On Empirical Bayes Variational Autoencoder: An Excess Risk Bound

Rong Tang RONGT3@ILLINOIS.EDU
University of Illinois at Urbana-Champaign
Yun Yang YY84@ILLINOIS.EDU
University of Illinois at Urbana-Champaign

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

In this paper, we consider variational autoencoders (VAE) via empirical Bayes estimation, referred to as Empirical Bayes Variational Autoencoders (EBVAE), which is a general framework including popular VAE methods as special cases. Despite the widespread use of VAE, its theoretical aspects are less explored in the literature. Motivated by this, we establish a general theoretical framework for analyzing the excess risk associated with EBVAE under the setting of density estimation, covering both parametric and nonparametric cases, through the lens of M-estimation. As an application, we analyze the excess risk of the commonly-used EBVAE with Gaussian models and highlight the importance of covariance matrices of Gaussian encoders and decoders in obtaining a good statistical guarantee, shedding light on the empirical observations reported in the literature.

1. Introduction

A wide variety of machine learning problems can be framed as directed probabilistic inference in generative models (Jebara and Meila, 2006), especially when we care about modeling and efficient sampling from complex distributions such as those over natural images and text (Yang et al., 2017; Brock et al., 2018; van den Oord et al., 2016). Variational autoencoder (VAE) (Kingma and Welling, 2013; Rezende et al., 2014) replaces conventional instance-specific local inference with a global inference network and therefore enables efficient training of deep generative models. In plain language, a latent variable generative model defines a joint density $p(x, z)$ over the data space $\mathcal{X}$ and the latent space $\mathcal{Z}$ by specifying a prior $\pi(z)$ over latent variables and a conditional density $p(x|z)$ of data given latent variables. Typically we aim at learning $p_{\mathcal{D}}(x)$ over data space, based on a finite number n of samples $\{x_i\}_{i=1}^n$, assumed to be drawn from it. In most cases, maximizing the average marginal log-likelihood of the data is difficult, as the marginal likelihood functions are intractable due to the integral for marginalizing out latent variables (Kingma and Welling, 2013). VAE overcomes this issue by introducing a family of inference distributions $q(z|x)$ for approximating the posterior of latent variables given the data and jointly optimizing the so-called evidence lower bound (ELBO, Ormerod and Wand, 2010) as in the variational Bayes methods. From a coding theory perspective, the unobserved latent variables can be interpreted as a latent representation or code (Kingma and Welling, 2013). Therefore, the inference distribution $q(z|x)$ can be interpreted as a probabilistic encoder, and the conditional distribution $p(x|z)$ of data given latent variables can be interpreted as a probabilistic decoder.

VAE has received great success in generating complicated data, including images (Gregor et al., 2015; Kulkarni et al., 2015), molecules (Segler et al., 2017), text (Yang et al., 2017), and predicting the future from static images (Walker et al., 2016). However, as empirically observed in Tomczak and Welling (2017), VAE with a standard multivariate Gaussian prior tends to underfit the data. We thus consider a broader class of VAE via empirical Bayes estimation. Specifically, we incorporate hyperparameters in the prior over latent variables, and jointly optimize the prior with the encoder and the decoder. We call this framework Empirical Bayes Variational Autoencoders (EBVAE), which includes popular VAE variants "VampVAE" (Tomczak and Welling, 2017) and "LARSVAE" (Bauer and Mnih, 2018) as two representative examples. In the statistical literature, density estimation (Silverman, 1986; Sheather, 2004) has been an important topic in both nonparametric statistics and parametric statistics, and its hardness in terms of minimax optimal rate of convergence has been understood fairly well for a wide range of density functions under smoothness constraints (Stone, 1982). Despite the celebrated empirical success, little general theory has been developed to investigate statistical properties of VAE or more broadly, EBVAE (Doersch, 2016). In this paper, we undertake this task and focus on the theoretical front to answer: how well can EBVAE learn the target density $p_{\mathcal{D}}(x)$ under different choices of prior families, encoder families, and decoder families.

1.1. Related Work

In the original formulation of VAE, the prior is chosen to be the standard multivariate Gaussian and the encoder is optimized over a Gaussian family (Kingma and Welling, 2013), which may lead to poor performance when applied to complex datasets because of model misspecification. Many approaches have been developed to increase the model capacity by either using a more flexible encoder family (Rezende and Mohamed, 2015; Kingma et al., 2016) or choosing a more expressive family of the priors (Chen et al., 2016; Guillemin and Pollack, 2010). Tomczak and Welling (2017) have shown that the prior minimizing the objective function of VAE is given by the corresponding aggregated posterior $\frac{1}{n}\sum_{i=1}^{n} q(z|x_i)$ with $q(z|x)$ being the encoder. In view of this fact, some studies (Tomczak and Welling, 2017; Bauer and Mnih, 2018) considered prior families that aim to approximate the aggregated posterior, which can be seen as special cases of prior parametrization within the framework of EBVAE.

On the theoretical side, Liang (2018) studied the rates of convergence for learning generative models using Generative Adversarial Networks (GAN, Goodfellow et al., 2014). They provided a comprehensive statistical treatment of GAN in which the generator and discriminator are parametrized by neural networks. Unlike GAN which aims at achieving an equilibrium between the generator and the discriminator, EBVAE aims at maximizing a variational lower bound to the data log-likelihood and possess an encoder-decoder type interpretation. In this work, we develop a general theoretical framework to characterize the excess risk of EBVAE as a generative model learning approach for density estimation covering both parametric and nonparametric cases. A most relevant work to ours is Doersch (2016), where they analyzed the approximation error associated with the population level objective function of VAE for one-dimensional data when Gaussian encoders and decoders are used, they found that the approximation error will go to zero if the standard deviation (noise level) of the data given latent variables vanishes, given that the approximation families of mean functions and covariance functions of the Gaussian models have enough capacity. In our study, we give a excess risk bound on the estimator arising from EBVAE with Gaussian models, which includes a term depend on the sample size due to random fluctuations and therefore enables us to study the finite sample performance of the EBVAE estimator (c.f. Theorem 7).

1.2. Summary of Contributions

Below is a summary of our main theoretical contributions in the paper.

1. *We provide the first rigorous theoretical analysis to the excess risk of EBVAE.*

Despite the empirical success of VAE, to the best of our knowledge, there is no general theory about the statistical properties of the resulting estimator. A systematic theoretical study on VAE enables practitioners to be aware of whether their resulting estimators are reliable and provide guidance on how to set the best hyperparameters and approximation families in concrete situations. In this study, we address the problem by giving a general statistical framework to analyze the excess risk for learning densities using EBVAE. The key insight of our work comes from representing the EBVAE estimator as an M-estimator (see for example, Chapter 5 of Vaart (1998)). Once we make the connection, we can leverage the rich toolkit of theoretical and methodological results available for this context. We develop novel oracle inequalities (c.f. Theorem 1) that provide general tools to verify the statistical accuracy of estimators arising from EBVAE and give insight about which decoder families, encoder families and prior families yield consistency.

2. *As an application, we analyze the risk of estimators derived from the commonly-used EBVAE with Gaussian encoders and decoders in Theorem 7.*

The theory we established for EBVAE estimators with Gaussian models highlights the importance of the covariance matrix of the Gaussian encoder, which is often chosen as a diagonal matrix in practice. For example, our theory suggests that the approximation error of EBVAE with Gaussian encoders is strictly related to the model of covariance matrices of encoders, misspecifying the off-diagonal elements will introduce extra errors. As an implication, the covariate parameters of Gaussian decoders, which are often chosen to be independent of the data in advance, should be jointly optimized with other parameters. This explains the reason why Vanilla VAE models tend to produce unrealistic, blurry samples when applied to complex datasets of natural images (Dosovitskiy and Brox, 2016). As another implication of our theory, the limited capacity of parametric families such as Gaussians suggests the necessity of using more complicated encoder/decoder models and thus we follow the classic nonparametric literature by considering a broad class of nonparametric families characterized by smoothness levels, and quantify the accompanied approximation error and estimation error.

3. *We build a uniform law with a data-dependent complexity specifically tailored to handle the unbounded loss function associated with EBVAE.*

Due to our delicate localization technique in the proof, we obtained a “fast rate” (i.e. n^{-1} rate in case of parametric models) without assuming the boundedness of loss function (w.r.t. data x) as opposed to a “slow rate” (i.e. $n^{-1/2}$). This is achieved by our key localization Lemma 12 and Lemma 13. Specifically, Lemma 12 provides a “maximal” type inequality for controlling the supreme of an unbounded empirical process specifically constructed for dealing with the loss function involving the Killback-Leibler divergence.This inequality captures the local fluctuation behavior of our empirical loss function via the variance of the increments of an empirical process. Its proof involves non-trivial applications of many empirical process techniques such as chaining and peeling. Lemma 13 provides an upper bound to the local Rademacher complexity (Bartlett et al., 2005) associated with unbounded functions, which enables us to deal with the unbounded loss function associated with EBVAE.

4. *We take the low-dimensional structure of data space into account and illustrate that EBVAE can benefit from the underlying submanifold structure.*

Specifically, our results for EBVAE with Gaussian encoders/decoders (c.f. Theorem 7) show the adaptiveness of EBVAE to lower-dimensional submanifold structures so that the bound does not suffer from the "curse of dimensionality". This is achieved by our Lemma 17 that provides an error bound of ReLU neural networks for approximating smooth functions with domain being close to a d_z-dimensional submanifold and Lemma 18 that gives an explicit dependence of the excess risk and approximation error of EBVAE estimators on the variance of the data given latent variables.

1.3. Notations.

We summarize some necessary notations and definitions here. We use $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$ to denote the data space and the latent space, $p(x|z)$ to denote the decoder, $q(z|x)$ to denote the encoder and $\pi(z)$ to denote the prior for the latent variable. To simplify the notation, we may also use shorthands p, q, π when no ambiguity may arise. In the parametric case, we use $\theta \in \Theta_\theta$, $\phi \in \Theta_\phi$ and $\beta \in \Theta_\beta$ to denote the parameters associated with the decoder family $\mathcal{F}_{dd}$, the encoder family $\mathcal{F}_{ed}$ and the prior family $\mathcal{F}_{prior}$ respectively, and use $p_\theta(x|z)$, $q_\phi(z|x)$ and $\pi_\beta(z)$ with shorthands $p_\theta, q_\phi, \pi_\beta$ to denote the decoder, encoder and prior in these families. We use $p_{\mathcal{D}}(x)$ to denote the target data distribution and $\{x_i\}_{i=1}^n$ to denote n i.i.d. copies generated from $p_{\mathcal{D}}(x)$.

