditional GAN loss to form a hybrid training objective for the GAN based NRG model. This paper provides a reasonable explanation to the unstable training process and unsatisfying results of GAN based NRG approaches, and meanwhile gives a different perspective to leverage the local data distribution to enhance classic GAN approaches.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.

Sanjeev Arora and Yi Zhang. Do gans actually learn the distribution? an empirical study. *arXiv preprint arXiv:1706.08224*, 2017.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 224–232. PMLR, 06–11 Aug 2017.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv e-prints*, abs/1409.0473, September 2014. URL `https://arxiv.org/abs/1409.0473`.

Ashutosh Baheti, Alan Ritter, Jiwei Li, and Bill Dolan. Generating more interesting responses in neural conversation models with distributional constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3970–3980, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

Jiachen Du, Wenjie Li, Yulan He, Ruifeng Xu, Lidong Bing, and Xuan Wang. Variational autoregressive decoder for neural response generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3154–3163, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. *arXiv preprint arXiv:1705.00106*, 2017.

Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1-3):70–87, 2006.

Karl Friston, Christopher Thornton, and Andy Clark. Free-energy minimization and the dark-room problem. *Frontiers in psychology*, 3:130, 2012.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems 27*, pp. 2672–2680, 2014.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimeoality of data with neural networks. *science*, 313(5786):504–507, 2006.

Geoffrey E Hinton and Richard S Zemel. Autoencoders, minimum description length and helmholtz free energy. In *Advances in neural information processing systems*, pp. 3–10, 1994.

Lou Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006.

Mahyar Khayatkhoei, Maneesh K Singh, and Ahmed Elgammal. Disconnected manifold learning for generative adversarial networks. In *Advances in Neural Information Processing Systems*, pp. 7343–7353, 2018.

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, 2014.

Abhishek Kumar, Prasanna Sattigeri, and Tom Fletcher. Semi-supervised learning with gans: Manifold invariance with improved inference. In *Advances in Neural Information Processing Systems*, pp. 5534–5544, 2017.

Nicolas Le Roux and Yoshua Bengio. Representational power of restricted boltzmann machines and deep belief networks. *Neural computation*, 20(6):1631–1649, 2008.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, pp. 110–119, 2016.

Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2157–2169, 2017.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 138–143, 2018.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2122–2132, 2016.

L. Mirsky. A trace inequality of john von neumann. *Monatshefte für Mathematik*, 79(4):303–306, Dec 1975. ISSN 1436-5081. doi: 10.1007/BF01647331. URL `https://doi.org/10.1007/BF01647331`.

Jiquan Ngiam, Zhenghao Chen, Pang W Koh, and Andrew Y Ng. Learning deep energy models. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 1105–1112, 2011.

Gaurav Pandey, Danish Contractor, Vineet Kumar, and Sachindra Joshi. Exemplar encoder-decoder for neural conversation generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1329–1338, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

K. B. Petersen and M. S. Pedersen. The matrix cookbook, nov 2012. URL `http://www2.imm.dtu.dk/pubdb/p.php?3274`. Version 20121115.

Ruslan Salakhutdinov and Geoffrey E. Hinton. Deep boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pp. 448–455, 2009.

Ruslan Salakhutdinov and Hugo Larochelle. Efficient learning of deep boltzmann machines. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 693–700, 2010.

Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pp. 1577–1586, 2015a.

Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pp. 1577–1586, 2015b.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2210–2219, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1235.

P. Smolensky. *Parallel Distributed Processing: Foundations*, volume 1, pp. 194–281. MIT Press, Cambridge, 1986.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 14th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 196–205, 2015.

Xingwu Sun, Jing Liu, Yajuan Lyu, Wei He, Yanjun Ma, and Shi Wang. Answer-focused and position-aware neural question generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3930–3939, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pp. 3104–3112, 2014.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, ukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. URL `http://arxiv.org/abs/1609.08144`.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. Topic aware neural response generation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. Diversity-promoting gan: A cross-entropy based generative adversarial network for diversified text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3940–3949, 2018.

Zhen Xu, Bingquan Liu, Baoxun Wang, SUN Chengjie, Xiaolong Wang, Zhuoran Wang, and Chao Qi. Neural response generation via gan with an approximate embedding layer. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 628–637, 2017.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, pp. 2852–2858, 2017.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, and Xueqi Cheng. Learning to control the specificity in neural response generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1108–1117, Melbourne, Australia, July 2018a. Association for Computational Linguistics.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems*, pp. 1810–1820, 2018b.

Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3901–3910, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

A The Graphical Model for Response Distribution Modeling

Different from the computer vision related scenario, training a DBM on a set of text vectors is not trivial, since the training procedure is difficult to converge due to the value scale of text vectors is much larger than image vectors. Moreover, the simple scaling methods are not effective enough for this issue. For this purpose, this paper adopts standard-scaler[6] to remove the mean and scale to

[6] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

unit variance. To valid the effectiveness of standard-scaler, We conduct experiments on the query-response matching task using normalized vectors. The experiment result show that the matching performance based normalized vectors is similar to that of CNN based architecture (Kim, 2014).

B Detailed Proof of Lemmas

Proof of Lemma 1. As mentioned before, we assume that $\mathbf{r} \sim N(\mathbf{r_c}, \Sigma)$. It is a reasonable assumption that response $\mathbf{r_i}$ is independent from other response $\mathbf{r_j}$, that is, a possible response to one query does not rely on other possible responses. More specific, $F(q, r_i)$ can be seen as an independent identically-distributed random variable.

Based on the law of large numbers,

$$\begin{aligned} F(q, R) &= \frac{1}{|R_q|} \sum_{r_i \in R_q} F(q, r_i) \\ &\to \mathrm{E}[F(q, \mathbf{r})] \qquad as\ |R_q| \to \infty \end{aligned} \tag{18}$$

Thus, when the possible responses are sufficient, which means R_q is large enough, we can say that for any small term $\epsilon > 0$

$$|F(q, R) - \mathrm{E}[F(q, \mathbf{r})]| \leq \epsilon$$

□

Proof of Lemma 2. For a fixed query, $F(q, \mathbf{r})$ can be seen as the the scalar function of vector $\mathbf{r}$. For simplicity, we denote $F(q, \mathbf{r})$ as $f(\mathbf{r})$.

Taylor expansions for the first moment of function of random variables are as follows.

$$\begin{aligned} \mathrm{E}[f(\mathbf{r})]] &= \mathrm{E}\left[f\left(\mathbf{r_c} + (\mathbf{r} - \mathbf{r_c})\right)\right] \\ &\approx \mathrm{E}\left[f\left(\mathbf{r_c}\right) + (\mathbf{r} - \mathbf{r_c})^T Df(\mathbf{r_c}) + \frac{1}{2}(\mathbf{r} - \mathbf{r_c})^\top \left\{D^2 f(\mathbf{r_c})\right\} (\mathbf{r} - \mathbf{r_c})\right] \end{aligned} \tag{19}$$

where $Df(\mathbf{r_c})$ is the gradient of f evaluated at $\mathbf{r} = \mathbf{r_c}$ and $D^2 f(\mathbf{r_c})$ is the Hessian matrix. Since $E\mathbf{r} = \mathbf{r_c}$, the second term $(\mathbf{r} - \mathbf{r_c})^T Df(\mathbf{r_c})$ disappears.

Now, simplify the third term.

- According to (Petersen & Pedersen, 2012), assume A is symmetric, $\mathbf{c} = E[\mathbf{x}]$ and $\mathbf{\Sigma} = \mathrm{Var}[\mathbf{x}]$, then
$$E\left[\mathbf{x}^T \mathbf{A} \mathbf{x}\right] = \mathrm{Tr}(\mathbf{A}\mathbf{\Sigma}) + \mathbf{c}^T \mathbf{A} \mathbf{c}.$$
- (Mirsky, 1975) states following theorem: If A, B are complex $n \times n$ matrices with singular values $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$ and $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n$ respectively, then
$$|\operatorname{Tr}(AB)| \leq \sum_{i=1}^{n} \alpha_i \beta_i$$
- A symmetric matrix is positive semi-definite if and only if all eigenvalues are non-negative.

On the one hand, since $\mathbf{r} - \mathbf{r_c} \sim N(0, \Sigma)$ and Σ is positive semi-definite matrix, the third term can be simplified as follows.

$$\begin{aligned} |\mathrm{E}[(\mathbf{r} - \mathbf{r_c})^\top \left\{D^2 f(\mathbf{r_c})\right\} (\mathbf{r} - \mathbf{r_c})]| &= |\operatorname{Tr}(\{D^2 f(\mathbf{r_c})\}\Sigma) + \mathbf{0}^\top D^2 f(\mathbf{r_c})\mathbf{0}| \\ &= |\operatorname{Tr}(\{D^2 f(\mathbf{r_c})\}\Sigma)| \\ &\leq \sum_{i=1}^{n} \alpha_i \beta_i \\ &\leq \alpha_1 \sum_{i=1}^{n} \beta_i \\ &= \alpha_1 \mathrm{Tr}(\Sigma) \end{aligned} \tag{20}$$

where $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$ and $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n$ denote the singular value of matrix $\{D^2 f(\mathbf{r_c})\}$ and Σ respectively.