For a d-dimensional Euclidean vector x, we use $\|x\|_p$ to denote its ℓ_p norm. For a function $f(x) : \mathbb{R}^{d_1} \mapsto \mathbb{R}^{d_2}$, $\nabla f(x)$ is a $d_2 \times d_1$ matrix, with $(\nabla f(x))_{i,j} = \frac{\partial f_i(x)}{\partial x_j}$. $D_{\mathrm{TV}}(p, q) = \frac{1}{2} \int |p(x) - q(x)| dx$ denotes the total variation distance and $D_{\mathrm{KL}}(p||q) = \int \log \frac{p(x)}{q(x)} p(x) dx$ denotes the Kullback-Leibler (KL) divergence. We use $\mathcal{N}(\mu, \Sigma)$ to denote the multivariate Gaussian distribution with mean vector $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. The symbols $\lesssim$ and $\gtrsim$ mean the corresponding inequality up to an n-independent constant. For multi-indexes $\gamma = (\gamma_1, \cdots, \gamma_d) \in \mathbb{N}_0^d$, a function f is said to be of class $\mathbf{C}^k$ ($k \in \mathbb{N}_{\geq 0}$) if all partial derivative of order γ ($\|\gamma\|_1 \leq k$) exist and are continuous. We use $C^\alpha(\Omega)$ to denote the Hölder space on Ω with Hölder exponent $\alpha > 0$ (see for example, Evans (2010)), and we use $B_r^\alpha(\Omega)$ to denote the closed ball in $C^\alpha(\Omega)$ with Hölder norm $\|\cdot\|_{C^\alpha(\Omega)}$ being bounded by r. We will also use the definition of Orlicz norms (see e.g. Dudley (1999)), recalled next. For $\alpha > 0$, define the function $\psi_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ with the formula $\psi_\alpha(x) = \exp(x^\alpha) - 1$. For a random variable X, we define its Orlicz norm with respect to ψ_α as

$$\|X\|_{\psi_\alpha} = \inf\left\{\lambda > 0 : \mathbb{E}\big[\psi_\alpha(|X|/\lambda)\big] \leq 1\right\}.$$

By standard analysis, we have for all $t > 0$,

$$\mathbb{P}(|X| \geq t) \leq 2\exp\left\{ -\Big(\frac{t}{\|X\|_{\psi_\alpha}}\Big)^\alpha \right\}.$$

1.4. Organization

The rest of the paper is organized as follows. In Section 2, we give a brief description of EBVAE. In Section 3, we develop an oracle inequality (Theorem 1); and in Section 4, we apply our oracle inequality to parametric and nonparametric cases. The paper is concluded with a discussion in Section 5. For the Appendix: a numerical study is included in Appendix A; the proofs of the main results are included in Appendix B; and the proofs of technical lemmas are included in Appendix C.

2. Empirical Bayes Variational Autoencoder

Suppose we have a dataset of x samples from a distribution that can be modelled by a generative model. Here, a generative model defines a joint distribution over the latent space $\mathcal{Z}$ and the data space $\mathcal{X}$. Usually we specify a simple prior distribution $\pi(z)$ over the latent variables, such as isotropic multivariate Gaussian or uniform, and model the data distribution by complex conditional distributions (decoders) $p(x|z) \in \mathcal{F}_{dd}$, where $\mathcal{F}_{dd}$ can either be a parametric or nonparametric family. The goal of VAE is to learn the true underlying marginal likelihood of the data $p_{\mathcal{D}}(x)$ in the generative process. Given data $\{x_i\}_{i=1}^n$, we typically aim at maximizing the average marginal log-likelihood $\frac{1}{n}\sum_{i=1}^n \log \int p(x_i|z)\pi(z)dz$. However, the optimization could be computationally infeasible due to the potentially high dimensional integral in the objective function, so it will be convenient to resort to VAE. Specifically, VAE overcomes this issue by introducing a family of encoders $q(z|x) \in \mathcal{F}_{ed}$ and jointly maximize a lower bound to the log likelihood (Kingma and Welling, 2013),

$$\frac{1}{n}\sum_{i=1}^n \left\{ \log \int p(x_i|z)\pi(z)dz - D_{\mathrm{KL}}\left(q(\cdot|x_i) \,\middle|\middle|\, \frac{p(x_i|\cdot)\pi(\cdot)}{\int p(x_i|z)\pi(z)dz} \right) \right\},$$

which is equivalent to (up to constants)

$$\frac{1}{n}\sum_{i=1}^n \left\{ \int \log p(x_i|z) q(z|x_i)dz - D_{\mathrm{KL}}\big(q(\cdot|x_i) \,||\, \pi(\cdot)\big) \right\}.$$

This objective function is computationally more friendly to optimize since the highest density region of $q(z|x)$ may be relatively smaller compared with the space of z under the prior.

In the original setting of VAE, the prior is chosen to be simple and data-independent and the decoder is chosen to be from a Gaussian family for continuous data, i.e, $\mathcal{N}(G_\theta(z), \sigma^2 I)$ with $G_\theta(z)$ being implemented with multi-layer perceptron (fully-connected neural networks with one hidden layer). Even though any d-dimensional distribution can be generated as a push forward measure through the standard d-dimensional Gaussian (Devroye, 2006), we may need a highly non-regular map to first map the fixed prior to a complicated distribution of latent variables. This may lead to underfitting if the decoder families have low capacity. To address this issue, we increase the model capacity by introducing hyperparameters in the prior and jointly training the prior distribution of the latent variable over a prior family $\mathcal{F}_{prior}$ with the encoder and decoder (see Appendix A for a numerical comparison). This lead to the EBVAE as the following optimization problem,

$$\min_{p\in\mathcal{F}_{dd}, q\in\mathcal{F}_{ed}, \pi\in\mathcal{F}_{prior}} \frac{1}{n}\sum_{i=1}^n \left[-\log \int p(x_i|z)\pi(z)dz + D_{\mathrm{KL}}\left(q(\cdot|x_i) \,\middle|\middle|\, \frac{p(x_i|\cdot)\pi(\cdot)}{\int p(x_i|z)\pi(z)dz} \right) \right].$$

The objective function of EBVAE can also be rewritten as $n^{-1}\sum_{i=1}^n \big(-\int \log p(x_i|z)q(z|x_i)dz + D_{\mathrm{KL}}(q(\cdot|x_i) \,||\, \pi(\cdot))\big)$ for facilitating computation. During the learning, we can apply Monte Carlo method to approximate the above objective function using draws sampled from $q(z|x)$.

3. Main Theoretical Results

Despite its popularity, the theoretical aspects of EBVAE are less explored in literature. In this section, we will study the general statistical properties of the EBVAE estimator through the lens of M-estimation. As introduced in Section 2, we define the following loss function for a single data x,

$$m(p, q, \pi, x) = \log \frac{p_{\mathcal{D}}(x)}{\int p(x|z)\pi(z)dz} + D_{\mathrm{KL}}\left(q(\cdot|x) \,\middle|\middle|\, \frac{p(x|\cdot)\pi(\cdot)}{\int p(x|z)\pi(z)dz} \right), \tag{1}$$

where we deliberately added the term $\log p_{\mathcal{D}}(x)$ which is independent of (p, q, π) to the loss function for the sake of theoretical analysis. With this notation, the EBVAE estimator can be casted as the following M-estimator,

$$(\hat{p},\ \hat{q},\ \hat{\pi}) = \underset{p\in\mathcal{F}_{dd}, q\in\mathcal{F}_{ed}, \pi\in\mathcal{F}_{prior}}{\arg\min} \left\{ n^{-1}\sum_{i=1}^{n} m(p, q, \pi, x_i) \right\}, \tag{2}$$

where recall that $\mathcal{F}_{dd}$ denotes the decoder family, $\mathcal{F}_{ed}$ denotes the encoder family and $\mathcal{F}_{prior}$ denotes the prior family. In the population level, we can also define

$$\Psi^* = \underset{p\in\mathcal{F}_{dd}, q\in\mathcal{F}_{ed}, \pi\in\mathcal{F}_{prior}}{\arg\min} \mathbb{E}_{p_{\mathcal{D}}(x)}\big[m(p, q, \pi, x)\big]$$

as the set of minimizers of the population level loss function.

The goal of this section is to study the finite sample performance of the point estimator obtained from EBVAE, which is captured by the so-called oracle inequality (Rigollet and Hütter, 2015). We prove a general oracle inequality for the EBVAE estimator (2) with risk function being chosen as the population level loss function $\mathbb{E}_{p_{\mathcal{D}}(x)}\big[m(p, q, \pi, x)\big]$ in the next theorem. According to the definition of the loss function in (1), the risk function can be decomposed into two components (c.f. Theorem 1). The first component of the risk function quantifies the difference between the target density and the marginal density $\int p(x|z)\pi(z)dz$ relative to the KL divergence, while the second component quantifies the difference between the encoder and the posterior $\frac{p(x|z)\pi(z)}{\int p(x|z)\pi(z)dz}$ relative to the KL divergence. Including the second term in the risk function brings several benefits. By writing the objective function of EBVAE as an empirical counterpart of the risk function as in (2), we can therefore leverage the existing theory of M-estimation to build an oracle inequality. In addition, since the second term in the risk function is always nonnegative, the risk function evaluated at $(\hat{p}, \hat{q}, \hat{\pi})$ also acts as an upper bound to the KL divergence between the fitted marginal density and the target distribution. On the computational side, according to Kingma and Welling (2013), the loss function defined in (1) can be regarded as a computationally efficient surrogate to $\log \int p(x|z)\pi(z)dz$ in the definition of maximum likelihood estimator (MLE) with error $D_{\mathrm{KL}}\big(q(\cdot|x)\,\big|\big|\,\frac{p(x|\cdot)\pi(\cdot)dz}{\int p(x|z)\pi(z)dz}\big)$, which is quantified by the second component of the risk function in the population level. We then impose the following assumption for controlling the tail for the suprema of an unbounded empirical process (Adamczak, 2008; Mendelson et al., 2007) appearing in the analysis of EBVAE.

Assumption A For a random variable X with density $p_{\mathcal{D}}(x)$, there exist some positive constants (α, D) such that

$$\left\| \sup_{\substack{p\in\mathcal{F}_{dd}, q\in\mathcal{F}_{ed}\\ \pi\in\mathcal{F}_{prior}}} \left\{ \left| \log \frac{\int p(X|z)\pi(z)dz}{p_{\mathcal{D}}(X)} \right| + D_{\mathrm{KL}}\left(q(\cdot|X)\,\Big|\Big|\, \frac{p(X|\cdot)\pi(\cdot)}{\int p(X|z)\pi(z)dz} \right) \right\} \right\|_{\psi_\alpha} \le D.$$

Roughly speaking, Assumption A is a tail condition on the loss function so that the population level loss function and its empirical counterpart can be proved to be close to each other uniformly. Similar assumptions are commonly made in the literature (Grünwald and Mehta, 2020). Our assumption is comparable to Grünwald and Mehta (2020) on fast rates for unbounded loss where the uniform boundedness is only in terms of parameters, but not data X. We show in Theorem 7 that Assumption A is applicable to commonly used encoder/decoder examples. Moreover, for parametric models (c.f. Section 4.1) where absolute values of logarithms of density functions $p_D(x)$, $p_\theta(x|z)$, $q_\phi(z|x)$ and $\pi_\beta(z)$ grow at most polynomially in $\|x\|_2$ and the parameters (θ,ϕ,β), if the parameter space and latent space are bounded and the data X has bounded Orlicz norm with a suitable $\alpha>0$ (e.g. sub-Gaussian and sub-exponential), then Assumption A holds. This requirement holds for any regular exponential family. Note that Assumption A also holds when the quantity inside the norm is bounded. For any $(p^*,q^*,\pi^*)\in\Psi^*$, consider the shifted function class,

$$G^*=\big\{g(x)=m(p,q,\pi,x)-m(p^*,q^*,\pi^*,x)\,\big|\,p\in\mathcal{F}_{dd},\ q\in\mathcal{F}_{ed},\ \pi\in\mathcal{F}_{prior}\big\}.$$

Define the star hull of G^* as $\overline{G}^*=\{ag\,|\,a\in(0,1],\ g\in G^*\}$. To bound the estimation error, we need certain data-dependent estimate of the complexity of $\overline{G}^*$, namely, the local Rademacher complexity (Bartlett et al., 2005), defined by

$$\overline{R}_n(\delta,\overline{G}^*)=\mathbb{E}_{p_\mathcal{D}(x)}\mathbb{E}_\varepsilon\left[\sup_{g\in\overline{G}^*,\|g\|_2\le\delta}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i g(x_i)\right|\right],$$

where $\{\varepsilon_i\}_{i=1}^n$ are n i.i.d. copies from Rademacher distribution, i.e. $P(\varepsilon_i=1)=P(\varepsilon_i=-1)=\frac{1}{2}$ and $\|g\|_2^2=\int_\mathcal{X}g^2(x)p_\mathcal{D}(x)dx$. We are then ready to state the following theorem that provides oracle result of EBVAE estimator.

Theorem 1 *Consider the EBVAE estimator $\hat p$, $\hat q$ and $\hat\pi$ defined in (2). Under Assumption A, if there exist $\delta_n>0$ and $(p^*,q^*,\pi^*)\in\Psi^*$, such that: (1) $\overline{R}_n(\delta_n,\overline{G}^*)\le\delta_n^2/(D\log^{\frac{1}{\alpha}}n)$; (2) $(n\delta_n^2/(D^2\log^{\frac{2}{\alpha}}n))^{\min\{\alpha,1\}}\ge\log(\log\frac{D}{\delta_n})$, then there exist constants (c_0,c_1,c_2) only dependent of α, such that it holds with probability at least $1-c_0\exp\Big\{-c_1\Big(\frac{n\delta_n^2}{D^2\log^{\frac{2}{\alpha}}n}\Big)^{\min\{\alpha,1\}}\Big\}$ that,*

$$\mathbb{E}_{p_\mathcal{D}(x)}\big[m(\hat p,\hat q,\hat\pi,x)\big]\le\inf_{\gamma>0}\left\{(1+\gamma)\min_{\substack{p\in\mathcal{F}_{dd},q\in\mathcal{F}_{ed},\\ \pi\in\mathcal{F}_{prior}}}\mathbb{E}_{p_\mathcal{D}(x)}\big[m(p,q,\pi,x)\big]+c_2\Big(1+\frac{1}{\gamma}\Big)\delta_n^2\frac{\log^{-\frac{1}{\alpha}}n}{D}\right\},$$

where we can decompose $\mathbb{E}_{p_\mathcal{D}(x)}\big[m(p,q,\pi,x)\big]=D_{\mathrm{KL}}\Big(p_\mathcal{D}(\cdot)\,\Big|\Big|\int p(\cdot|z)\pi(z)dz\Big)+$ $\mathbb{E}_{p_\mathcal{D}(x)}\Big[D_{\mathrm{KL}}\Big(q(\cdot|x)\,\Big|\Big|\,\frac{p(x|\cdot)\pi(\cdot)}{\int p(x|z)\pi(z)dz}\Big)\Big]$.

Remark 2 *The constant c_2 has a polynomial dependence on $\lceil 1/\alpha\rceil!=\prod_{j=1}^{\lceil 1/\alpha\rceil}j$ when $\alpha\le1$, so there is a super-exponential dependence on α. The main tool for proving Theorem 1 is the tail inequality for the suprema of an unbounded empirical process (Adamczak, 2008). One major difficulty is that the tail bound applies only to a deterministic radius δ, as opposed to the random radius $\|\hat g\|_2=\|m(\hat p,\hat q,\hat\pi,\cdot)-m(p^*,q^*,\pi^*,\cdot)\|_2$. This issue can be solved by using the "peeling" argument (Wainwright, 2019), i.e., considering sets $\mathcal{S}_m=\{2^{m-1}\delta_n\le\|\hat g\|_2\le2^m\delta_n\}$ with $m=1,\cdots\log(D/\delta_n)$. See Appendix B.1 for further details.*

The result in Theorem 1 can be used to determine a set of sufficient conditions under which the EBVAE estimator is consistent. An estimator is called consistent if it converges to its estimand as sample size increases, which gives a guarantee that we could get the right answer of parameters of interest based on the estimator for large sample sizes. The first term of the bound in

Theorem 1 corresponds to the approximation error and tends to be small as the encoder family, decoder family and prior family become richer. In the next section, we give instances of $\mathcal{F}_{dd}$, $\mathcal{F}_{ed}$ and $\mathcal{F}_{prior}$ leading to zero approximation error in concrete examples. In particular, in the usual setting where the target data distribution $p_{\mathcal{D}}(x)$ can be expressed as $\int p^*(x|z)\pi^*(z)dz$ with $p^*(x|z) \in \mathcal{F}_{dd}$ and $\pi^*(z) \in \mathcal{F}_{prior}$. The approximation error can be further upper bounded by $\min_{q\in\mathcal{F}_{ed}} \mathbb{E}_{p_{\mathcal{D}}(x)}\big[D_{\mathrm{KL}}\big(q(\cdot|x) \,\big|\big|\, p^*(x|\cdot)\pi^*(\cdot)/p_{\mathcal{D}}(x)\big)\big]$, which validates the importance of choosing a suitable encoder family. In practice, many approaches have been developed to increase the empirical performance of VAE by using flexible encoder families, e.g. NF (Rezende and Mohamed, 2015), IAF (Kingma et al., 2016), which outperform the Vanilla VAE. The second term of the bound in Theorem 1 corresponds to the estimation error which tends to be small as complexities of the encoder family, decoder family and prior family decrease. In particular, the deterministic radius δ_n in the estimation error term is called the critical radius associated with $\overline{G}^*$ (Wainwright, 2019), which is commonly used to specify bounds on the excess risk in M-estimation problems. We will determine δ_n in some representative examples under different choices of $\mathcal{F}_{dd}$, $\mathcal{F}_{ed}$ and $\mathcal{F}_{prior}$ in Section 4. Ideally, we want to make a choice to $\mathcal{F}_{dd}$, $\mathcal{F}_{ed}$ and $\mathcal{F}_{prior}$ such that the approximation error and estimation error are well-balanced.

4. Applications

In this section, we apply Theorem 1 to some representative examples. In each case, we will determine the approximation error and solve the δ_n in Theorem 1 via bounding the local Rademacher complexity from above by Dudley's integral (see, for example, (8.13) of Vershynin, 2018) to obtain an explicit excess risk bound in terms of model characteristics.

4.1. Parametric Models

In this subsection, we consider the case when $\mathcal{F}_{dd}$, $\mathcal{F}_{ed}$ and $\mathcal{F}_{prior}$ are parametric families. Recall that to explicitly express the dependence of the encoder and decoder on the parameters, we adopt the notation in (Kingma and Welling, 2013) to use $p_\theta(x|z)$, $p_\phi(z|x)$ and $\pi_\beta(z)$ with shorthands p_θ, q_ϕ and π_β to denote the decoder, the encoder and the prior respectively. To begin with, we impose the following Lipschitz condition for bounding the Rademacher complexity associated with $\overline{G}^*$, which is a common regularity condition in M-estimation problem (Vaart, 1998).

Condition A For $\mathcal{F}_{dd} = \{p_\theta(x|z) \,|\, \theta \in \Theta_\theta \subseteq \mathbb{R}^{d_\theta}\}$, $\mathcal{F}_{ed} = \{q_\phi(z|x) \,|\, \phi \in \Theta_\phi \subseteq \mathbb{R}^{d_\phi}\}$ and $\mathcal{F}_{prior} = \{\pi_\beta(z) \,|\, \beta \in \Theta_\beta \subseteq \mathbb{R}^{d_\beta}\}$, there exist some constants (a_0, a_1) such that for any $\theta, \theta' \in \Theta_\theta$, $\phi, \phi' \in \Theta_\phi$, $\beta, \beta' \in \Theta_\beta$ and $x \in \mathcal{X}$,

$$\|\theta\|_\infty + \|\phi\|_\infty + \|\beta\|_\infty \le a_0,$$
$$\big|m(p_\theta, q_\phi, \pi_\beta, x) - m(p_{\theta'}, q_{\phi'}, \pi_{\beta'}, x)\big| \le b(x)\|(\theta, \phi, \beta) - (\theta', \phi', \beta')\|_2,$$

with $\mathbb{E}_{p_{\mathcal{D}}(x)}\big[b^2(x)\big] \le a_1$, where $m(p_\theta, q_\phi, \pi_\beta, x)$ is the loss function for single data point defined in equation (1).

We are then ready to state the following theorem that provides an oracle inequality for the EBVAE estimators with parametric models.

Theorem 3 *Consider the EBVAE estimator $p_{\hat{\theta}}$, $q_{\hat{\phi}}$ and $\pi_{\hat{\beta}}$ defined in (2), and let $d^* = d_\theta + d_\phi + d_\beta$. If Assumption A and Condition A hold, then there exist some constants (c_0, c_1, c_2) that only depend* *on (α, a_0, a_1) so that it holds with probability at least $1 - c_0 \exp\big\{ - c_1 \, (d^* \log n)^{\min\{\alpha,1\}} \big\}$ that,*

$$\mathbb{E}_{p_{\mathcal{D}}(x)}\big[m(p_{\hat\theta}, q_{\hat\phi}, \pi_{\hat\beta}, x)\big] \le \inf_{\gamma>0} \Big\{(1+\gamma) \min_{\theta\in\Theta_\theta, \phi\in\Theta_\phi, \beta\in\Theta_\beta} \mathbb{E}_{p_{\mathcal{D}}(x)}\big[m(p_\theta, q_\phi, \pi_\beta, x)\big] + c_2\Big(1+\frac{1}{\gamma}\Big)\frac{Dd^*}{n}\log(nd^*)\log^{\frac{1}{\alpha}} n\Big\}.$$

The estimation error (second term) of Theorem 3 scales as $O(1/n)$ up to a logarithmic term, which matches the minimax optimal rate of parametric density estimation (Rigollet and Hütter, 2015; Silverman, 1986). The approximation error term of the risk bound in Theorem 3 is zero if the model is well-specified, that is, there exist some $p_{\theta^*} \in \mathcal{F}_{dd}$, $q_{\phi^*} \in \mathcal{F}_{ed}$ and $\pi_{\beta^*} \in \mathcal{F}_{prior}$, such that $p_{\mathcal{D}}(x) = \int p_{\theta^*}(x|z)\pi_{\beta^*}(z)dz$ and $q_{\phi^*}(z|x)$ is the posterior density with likelihood $p_{\theta^*}(x|z)$ and prior $\pi_{\beta^*}(z)$. Moreover, enriching the prior distribution family $\mathcal{F}_{prior}$ via hyperparameters may greatly reduce the approximation error term when $\mathcal{F}_{dd}$ and $\mathcal{F}_{ed}$ have limited capacities. Conversely, the estimation error is positively correlated with the number of parameters d^*. Suitable choices of $\mathcal{F}_{dd}$, $\mathcal{F}_{ed}$ and $\mathcal{F}_{prior}$ should minimize the risk upper bound, i.e., the approximation error and estimation error are balanced. In fact, Tomczak and Welling (2017) empirically shows that when the "Vamp prior" $\frac{1}{K}\sum_{k=1}^{K}\mathcal{N}(\mu_\phi(\mathbf{u}_k), \mathrm{diag}(\sigma^2_\phi(\mathbf{u}_k))$ is used as the parametric family of the prior, a most suitable choice of K is 500, either decreasing it or increasing it will result in significant deterioration of the performance (c.f. Appendix A for some numerical results). Theorem 3 provides a theoretically explanation to this phenomenon. When K is small, the first term (approximation error) in the risk bound dominates and when K is large, the second term (estimation error) dominates.

It has been shown that the prior which minimizes (2) is given by the corresponding aggregated posterior $\frac{1}{n}\sum_{i=1}^{n} q_{\hat\phi}(z|x_i)$ (Tomczak and Welling, 2017). The next corollary offer theoretical guarantees to methods that parameterize the prior for approximating the aggregated posterior (Tomczak and Welling, 2017; Bauer and Mnih, 2018) via giving an upper bound to the total variation distance between the target distribution and the distribution generated from a latent space model with prior being the aggregated posterior and conditional distribution being the fitted decoder.