On the other hand, the expected Euclidean distance between random variable $\mathbf{r}$ and constant variable $\mathbf{r_c}$ can be written as

$$\mathrm{E}\left[\|\mathbf{r} - \mathrm{E}[\mathbf{r}]\|_2^2\right] = \sum_{i=i}^{n} \Sigma_{ii} = \mathrm{Tr}(\Sigma).$$

Thus, for a small value $\epsilon > 0$,

$$\mathrm{E}\left[\|\mathbf{r} - \mathrm{E}[\mathbf{r}]\|_2^2\right] \leq \epsilon \Longleftrightarrow \mathrm{Tr}(\Sigma) \leq \epsilon$$

Hence, under the assumption that the Euclidean distance between random variable $\mathbf{r}$ and constant variable $\mathbf{r_c}$ is small enough, we have

$$\begin{aligned} |\mathrm{E}[f(\mathbf{r})] - f(\mathbf{r_c})| &\approx |\frac{1}{2}\mathrm{E}[(\mathbf{r} - \mathbf{r_c})^\top \left\{D^2 f(\mathbf{r_c})\right\}(\mathbf{r} - \mathbf{r_c})]| \\ &\leq \frac{1}{2}\alpha_1 \mathrm{Tr}(\Sigma) \\ &\leq \frac{1}{2}\alpha_1 \epsilon \end{aligned} \tag{21}$$

To observe the maximum singular value α_1 of hessian matrix of $f(r)$, we first substitute the definition of $E(q, r, H)$ into $F(q, r)$ and have following equation:

$$f(r) = F(q, r) = -\log \sum_{h_q, h} exp(h_q^T W_{qr} r + h_q^T W_{qq} q + h^T W_{qh} h_q).$$

After that, its first-order and second-order partial derivative are calculated as follows:

$$\frac{\partial f(r)}{\partial r_i} = -\frac{\sum_{h_q,h} exp(h_q^T W_{qr} r + h_q^T W_{qq} q + h^T W_{qh} h_q) \times (h_q^T W_{qr})_i}{\sum_{h_q,h} exp(h_q^T W_{qr} r + h_q^T W_{qq} q + h^T W_{qh} h_q)}$$

$$\frac{\partial^2 f(r)}{\partial r_i \partial r_j} = \sum_{h_q,h} a(h_q, h) \times b(h_q, i) \times \left[\sum_{\tilde{h}_q, \tilde{h}} a(\tilde{h}_q, \tilde{h}) \times b(\tilde{h}_q, j) - b(h_q, j)\right]$$

where $a(h_q, h) = \frac{exp(h_q^T W_{qr} r + h_q^T W_{qq} q + h^T W_{qh} h_q)}{\sum_{\tilde{h}_q \tilde{h}} exp(\tilde{h}_q^T W_{qr} r + \tilde{h}_q^T W_{qq} q + \tilde{h}^T W_{qh} \tilde{h}_q)}$ and $b(h_q, i) = (h_q^T W_{qr})_i$ representing the i-th element of vector $h_q^T W_{qr}$. Meanwhile, according to the definition of $a(h_q, h)$, it is obvious that $a(h_q, h) > 0$ and $\sum_{h_q,h} a(h_q, h) = 1$.

Denoting $\max_{h_q,i} |b(h_q, i)|$ as M (h_q follows multinomial distribution), the following inequality holds:

$$\begin{aligned} \left|\frac{\partial^2 f}{\partial r_i \partial r_j}\right| &\leq \sum_{h_q,h} |a(h_q, h)| \times |b(h_q, i)| \times \left[\sum_{\tilde{h}_q, \tilde{h}} \left|a(\tilde{h}_q, \tilde{h}) \times b(\tilde{h}_q, j)\right| + |b(h_q, j)|\right] \\ &\leq \sum_{h_q,h} a(h_q, h) \times M \times 2M = 2M^2 \end{aligned}$$

Therefore, we have

$$\alpha_1 = \sigma_{\max}(A) \leq \|A\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} |\frac{\partial^2 f}{\partial r_i \partial r_j}|^2} \leq 2nM^2$$

where n is the vector dimension.

As mentioned in Appendix A, DBM inputs are normalized by standard-scaler, and thus M is less than 1. Consequently, $2nM^2$ can be approximate to a controllable constant.

To conclude, if the Euclidean distance between random variable $\mathbf{r}$ and constant variable $\mathbf{r_c}$ is small enough, $|F(q, \mathbf{r_c}) - \mathrm{E}[F(q, \mathbf{r})]|$ is small enough as well. □