Corollary 4 *Consider the EBVAE estimator $p_{\hat\theta}$, $q_{\hat\phi}$ and $\pi_{\hat\beta}$ defined in (2). Let $d^* = d_\theta + d_\phi + d_\beta$. If Assumption A and Condition A hold, and for any $z \in \mathcal{Z}$, $\|z\|_2 \le a_2$, $x \in \mathcal{X}$, $(\phi, \phi') \in \Theta_\phi$ and $(z, z') \in \mathcal{Z}$, the support of z under $q_\phi(z|x)$ is contained in $\mathcal{Z}$, and $|q_\phi(z|x) - q_{\phi'}(z'|x)| \le a_3(\|\phi - \phi'\|_2 + \|z - z'\|_2)$, then for some constants (c_0, c_1, c_2, c_3) only dependent of $(d_z, \alpha, a_0, a_1, a_2, a_3)$, such that it holds with probability at least $1 - c_0 \exp\big\{ - c_1 \, (\log n)^{\min\{\alpha,1\}} \big\}$ that*

$$D^2_{\mathrm{TV}}\Big(p_{\mathcal{D}}(\cdot), \int_{\mathcal{Z}} \Big(\frac{1}{n}\sum_{i=1}^{n} q_{\hat\phi}(z|x_i)\Big) p_{\hat\theta}(\cdot|z)dz\Big) \le c_2 \min_{\substack{\theta\in\Theta_\theta \\ \phi\in\Theta_\phi, \beta\in\Theta_\beta}} \mathbb{E}_{p_{\mathcal{D}}(x)}\big[m(p_\theta, q_\phi, \pi_\beta, x)\big] + c_3\frac{Dd^*}{n}\log(nd^*)\log^{\frac{1}{\alpha}} n.$$

Remark 5 *Here we state the risk bound in terms of total variation distance since the total variation distance is a metric satisfying the triangle inequality. Corollary 4 is proved by the triangle inequality and the fact that the aggregated posterior is close to the fitted prior with high probability.*

4.2. Gaussian Encoder and Decoder

In this subsection, we study the theoretical properties of the commonly used Gaussian encoder and decoder (Kingma and Welling, 2013; Doersch, 2016). Same as Section 4.1, we use $p_\theta(x|z)$ (p_θ)

and $q_\phi(z|x)$ (q_ϕ) to denote the decoder and encoder. We consider $p_\theta(x|z) = \mathcal{N}(G_{\theta_1}(z), \sigma^2 I_{d_x})$ and $q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \Sigma_\phi(x))$, where $G_{\theta_1}(z)$, $\mu_\phi(x)$ and $\Sigma_\phi(x)$ are functions parametrized by θ_1 and ϕ, and σ is a unknown parameter jointly trained with others. Adopting the Gaussian encoder family for $q_\phi(z|x)$ makes the optimization problem (2) in EBVAE computationally simpler. Unfortunately, even if we assume high capacity for $\mu_\phi(x)$ and $\Sigma_\phi(x)$, the approximation error from $\mathbb{E}_{p_\mathcal{D}(x)}\left[D_{\mathrm{KL}}\left(q_\phi(\cdot|x)\,\middle|\middle|\,\frac{p_\theta(x|\cdot)\pi_\beta(\cdot)}{\int p_\theta(x|z)\pi_\beta(z)dz}\right)\right]$ is nonvanishing, since the posterior is not necessarily Gaussian. However, if we assume the true data X to be generated by some low dimensional latent variable Z, with a deterministic and invertible generative function $G_\mathcal{D}(z)$, plus a random Gaussian error vector with mean 0 and covariance matrix $\sigma^{*2} I_{d_x}$ where σ^* is small enough, i.e., using $T_{\#}\mu$ to denote the image measure (or push-forward) of μ by T and $\mu * \nu$ to denote the convolution of μ and ν, so that the model of X can be expressed as $(G_{\mathcal{D}\#}\pi_\mathcal{D}) * \mathcal{N}(0, \sigma^{*2} I_{d_x})$, then Z becomes nearly "deterministic" given X and the approximation error vanishes, which is consistent with the finding in Doersch (2016). We then state the our conditions on the approximation family $\mathcal{F}_{prior}$ and assumptions on the true model $p_\mathcal{D}$. For a vector-valued function $f(x)$, we use $\|f(x)\|_p$ to denote its vector ℓ_p norm at input x.

Condition B The family of prior $\mathcal{F}_{prior} = \left\{\pi_\beta(z) \,|\, \beta \in \Theta_\beta \subseteq \mathbb{R}^{d_\beta}\right\}$ has a compact parameter space Θ_β. In addition, there exist some constants (b_2, b_3) such that for any $\beta, \beta' \in \Theta_\beta$ and $z \in \mathbb{R}^{d_z}$, $\|\beta\|_2 \leq b_2$, $|\log \pi_\beta(0)| \leq b_2$, $\|\nabla_z \log \pi_\beta(z)\|_2 \leq b_2\left(\|z\|_2 + 1\right)$ and $|\log \pi_\beta(z) - \log \pi_{\beta'}(z)| \leq b(z)\,\|\beta - \beta'\|_2$ with $\|b(z)\|_2 \leq b_2\left(\|z\|_2^{b_3} + 1\right)$.

Assumption B Assume the followings:
B.1: The data distribution $p_\mathcal{D} = (G_{\mathcal{D}\#}\pi_\mathcal{D}) * \mathcal{N}(0, \sigma^{*2} I_{d_x})$ $(\sigma_1 \leq \sigma^* \leq \frac{1}{2e})$, where $\pi_\mathcal{D}(z)$ is a probability density function (w.r.t. the Lebesgue measure on $\mathbb{R}^{d_z}$) that belongs to $\mathcal{F}_{prior}$. For a random variable Z with probability density $\pi_\mathcal{D}$, it holds that $\|\sum_{i=1}^{d_z}(Z_i)^2\|_{\psi_1} \leq b_5$ with some constant $b_5 > 0$. Moreover, $\forall\, z \in \mathbb{R}^{d_z}$, $\nabla \pi_\mathcal{D}(z)$ exists and $\|\nabla \pi_\mathcal{D}(z)\|_2 \leq b_5$.
B.2: There exists an integer $k \geq 2$, so that $G_\mathcal{D}(z) : \mathbb{R}^{d_z} \mapsto \mathbb{R}^{d_x} (d_z \leq d_x)$ is a $\mathbf{C}^k$ map, and there exists a $\mathbf{C}^k$ map $Q_\mathcal{D}(x) : \mathbb{R}^{d_x} \mapsto \mathbb{R}^{d_z}$ such that $\forall\, z \in \mathbb{R}^{d_z}$, $Q_\mathcal{D} \circ G_\mathcal{D}(z) = z$. Also, there exist some constants (α, b_6) where $0 < \alpha \leq 2$, such that for any $1 \leq i \leq d_x$, $1 \leq j \leq d_z$, $z \in \mathbb{R}^{d_z}$ and $x \in \mathbb{R}^{d_x}$, it holds that $\sum_{|\gamma| \leq k} |D^\gamma G_{\mathcal{D},i}(z)| \leq b_6(\|z\|_2^{\frac{2}{\alpha}} + 1)$ and $\sum_{|\gamma| \leq k} |D^\gamma Q_{\mathcal{D},j}(x)| \leq b_6(\|x\|_2^{\frac{2}{\alpha}} + 1)$, where $G_{\mathcal{D},i}(z)$ and $Q_{\mathcal{D},j}(x)$ are the elements of the ith and the jth dimension of $G_\mathcal{D}(z)$ and $Q_\mathcal{D}(x)$, γ is a multi-index $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_d) \in \mathbb{N}_0^d$ and D^γ denotes the mixed partial derivative operator.

Remark 6 *Condition B requires the priors in $\mathcal{F}_{prior}$ to behave like (mixture of) Gaussian distributions. Assumption B requires the latent variable to have a density function that is sub-Gaussian and sufficiently smooth, and demands some regularity conditions on the map $G_\mathcal{D}$. It states that $G_\mathcal{D}(z)$ is a $\mathbf{C}^k$ map with a $\mathbf{C}^k$ inverse $Q_\mathcal{D}(x)$ and the mixed partial derivatives of $G_\mathcal{D}(z)$ and $Q_\mathcal{D}(x)$ are upper bounded by polynomial functions of z and x with order up to $\frac{2}{\alpha}$ respectively. The invertibility of $G_\mathcal{D}(z)$ is also assumed in Doersch (2016). The assumptions on the mixed partial derivatives of $G_\mathcal{D}$ and $Q_\mathcal{D}$ ensure that $G_\mathcal{D}$ and $Q_\mathcal{D}$ can be well approximated by ReLU neural networks.*

Theorem 7 *Choose $\sigma_1 \in (0, \frac{1}{2e}]$, and consider $\mathcal{F}_{dd} = \left\{p_\theta(x|z) = \mathcal{N}(G_{\theta_1}(z), \sigma^2 I_{d_x}) \,|\, G_{\theta_1}(z) \in \mathcal{F}_G,\, \sigma \in [\sigma_1, 1]\right\}$, $\mathcal{F}_{ed} = \left\{q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \Sigma_\phi(x)) \,|\, (\mu_\phi(x), \Sigma_\phi(x)) \in \mathcal{F}_{\mu,\Sigma}\right\}$ and $\mathcal{F}_{prior} = \left\{\pi_\beta(z) \,|\, \beta \in \Theta_\beta \subseteq \mathbb{R}^{d_\beta}\right\}$. If Condition B holds for $\mathcal{F}_{prior}$, then there exists a choice of $\mathcal{F}_G$ and $\mathcal{F}_{\mu,\Sigma}$ so that for any target distribution $p_\mathcal{D} = (G_{\mathcal{D}\#}\pi_\mathcal{D}) * \mathcal{N}(0, \sigma^{*2} I_{d_x})$ satisfying Assumption B, the EBVAE estimator $p_{\hat{\theta}}$, $q_{\hat{\phi}}$ and $\pi_{\hat{\beta}}$ defined in (2) satisfies that there exist some constants (c, c_1, c_2)*

that only depend on (d_z, d_x) and $(\alpha, k, b_2, b_3, b_5, b_6)$ in Assumption B and Condition B, such that it holds with probability at least $1 - n^{-c}$ that,

$$\mathbb{E}_{p_{\mathcal{D}}(x)}\big[m(p_{\hat{\theta}}, q_{\hat{\phi}}, \pi_{\hat{\beta}}, x)\big] \leq c_1 \log^{\tilde{\alpha}_1} \frac{1}{\sigma^*}\, \sigma^{*2} + c_2 \log^{\frac{2}{\alpha}} n \, \log^{\tilde{\alpha}_2} \frac{1}{\sigma_1} \, \frac{d_\beta + {\sigma_1}^{-\frac{2d_z}{k}}}{n\sigma_1^2}, \tag{3}$$

where $\tilde{\alpha}_1 = \frac{28+10\alpha+3\alpha^2}{\alpha^2}$ and $\tilde{\alpha}_2 = \frac{2}{\alpha} + \frac{d_z}{\alpha(k-1)} + \frac{d_z}{2} + 6$. Recall that σ^* is the standard deviation of each component of the data X given the latent variable Z with $\sigma^* \in [\sigma_1, \frac{1}{2e}]$.

Remark 8 *In the decoder, we use a Gaussian distribution to approximate the posterior, so that $p_{\mathcal{D}}(z|x) = \frac{p_{\mathcal{D}}(x|z)\pi_{\mathcal{D}}(z)}{p_{\mathcal{D}}(x)}$ can be well approximated by a Gaussian either. When σ^* is small, the invertibility assumption on $G_{\mathcal{D}}$ guarantees that the highest density region of $p_{\mathcal{D}}(z|x)$ is concentrated around $Q_{\mathcal{D}}(x)$. By applying the first order Taylor expansion of $G_{\mathcal{D},j}(z)$ at $z = Q_{\mathcal{D}}(x)$, we have that $p_{\mathcal{D}}(z|x) \propto \pi_{\mathcal{D}}(z)\exp(-\sum_{j=1}^{d_x}(x_j - G_{\mathcal{D},j}(z))^2/2\sigma^{*2})$ is approximately a Gaussian distribution with mean $Q_{\mathcal{D}}(x) + \Sigma_{\mathcal{D}}(x)\nabla G_{\mathcal{D}}(z)^T|_{z=Q_{\mathcal{D}}(x)}(x - G_{\mathcal{D}}(Q_{\mathcal{D}}(x)))$ and covariance matrix $\sigma^{*2}\Sigma_{\mathcal{D}}(x)$ with $\Sigma_{\mathcal{D}}(x) = (\nabla G_{\mathcal{D}}(z)^T|_{z=Q_{\mathcal{D}}(x)}\nabla G_{\mathcal{D}}(z)|_{z=Q_{\mathcal{D}}(x)})^{-1}$. $\mathcal{F}_G$ and $\mathcal{F}_{\mu,\Sigma}$ are realized through feedforward ReLU neural networks with sizes depend on σ_1 in the proof of Theorem 7 to achieve the rate in equation (3). Since the data X lie approximately on a d_z-dimensional manifold, the result does not suffer from the "the curse of dimensionality", i.e., the dimension of X (d_x) does not occur in the exponent of the approximation error of a ReLU neural network with given size for approximating functions of X with certain smoothness constraints, see Appendix B.3.1 for further details.*

The constants (c_1, c_2) has an exponential dependence on d_z and a polynomial dependence on d_x, scale as $(c_3)^{d_z}$ and $(d_x)^{c_4}$ for positive constants c_3, c_4 independent of (n, d_x, d_z). The occurrence of σ^{*2} in the above theorem is from the fact that $p_{\mathcal{D}}(z|x)$ is not necessarily a Gaussian distribution, which theoretically explains the reason why VAE models tend to produce unrealistic, blurry samples when applied to complex datasets of natural images (Dosovitskiy and Brox, 2016). In particular, when $\sigma_1 \asymp \sigma^*$, regardless of the logarithmic term, the risk bound in above thorem scales as $\sigma^{*2} + \frac{1}{n\sigma^{*2}}(d_\beta + {\sigma^*}^{-\frac{2d_z}{k}})$, where the first term corresponds to an upper bound for the approximation error and the second term correspond to an excess risk bound. In particular, if the noise level σ^* decreases with the sample size at the rate $\sigma^* \asymp n^{-k_1}$, where $0 < k_1 < \frac{k}{2(k+d_z)}$, the EBVAE estimator will be consistent relative to the KL risk function, which give theoretical guarantee to EBVAE estimators and theoretically explains the phenomenon that Vanilla VAE still achieves good performance for some simple dataset (e.g. MNIST dataset) even if the encoder model is misspecified as a simple gaussian model. Here, we emphasize that we need k_1 to be upper bounded since d_z can be smaller than d_x and the KL divergence may diverge to infinity if the supports of the two distributions are not the same. In addition, for fixed σ^* and n, the above bound depends on the number of parameters in the prior family and the smoothness of $G_{\mathcal{D}}$. When a pre-specified data independent prior is used, we may need a highly complicated $G_{\mathcal{D}}$ to first map the chosen prior to a highly irregular distribution of latent variables, which increases the capacity demand for $G_{\mathcal{D}}$.

In practice, the covariance matrix of the encoder model $\Sigma_\phi(x)$ is often chosen to be diagonal and characterized by a variance vector (Kingma and Welling, 2013; Tomczak and Welling, 2017). However, Remark 8, when σ^* converges to 0, the posterior of the latent variable converges to a Gaussian distribution with covariance matrix $\sigma^{*2}(\nabla G_{\mathcal{D}}(z)^T|_{z=Q_{\mathcal{D}}(x)}\nabla G_{\mathcal{D}}(z)|_{z=Q_{\mathcal{D}}(x)})^{-1}$, which may not be diagonal. Misspecifying the off-diagonal elements of the covariance matrix introduces extra approximation errors and thus deteriorates the performance of the EBAVE estimator. In order to achieve the smallest risk, we should model the full $d_z \times d_z$ covariance matrix instead of through a variance vector. A natural practical choice of $\Sigma_\phi(x)$ is $\tilde{\Sigma}_\phi(x)^T \tilde{\Sigma}_\phi(x) + \varepsilon^2 I_{d_z}$, where $\tilde{\Sigma}_\phi(x)$ is a $d_z \times d_z$ matrix modelled by a neural network and ε is small number to guarantee the positive definiteness of the covariance matrix. The theory we established for EBVAE estimators with Gaussian encoders and decoders also validates the importance of the variance parameter σ in the decoder family, which is often chosen as a predefined weighting factor depending on the target accuracy level for reconstructing. However, our theory suggests that misspecifying the conditional variance of the data will lead to a large approximation error. Consequently, the variance parameter of the decoder family should be jointly optimized instead of being prespecified. Moreover, if the decoder family is correctly specified, i.e. the conditional distribution of data is $\mathcal{N}(G_{\theta^*}(z), {\sigma^*}^2 I_{d_x})$, then the parameter σ should be constrained by a lower bound that is close to σ^* up to a multiplicative constant.

4.3. Nonparametric Models

The risk bound in the previous subsection demands us to consider more complicated encoder and decoder families to reduce the approximation error. Motivated by this, we consider nonparametric families in this subsection. We assume the data space $\mathcal{X}$ and the latent space $\mathcal{Z}$ are $[0,1]^{d_x}$ and $[0,1]^{d_z}$ respectively. To begin with, we consider the following densities on $\mathcal{X}$ characterized by an undirected graphical model (Markov network) (Koller and Friedman, 2009) with clique sizes being bounded by $\overline{p}$ as our decoder family: $\bar{\mathcal{F}}_{dd} = \big\{p_{dd}(x|z) : p(x|z) \propto \exp\big(\sum_{j=1}^{k_1} l_j(\overline{x}_j, z)\big) \,\big|\, l_j(\overline{x}_j, z) \in B_{r_1}^{\overline{\alpha}}([0,1]^{|\overline{x}_j|+d_z}),\ |\overline{x}_j| \le \overline{p}\big\}$, where $\overline{x}_j$ is a subvector of $x = (x_1, \cdots, x_{d_x})$ with $|\overline{x}_j|$ being its dimension. Similarly, we consider the following encoder family on $\mathcal{Z}$: $\bar{\mathcal{F}}_{ed} = \big\{q(z|x) : q(z|x) \propto \exp\big(\sum_{j=1}^{k_2} f_j(z, \overline{x}_j)\big) \,\big|\, f_j(z, \overline{x}_j) \in B_{r_2}^{\overline{\alpha}}([0,1]^{|\overline{x}_j|+d_z}),\ |\overline{x}_j| \le \overline{p}\big\}$. We then state our condition on the approximation families and assumption on the true model $p_{\mathcal{D}}$ for deriving the Lipschitzness of the loss function in (1).

Condition C $\mathcal{F}_{dd} \subseteq \bar{\mathcal{F}}_{dd}$ and $\mathcal{F}_{ed} \subseteq \bar{\mathcal{F}}_{ed}$. For the family of prior $\mathcal{F}_{prior} = \big\{\pi_\beta(z) \,|\, \beta \in \Theta_\beta\big\}$, Θ_β is a compact set so that for any $\beta, \beta' \in \Theta_\beta$ and $z \in \mathcal{Z}$, $|\log \pi_\beta(z) - \log \pi_{\beta'}(z)| \le c_1 \|\beta - \beta'\|_2$ with a constant c_1 and the support of π_β is contained in $\mathcal{Z}$. Moreover, we have $\sup_{\pi_\beta \in \mathcal{F}_{prior}} \sup_{z \in \mathcal{Z}} |\log \pi_\beta(z)| \le c_2$ with some positive constant c_2.

Assumption C There exists a positive constant c such that $\sup_{x \in \mathcal{X}} |\log p_{\mathcal{D}}(x)| \le c$.

For ease of notation, we define δ_n as: if $d_z + \overline{p} < 2\overline{\alpha}$, then $\delta_n = n^{-\frac{\overline{\alpha}}{2\overline{\alpha}+d_z+\overline{p}}}$; if $d_z + \overline{p} = 2\overline{\alpha}$, then $\delta_n = n^{-\frac{1}{4}}\sqrt{\log n}$; if $d_z + \overline{p} > 2\overline{\alpha}$, then $\delta_n = n^{-\frac{\overline{\alpha}}{2(d_z+\overline{p})}}$.

Theorem 9 *Consider the EBVAE estimator $\hat{p}$, $\hat{q}$ and $\pi_{\hat{\beta}}$ defined in (2). If Condition C and Assumption C hold, then for some constants (c_0, c_1, c_2) independent of n, it holds with probability at least $1 - c_0 \exp\left(-c_1 n \delta_n^2\right)$ that*

$$\mathbb{E}_{p_{\mathcal{D}}(x)}\big[m(\hat{p}, \hat{q}, \pi_{\hat{\beta}}, x)\big] \le \inf_{\gamma > 0} \left\{(1+\gamma) \min_{\substack{p \in \mathcal{F}_{dd}, q \in \mathcal{F}_{ed} \\ \pi_\beta \in \mathcal{F}_{prior}}} \mathbb{E}_{p_{\mathcal{D}}(x)}\big[m(p, q, \pi_\beta, x)\big] + c_2\Big(1 + \frac{1}{\gamma}\Big)\delta_n^2\right\}.$$

Remark 10 *If the target distribution $p_{\mathcal{D}}(x)$ can be expressed as $\int_{\mathcal{Z}} p_{\mathcal{D}}(x|z)\pi_{\mathcal{D}}(z)dz$ with some $p_{\mathcal{D}}(x|z)$ and $\pi_{\mathcal{D}}(z)$, where $p_{\mathcal{D}}(x|z) \in \bar{\mathcal{F}}_{dd}$, $\log \pi_{\mathcal{D}}(z) \in C_{r_1}^{\overline{\alpha}}([0,1]^{d_z})$ and $\pi_{\mathcal{D}}(z) \in \mathcal{F}_{prior}$, then by choosing (k_2, r_2) in $\bar{\mathcal{F}}_{ed}$ to be (k_1, cr_1) with some constants c and $(\mathcal{F}_{dd}, \mathcal{F}_{ed}) = (\bar{\mathcal{F}}_{dd}, \bar{\mathcal{F}}_{ed})$, the* *approximation error term in the above risk bound is zero. Moreover, when $\overline{p} \ll d_x$ (e.g. given the latent variable, the component of each dimension of the data is independent of each other), the additive structure in the encoder family and decoder family prevents the risk bound from suffering from "the curse of dimensionality".*

5. Discussion

In this paper, we consider variational autoencoders via empirical Bayes estimation, referred to as Empirical Bayes Variational Autoencoders (EBVAE), which is a general framework including popular VAE methods as special cases. Theoretically, we give a general statistical framework to analyze the convergence rate for learning densities using EBVAE. We develop novel oracle inequalities which quantitively capture impacts of prior families, encoder families, and decoder families on excess risks of the estimators arising from the EBVAE. The key idea in our proof comes from representing the EBVAE estimator as an M-estimator. Once making this connection, we can leverage the general theoretical machinery of M-estimation for obtaining a risk bound. Our theory gives sufficient conditions under which the EBVAE estimators are consistent in both parametric cases and nonparametric cases. In particular, we carefully analyze the estimator derived from EBVAE with Gaussian encoders and decoders, we show that it is consistent if the conditional variance of data given latent variables decreases with sample size under suitable rates. Our result highlights the importance of covariance matrices of encoders and decoders in obtaining a good statistical guarantee.

The risk bound we derived for the EBVAE estimators under Gaussian models does not apply to the case that the data is deterministic given latent variables, for the reason that the dimension of latent variables can be smaller than the dimension of data and the KL divergence may diverge to infinity if the supports of the two distributions are not the same. We suspect that this issue can be resolved by stating the risk bound in terms of some adversarial losses that is insensitive to small fluctuations compared with KL divergence (e.g. Wasserstein distance, Santambrogio, 2015); we leave this for future work. Moreover, the proposal of encoder and decoder families that yield consistency in more general cases without adding significant computational burden is another important topic of future research.

References

Radoslaw Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to markov chains. *Electron. J. Probab.*, 13:1000–1034, 2008. doi: 10.1214/EJP.v13-521. URL https://doi.org/10.1214/EJP.v13-521.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999. doi: 10.1017/CBO9780511624216.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 08 2005. doi: 10.1214/009053605000000282. URL https://doi.org/10.1214/009053605000000282.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019. URL http://jmlr.org/papers/v20/17-612.html.

Matthias Bauer and Andriy Mnih. Resampled priors for variational autoencoders, 2018.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2018.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Xi Chen, Diederik P. Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder, 2016.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Luc Devroye. Nonuniform random variate generation. *Handbooks in operations research and management science*, 13:83–121, 2006.

Carl Doersch. Tutorial on variational autoencoders, 2016.

Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks, 2016.

R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1999. doi: 10.1017/CBO9780511665622.

Lawrence C. Evans. *Partial differential equations*. American Mathematical Society, Providence, R.I., 2010.

Clark R. Givens and Rae Michael Shortt. A class of wasserstein metrics for probability distributions. *Michigan Math. J.*, 31(2):231–240, 1984. doi: 10.1307/mmj/1029003026. URL https://doi.org/10.1307/mmj/1029003026.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation, 2015.

Peter D. Grünwald and Nishant A. Mehta. Fast rates for general unbounded loss functions: From erm to generalized bayes. *Journal of Machine Learning Research*, 21(56):1–80, 2020. URL http://jmlr.org/papers/v21/18-488.html.

Victor Guillemin and Alan Pollack. *Differential topology*, volume 370. American Mathematical Soc., 2010.

Tony Jebara and Marina Meila. Machine learning: Discriminative and generative. *The Mathematical Intelligencer*, 28(1):67–69, 2006.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pages 4743–4751, 2016.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Tejas D. Kulkarni, Will Whitney, Pushmeet Kohli, and Joshua B. Tenenbaum. Deep convolutional inverse graphics network, 2015.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1991. ISBN 978-3-642-20212-4. doi: 10.1007/978-3-642-20212-4_8. URL https://doi.org/10.1007/978-3-642-20212-4_8.

Tengyuan Liang. How well generative adversarial networks learn distributions, 2018.

Pascal Massart. Concentration inequalities and model selection. 2007.

Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis*, 17(4):1248–1282, 2007.

John T Ormerod and Matt P Wand. Explaining variational approximations. *The American Statistician*, 64(2):140–153, 2010.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models, 2014.

Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 2015.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Marwin H. S. Segler, Thierry Kogej, Christian Tyrchan, and Mark P. Waller. Generating focussed molecule libraries for drug discovery with recurrent neural networks, 2017.

Simon J. Sheather. Density estimation. *Statist. Sci.*, 19(4):588–597, 11 2004. doi: 10.1214/088342304000000297. URL https://doi.org/10.1214/088342304000000297.

Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053, 12 1982. doi: 10.1214/aos/1176345969. URL https://doi.org/10.1214/aos/1176345969.

Jakub M. Tomczak and Max Welling. Vae with a vampprior, 2017.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio, 2016.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. ISBN 9781108415194. URL https://books.google.com/books?id=NDdqDwAAQBAJ.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders, 2016.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions, 2017.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94: 103–114, 2017.

Appendix

Notations: We adopt the notations in the manuscript, and further introduce the following additional notations for technical proof. We write $a \asymp b$ if $a \lesssim b$ and $a \gtrsim b$. $a = O(b)$ if $a \asymp b$. For a matrix $A \in \mathbb{R}^{m \times n}$, we use $\|A\|_F$ and $\|A\|_{op}$ to denote its Frobenius norm and operator norm respectively. When $m = n$, we use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote its minimal and maximal eigenvalues. Unless otherwise specified, for a matrix $A \in \mathbb{R}^{n \times n}$, $|A|$ denotes the determinant. $\mathbf{N}(G, d, \varepsilon)$ denotes the ε-covering number of G under metric d. $\mathbb{O}(n, d)$ denotes the set of $n \times d$ matrices U such that $U^T U = I_d$, $\mathbb{O}(d)$ denotes the set of $d \times d$ orthogonal matrices.

Appendix A. Numerical study

A.1. Set up

In the experiments we aim at: (1) verifying empirically whether the EBVAE outperform VAE, (2) investigate the influence of the choice of the prior family on the performance of data generation and (3) showing the validity of our theory. We carry out experiment using two models: “Vanilla VAE” (Kingma and Welling, 2013) and “VampPrior” (VP) (Tomczak and Welling, 2017). The “Vanilla VAE” model use a predefined isotropic gaussian prior. The “VampPrior” model consider prior and amortized inference distribution

$$
\begin{aligned}
\pi_{\phi,\mathbf{u}}(z) &= \frac{1}{K} \sum_{k=1}^{K} N(\mu_\phi(\mathbf{u}_k), \mathrm{diag}(\sigma_\phi^2(\mathbf{u}_k))) \\
q_\phi(z|x) &= N(\mu_\phi(x), \mathrm{diag}(\sigma_\phi^2(x)),
\end{aligned}
$$

where K is the number of pseudo-inputs, and $\mathbf{u}_k$ is a D-dimensional vector we refer to as a pseudo-input. We then apply the two model to the dynamic MNIST dataset. In the experiments we modeled all distributions using MLPs with two hidden layers of 300 hidden units. The dimension of the latent variable is choose to be 40, and for “VampPrior” model, we choose $K = (1, 10, 100, 300, 400, 500, 600)$.

A.2. Results

We quantitatively evaluate the three method using the test marginal log-likelihood (LL) estimated using the Importance Sampling (Burda et al., 2015). The LL values and the digits generated by the two models is given in Figure 1 and Figure 2.

We can see that the supremacy of EBVAE is visible not only in LL values but in image generations as well. According to our results on the parametric rate, the estimation error includes two terms: Approximation error and the dimension of parameters. The “Vanilla VAE” model use a predefined prior, so $d_\beta = 0$, but the approximation error is larger than VP model, which result in a poor performance. Also, for the VP model, when the number of pseudo-inputs is large enough, increasing the number of pseudo-inputs will actually result in drop of the performance, which is consistent with our bound, since when the parameter space of prior is too large, the $\frac{d_\theta + d_\phi + d_\beta}{n}$ term will dominate.

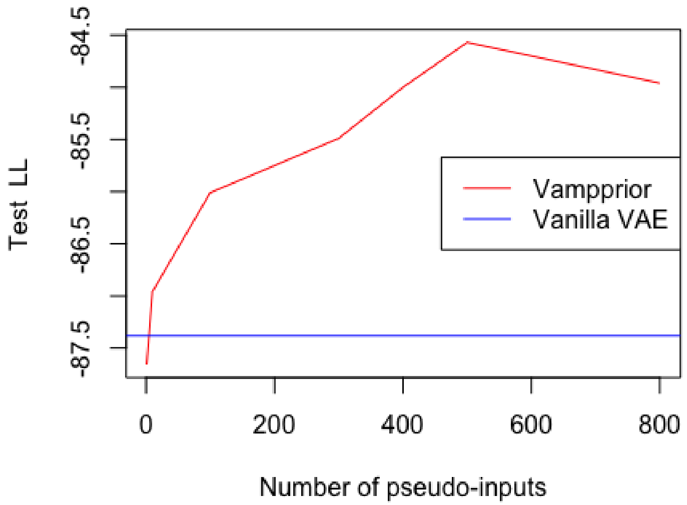
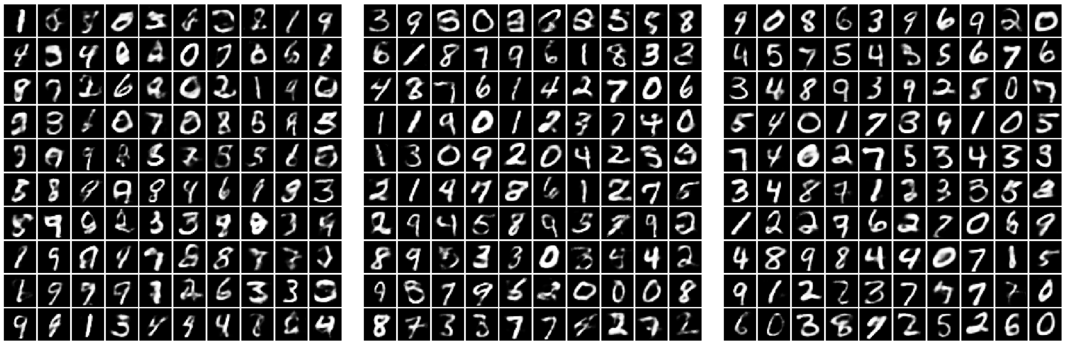


Figure 1: Test LL between different models



(*a*) Vanilla VAE (*b*) VP-K=300 (*c*) VP-K=500

Figure 2: Digits generated by different models

Appendix B. Proof of Main Results

B.1. Main Theoretical Results

Define

$$M_n(p, q, \pi) = \frac{1}{n}\sum_{i=1}^{n} m(p, q, \pi, x_i);$$
$$M^*(p, q, \pi) = \mathbb{E}_{p_{\mathcal{D}}(x)} m(p, q, \pi, x),$$

where $m(p, q, \pi, x)$ is defined in equation (1). We also use the notation $p(x)$ to denote the marginal $\int p(x|z)\pi(z)dz$ and $p(z|x)$ to denote the posterior $\frac{p(x|z)\pi(z)}{\int p(x|z)\pi(z)dz}$ when no ambiguity may arise. We begin the proof of Theorem 1 with the following two lemmas for controlling the supreme of an unbounded empirical process.

Lemma 11 *Suppose Assumption A holds, then there exist some constants (c_1, c_2) only depend on α, such that for any $p(x|z), p'(x|z) \in \mathcal{F}_{dd}$, $q(z|x), q'(z|x) \in \mathcal{F}_{ed}$ and $\pi(z), \pi'(z) \in \mathcal{F}_{prior}$,*

$$
\begin{aligned}
&\mathbb{E}_{p_{\mathcal{D}}(x)}\left[\left(\log\frac{p_{\mathcal{D}}(x)}{p(x)}\right)^2\right] \le c_1\left((D\log^{\frac{1}{\alpha}} n) D_{\mathrm{KL}}(p_{\mathcal{D}}(\cdot)||p(\cdot)) + \frac{D^2\log^{\frac{2}{\alpha}} n}{n}\right);\\
&\mathbb{E}_{p_{\mathcal{D}}(x)}\left[(D_{\mathrm{KL}}(q(\cdot|x)||p(\cdot|x)) - D_{\mathrm{KL}}(q'(\cdot|x)||p'(\cdot|x)))^2\right] \le\\
&c_2\left((D\log^{\frac{1}{\alpha}} n)\mathbb{E}_{p_{\mathcal{D}}(x)}\left[\left(\sqrt{D_{\mathrm{KL}}(q(\cdot|x)||p(\cdot|x))} - \sqrt{D_{\mathrm{KL}}(q'(\cdot|x)||p'(\cdot|x))}\right)^2\right] + \frac{D^2\log^{\frac{2}{\alpha}} n}{n}\right).
\end{aligned}
$$

Lemma 12 *Under Assumption A, if there exist $(p^*, q^*, \pi^*) \in \Psi^*$ and δ_n satisfying conditions defined in Theorem 1, then there exist some constants (c_0, c_1, c_2) that only depend on α , such that it holds with probability larger than $1 - c_0\exp\left(-c_1\left(\frac{n\delta_n^2}{D^2\log^{\frac{2}{\alpha}} n}\right)^{\min\{\alpha,1\}}\right)$ that,*

$$
\forall p(x|z) \in \mathcal{F}_{dd}, q(z|x) \in \mathcal{F}_{ed}, \pi(z) \in \mathcal{F}_{prior},
$$

$$
\begin{aligned}
&\frac{|M_n(p,q,\pi) - M_n(p^*,q^*,\pi^*) - M^*(p,q,\pi) + M^*(p^*,q^*,\pi^*)|}{\delta_n + \|m(p,q,\pi,\cdot) - m(p^*,q^*,\pi,\cdot)\|_2}\\
&\le c_2\delta_n/(D\log^{\frac{1}{\alpha}} n).
\end{aligned}
$$

B.1.1. PROOF OF THEOREM 1

$$
\begin{aligned}
&\|m(\hat{p},\hat{q},\hat{\pi},\cdot) - m(p^*,q^*,\pi^*,\cdot)\|_2^2\\
&= \mathbb{E}_{p_{\mathcal{D}}(x)}\left[\left(-\log\frac{\hat{p}(x)}{p_{\mathcal{D}}(x)} + D_{\mathrm{KL}}(\hat{q}(\cdot|x)||\hat{p}(\cdot|x)) + \log\frac{p^*(x)}{p_{\mathcal{D}}(x)} - D_{\mathrm{KL}}(q^*(\cdot|x)||p^*(\cdot|x))\right)^2\right]\\
&\le 4\mathbb{E}_{p_{\mathcal{D}}(x)}\left[\left(\log\frac{\hat{p}(x)}{p_{\mathcal{D}}(x)}\right)^2\right] + 4\mathbb{E}_{p_{\mathcal{D}}(x)}\left[\left(\log\frac{p^*(x)}{p_{\mathcal{D}}(x)}\right)^2\right]\\
&+ 2\mathbb{E}_{p_{\mathcal{D}}(x)}\left[(D_{\mathrm{KL}}(\hat{q}(\cdot|x)||\hat{p}(\cdot|x)) - D_{\mathrm{KL}}(q^*(\cdot|x)||p^*(\cdot|x)))^2\right].
\end{aligned}
$$

Therefore by Lemma 11 and $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}\,(a, b \ge 0)$, there exists a constant $C_0 = C_0(\alpha)$ such that,

$$
\begin{aligned}
&\|m(\hat{p},\hat{q},\hat{\pi},\cdot) - m(p^*,q^*,\pi^*,\cdot)\|_2\\
&\le C_0\left(\log^{\frac{1}{2\alpha}} n\sqrt{D}\left(\sqrt{M^*(\hat{p},\hat{q},\hat{\pi})} + \sqrt{\min_{p\in\mathcal{F}_{dd},q\in\mathcal{F}_{ed},\pi\in\mathcal{F}_{prior}} M^*(p,q,\pi)}\right) + \frac{D\log^{\frac{1}{\alpha}} n}{\sqrt{n}}\right).
\end{aligned}
$$

Therefore by Lemma 12 and the fact that $M_n(\hat{p},\hat{q},\hat{\pi}) \leq M_n(p^*,q^*,\pi^*)$, under the high probability set of Lemma 12, there exists a constant $C = C(\alpha)$ such that,

$$\begin{aligned}
&M^*(\hat{p},\hat{q},\hat{\pi}) - \min_{p\in\mathcal{F}_{dd},q\in\mathcal{F}_{ed},\pi\in\mathcal{F}_{prior}} M^*(p,q,\pi)\\
&= M^*(\hat{p},\hat{q},\hat{\pi}) - M^*(p^*,q^*,\pi^*)\\
&\leq |M_n(\hat{p},\hat{q},\hat{\pi}) - M_n(p^*,q^*,\pi^*) - M^*(\hat{p},\hat{q},\hat{\pi}) + M^*(p^*,q^*,\pi^*)|\\
&\leq C\delta_n\left(\frac{\log^{-\frac{1}{2\alpha}} n}{\sqrt{D}}\left(\sqrt{M^*(\hat{p},\hat{q},\hat{\pi})} + \sqrt{\min_{p\in\mathcal{F}_{dd},q\in\mathcal{F}_{ed},\pi\in\mathcal{F}_{prior}} M^*(p,q,\pi)}\right) + \frac{1}{\sqrt{n}} + \delta_n\frac{\log^{-\frac{1}{\alpha}} n}{D}\right).
\end{aligned}$$

By the fact that $\left(\frac{n\delta_n^2}{D^2\log^{\frac{2}{\alpha}} n}\right)^{\min\{\alpha,1\}} \gtrsim \log(\log\frac{D}{\delta_n})$ and the inequalities that $2\sqrt{ab} \leq \gamma a + \frac{1}{\gamma}b$ $(a,b,\gamma > 0)$, there exist some constant (c_0,c_1,c_2) that only depend on α such that it holds with probability larger than $1 - c_0\exp\left(-c_1\left(\frac{n\delta_n^2}{D^2\log^{\frac{2}{\alpha}} n}\right)^{\min\{\alpha,1\}}\right)$ that,

$$\begin{aligned}
&D_{\mathrm{KL}}(p_{\mathcal{D}}(\cdot)||\hat{p}(\cdot)) + \mathbb{E}_{p_{\mathcal{D}}(x)}\left[D_{\mathrm{KL}}(\hat{q}(\cdot|x)||\hat{p}(\cdot|x))\right] = M^*(\hat{p},\hat{q},\hat{\pi})\\
&\leq \min_{\gamma>0}\Big((1+\gamma)\min_{p\in\mathcal{F}_{dd},q\in\mathcal{F}_{ed},\pi\in\mathcal{F}_{prior}}\big(D_{\mathrm{KL}}(p_{\mathcal{D}}(\cdot)||p(\cdot)) + \mathbb{E}_{p_{\mathcal{D}}(x)}\left[D_{\mathrm{KL}}(q(\cdot|x)||p(\cdot|x))\right]\big)\\
&+ c_2\big(1+\frac{1}{\gamma}\big)\delta_n^2\frac{\log^{-\frac{1}{\alpha}} n}{D}\Big).
\end{aligned}$$

B.1.2. PROOF OF LEMMA 12

For $G^* = \{g(x) = m(p,q,\pi,x) - m(p^*,q^*,\pi^*,x)\,|\,p\in\mathcal{F}_{dd}, q\in\mathcal{F}_{ed}, \pi\in\mathcal{F}_{prior}\}$, it holds that

$$\sup_{g\in G^*}|g(x)| \leq 2\sup_{p\in\mathcal{F}_{dd},q\in\mathcal{F}_{ed},\pi\in\mathcal{F}_{prior}}\left(\left|\log\frac{p(x)}{p_{\mathcal{D}}(x)}\right| + D_{\mathrm{KL}}(q(\cdot|x)||p(\cdot|x))\right).$$

Therefore $\left\|\sup_{g\in G^*}|g(x)|\right\|_{\psi_\alpha} < +\infty$ and $\left\|\sup_{g\in G^*}\left|g(x) - \mathbb{E}_{p_{\mathcal{D}}(x)}g(x)\right|\right\|_{\psi_\alpha} < +\infty$. Define

$$Z_n(\delta, G^*) = \sup_{\substack{g\in G^*\\ \|g\|_2\leq\delta}}\left|\frac{1}{n}\sum_{i=1}^n g(x_i) - \mathbb{E}_{p_{\mathcal{D}}(x)}g(x)\right|.$$

Since $\frac{1}{n}\sup_{\substack{g\in G^*\\ \|g\|_2\leq\delta}}\sum_{i=1}^n var(g(x_i)) \leq \delta^2$, by the tail inequality for suprema of unbounded empirical processes (see for example, Theorem 4 and Lemma 1 of Adamczak (2008)), it holds that

$$\begin{aligned}
&P(Z_n(\delta,G^*) \geq (1+\eta)\mathbb{E}_{p_{\mathcal{D}}(x)}(Z_n(\delta,G^*)) + s^2)\\
&\leq c_0(\eta,\alpha)\exp\left(-c_1(\eta,\alpha)\min\left\{\frac{ns^4}{\delta^2}, \frac{n^\alpha s^{2\alpha}}{D^\alpha\log n}, \frac{ns^2}{D\log^{\frac{1}{\alpha}} n}\right\}\right).
\end{aligned} \tag{4}$$

Using the standard symmetrization (see, for example, Proposition 4.11 of Wainwright (2019)), we can get

$$\begin{aligned}\mathbb{E}_{p_{\mathcal{D}}(x)}\left[Z_n(\delta, G^*)\right] &\leq \mathbb{E}_{p_{\mathcal{D}}(x)}\mathbb{E}_{\varepsilon}\left[\sup_{\substack{g\in G^* \\ \|g\|_2\leq\delta}}\left|\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i g(x_i)\right|\right] \\ &= 2\overline{R}_n(\delta, G^*) \leq 2\overline{R}_n(\delta, \overline{G}^*),\end{aligned}$$

where recall that $\overline{G}^* = \{ag | a \in (0,1], g \in G^*\}$. Therefore by $\overline{R}_n(\delta_n, \overline{G}^*) \leq \delta_n^2/(D\log^{\frac{1}{\alpha}} n)$, it holds that

$$\begin{aligned}\forall r \geq \delta_n, \quad \mathbb{E}_{p_{\mathcal{D}}(x)}\left[Z_n(r, G^*)\right] &\leq 2\overline{R}_n(r, \overline{G}^*) \\ &= 2\mathbb{E}_{p_{\mathcal{D}}(x)}\mathbb{E}_{\varepsilon}\left[\sup_{\substack{g\in\overline{G}^* \\ \|\frac{\delta_n}{r}g\|_2\leq\delta_n}}\frac{r}{\delta_n}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\frac{\delta_n}{r}g(x_i)\right|\right] \\ &\leq 2\frac{r}{\delta_n}\overline{R}_n(\delta_n, \overline{G}^*) \\ &\leq 2\frac{r\delta_n}{D\log^{\frac{1}{\alpha}} n}.\end{aligned}$$

Define the events

$$\begin{aligned}&\mathcal{A}_0 = \{Z_n(\delta_n, G^*) \geq c_2\delta_n^2/(D\log^{\frac{1}{\alpha}} n)\}; \\ &\mathcal{A}_1 = \{\exists g \in G^*, \text{such that}\left|\frac{1}{n}\sum_{i=1}^{n} g(x_i) - \mathbb{E}_{p_{\mathcal{D}}(x)}g(x)]\right| \geq c_2\delta_n\|g\|_2/(D\log^{\frac{1}{\alpha}} n) \\ &\text{and } \|g\|_2 \geq \delta_n\}.\end{aligned}$$

Using equation (4), there exist some constants (c_0', c_1', c_2) that only depend on α such that

$$P(\mathcal{A}_0) \leq c_0' \exp\left(-c_1'\left(\frac{n\delta_n^2}{D^2\log^{\frac{2}{\alpha}} n}\right)^{\min\{\alpha,1\}}\right).$$

Define $\mathcal{S}_m = \left\{2^{m-1}\delta_n \leq \|g\|_2 \leq 2^m\delta_n\right\}$ with $m = 1, \cdots M$, since $\frac{\|g\|_2}{D}$ is upper bounded by some constant less than infinity, we have $M \lesssim \log(\frac{D}{\delta_n})$.

Under $\mathcal{A}_1 \cap \mathcal{S}_m$, it holds that $Z_n(2^m\delta_n) \geq c_2 2^{m-1}\delta_n^2/(D\log^{\frac{1}{\alpha}} n)$. Therefore by $\left(\frac{n\delta_n^2}{D^2\log^{\frac{2}{\alpha}} n}\right)^{\min\{\alpha,1\}} \gtrsim \log(\log\frac{D}{\delta_n})$, we know, for some constants (c_3, c_4) that only depend on α,

$$P(\mathcal{A}_1) = \sum_{m=1}^{M} P(\mathcal{A}_1 \cap \mathcal{S}_m) \leq c_3 \exp\left(-c_4\left(\frac{n\delta_n^2}{D^2\log^{\frac{2}{\alpha}} n}\right)^{\min\{\alpha,1\}}\right).$$

Moreover, under $\mathcal{A}_0^c \cap \mathcal{A}_1^c$, we have

$$\sup_{g\in G^*}\frac{\left|\frac{1}{n}\sum_{i=1}^{n} g(x_i) - \mathbb{E}_{p_{\mathcal{D}}(x)}g(x)\right|}{\delta_n + \|g\|_2} \leq c_2\delta_n/(D\log^{\frac{1}{\alpha}} n).$$

We can then get the desired conclusion.

B.2. Parametric Models

We use $p_{\theta,\beta}(x)$ to denote the marginal $\int p_\theta(x|z)\pi_\beta(z)dz$ and $p_{\theta,\beta}(z|x)$ to denote the posterior $\frac{p_\theta(x|z)\pi_\beta(z)}{\int p_\theta(x|z)\pi_\beta(z)dz}$. We begin the proof of Theorem 3 with the following lemma for dealing with the unboundedness of the objective function. The Proof of Lemma 13 is based on the proof of Proposition 6.7 of Ledoux and Talagrand (1991).

Lemma 13 *Consider G^* and $\overline{G}^*$ defined in Section 3. If $\left\| \sup_{g \in G^*} |g(x)| \right\|_{\psi_\alpha} \leq 2D$, then there exists $\rho \leq c_0 D \log^{\frac{1}{\alpha}} n$ and a constant c, where (c, c_0) only depend on α, such that $\forall \delta > 0$,*

$$\begin{aligned}
&\mathbb{E}_{p_{\mathcal{D}}(x)} \left[\sup_{\substack{g \in \overline{G}^* \\ \|g\|_2 \leq \delta}} \left| \frac{1}{n} \sum_{i=1}^n g(x_i) - \mathbb{E}_{p_{\mathcal{D}}(x)} g(x) \right| \right] \\
&\leq \mathbb{E}_{p_{\mathcal{D}}(x)} \left[\sup_{\substack{g \in \overline{G}^* \\ \|g\|_2 \leq \delta}} \left| \frac{1}{n} \sum_{i=1}^n g(x_i) \boldsymbol{1}_{\mathcal{A}}(x_i) - \mathbb{E}_{p_{\mathcal{D}}(x)} \left[g(x) \boldsymbol{1}_{\mathcal{A}}(x) \right] \right| \right] + c \frac{D \log^{\frac{1}{\alpha}} n}{n}.
\end{aligned}$$

where $\mathcal{A}$ denotes the event $\left\{ \sup_{g \in G^*} |g(x)| \leq \rho \right\}$, and $\boldsymbol{1}_{\mathcal{A}}(x)$ denotes the indicator function of event $\mathcal{A}$.

B.2.1. PROOF OF THEOREM 3

Choose $\rho \leq c_0 D \log^{\frac{1}{\alpha}} n$ in Lemma 13 and define $\mathcal{A} = \left\{ \sup_{g \in G^*} |g(x)| \leq \rho \right\}$. Define

$$\begin{aligned}
\overline{G}^*_{\mathcal{A}} &= \left\{ g_{\mathcal{A}}(x) = g(x) \mathbf{1}_{\mathcal{A}}(x), g(x) \in \overline{G}^* \right\}; \\
G^*_{\mathcal{A}} &= \left\{ g_{\mathcal{A}}(x) = g(x) \mathbf{1}_{\mathcal{A}}(x), g(x) \in G^* \right\},
\end{aligned} \tag{5}$$

with $\overline{G}^*$ and G^* being defined in Section 3. Using standard symmetrization, we can get

$$\mathbb{E}_{p_{\mathcal{D}}(x)} \left[\sup_{\substack{g \in \overline{G}^* \\ \|g\|_2 \leq r}} \left| \frac{1}{n} \sum_{i=1}^n g(x_i) \mathbf{1}_{\mathcal{A}}(x_i) - \mathbb{E}_{p_{\mathcal{D}}(x)} \left[g(x) \mathbf{1}_{\mathcal{A}}(x) \right] \right| \right] \leq \mathbb{E}_{p_{\mathcal{D}}(x)} \mathbb{E}_{\varepsilon} \left[\sup_{\substack{g_{\mathcal{A}} \in \overline{G}^*_{\mathcal{A}} \\ \|g_{\mathcal{A}}\|_2 \leq r}} \left| \frac{2}{n} \sum_{i=1}^n \varepsilon_i g_{\mathcal{A}}(x_i) \right| \right].$$

Define $d_n(g_{\mathcal{A}}, g'_{\mathcal{A}}) = \sqrt{\frac{1}{n} \sum_{i=1}^n (g_{\mathcal{A}}(x_i) - g'_{\mathcal{A}}(x_i))^2}$, then

$$r_n = \max_{\substack{g_{\mathcal{A}}, g'_{\mathcal{A}} \in \overline{G}^* \\ \|g_{\mathcal{A}}\|_2, \|g'_{\mathcal{A}}\|_2 \leq r}} d_n(g_{\mathcal{A}}, g'_{\mathcal{A}}) \leq 2\rho.$$

By equation (3.84) of Wainwright (2019), there exists a constant c such that,

$$
\begin{aligned}
\mathbb{E}_{p_{\mathcal{D}}(x)} r_n^2 &\leq \mathbb{E}_{p_{\mathcal{D}}(x)} \left[\sup_{\substack{g_{\mathcal{A}} \in \overline{G}^*_{\mathcal{A}} \\ \|g_{\mathcal{A}}\|_2 \leq r}} \frac{2}{n} \sum_{i=1}^n g^2_{\mathcal{A}}(x_i) \right] \\
&\leq \mathbb{E}_{p_{\mathcal{D}}(x)} \left[\sup_{\substack{g_{\mathcal{A}} \in \overline{G}^*_{\mathcal{A}} \\ \|g_{\mathcal{A}}\|_2 \leq r}} \frac{4}{n} \sum_{i=1}^n \left(g_{\mathcal{A}}(x_i) - \mathbb{E}_{p_{\mathcal{D}}(x)} g_{\mathcal{A}}(x) \right)^2 \right] + 4r^2 \\
&\leq c(r^2 + \rho \overline{R}_n(r, \overline{G}^*_{\mathcal{A}})).
\end{aligned}
$$

Since $G^*_{\mathcal{A}}$ is uniformly bounded by ρ, it holds that $\overline{R}_n(r, \overline{G}^*_{\mathcal{A}}) \leq \rho$ and we only need to consider $r \leq \rho$. Therefore we can get $c(r^2 + \rho \overline{R}_n(r, \overline{G}^*_{\mathcal{A}})) \leq c_0 \rho^2$. Moreover, for any $g_{\mathcal{A}} \in G^*_{\mathcal{A}}$ and $a \in (0, 1]$, there exists a $k \in \mathbb{N}$, such that $k\frac{\varepsilon}{2\rho} < a \leq (k+1)\frac{\varepsilon}{2\rho}$ and $d_n((k+1)\frac{\varepsilon}{2\rho} g_{\mathcal{A}}, a g_{\mathcal{A}}) \leq \frac{\varepsilon}{2\rho}\rho = \frac{\varepsilon}{2}$. Therefore it follows that the ε- covering number of $\overline{G}^*_{\mathcal{A}}$ satisfies that, $\mathbf{N}(\overline{G}^*_{\mathcal{A}}, d_n, \varepsilon) \leq \mathbf{N}(G^*_{\mathcal{A}}, d_n, \frac{\varepsilon}{2})\frac{2\rho}{\varepsilon}$ and $\log \mathbf{N}(\overline{G}^*_{\mathcal{A}}, d_n, \varepsilon) \leq \log \mathbf{N}(G^*_{\mathcal{A}}, d_n, \frac{\varepsilon}{2}) + \log \frac{2\rho}{\varepsilon}$. Recall the definition of $G^*_{\mathcal{A}}$ in equation (5), it follows that

$$
\begin{aligned}
\forall g_{\mathcal{A}}, g'_{\mathcal{A}} \in G^*_{\mathcal{A}}, \quad d_n(g_{\mathcal{A}}, g'_{\mathcal{A}}) &= \sqrt{\frac{1}{n} \sum_{i=1}^n (m(p_\theta, q_\phi, \pi_\beta, x_i) - m(p_{\theta'}, q_{\phi'}, \pi_{\beta'}, x_i))^2 \mathbf{1}_{\mathcal{A}}(x_i)} \\
&\leq \sqrt{\frac{1}{n} \sum_{i=1}^n b^2(x_i) \|(\theta, \phi, \beta) - (\theta', \phi', \beta')\|_2^2} \\
&= \sqrt{\frac{1}{n} \sum_{i=1}^n b^2(x_i)} \|(\theta, \phi, \beta) - (\theta', \phi', \beta')\|_2 \\
&= \overline{d}_n((\theta, \phi, \beta), (\theta', \phi', \beta')).
\end{aligned}
$$

W.l.o.g, we can assume $\|\theta\|_\infty + \|\phi\|_\infty + \|\beta\|_\infty \leq 1$. By the fact that the ε-covering number of unit ball in R^d is smaller than $(\frac{3}{\varepsilon})^d$, let $d^* = d_\theta + d_\phi + d_\beta$, we have

$$
\log \mathbf{N}(G^*_{\mathcal{A}}, d_n, \frac{\varepsilon}{2}) \leq \log \mathbf{N}(\Theta, \overline{d}_n, \frac{\varepsilon}{2}) \leq d^* \log \left(\frac{3\sqrt{d^*}\sqrt{\frac{1}{n}\sum_{i=1}^n b^2(x_i)}}{\varepsilon} \right).
$$

We next analyze the Dudley entropy integral in the following lemma.

Lemma 14 *Given Condition A, there exists a constant c_1 that only depend on a_1 in Condition A such that,*

$$
\begin{aligned}
&\mathbb{E}_{p_{\mathcal{D}}(x)} \left[\int_0^{r_n} \sqrt{d^* \log \left(\frac{3\sqrt{\frac{d^*}{n}\sum_{i=1}^n b^2(x_i)}}{\varepsilon} \right) + \log \frac{2\rho}{\varepsilon}} d\varepsilon \right] \\
&\leq c_1 \left(\rho\sqrt{d^*} \sqrt{-\mathbb{E}_{p_{\mathcal{D}}(x)} \left[(\frac{r_n}{2\rho})^2 \log \mathbb{E}_{p_{\mathcal{D}}(x)} (\frac{r_n}{2\rho})^2 \right] + \mathbb{E}_{p_{\mathcal{D}}(x)} (\frac{r_n}{2\rho})^2} + \mathbb{E}_{p_{\mathcal{D}}(x)} \left[r_n \sqrt{d^* \log d^*} \right] \right).
\end{aligned}
$$

Since $r_n \le 2\rho$ and $\sqrt{-xlogx+x}$ is an increasing function when $x<1$, by $\mathbb{E}_{p_{\mathcal{D}}(x)} r_n^2 \le c(r^2 + \rho\overline{R}_n(r,\overline{G}^*_{\mathcal{A}})) \le c_0\rho^2$ and Dudley inequality (see, for example, (8.13) of Vershynin (2018)), we have

$$\overline{R}_n(r,\overline{G}^*_{\mathcal{A}}) \lesssim \frac{1}{\sqrt{n}}(r^2+\rho\overline{R}_n(r,\overline{G}^*_{\mathcal{A}}))^{\frac{1}{2}}\sqrt{\log\frac{\rho}{r}+\log d^*}\sqrt{d^*}.$$

Choose $\delta_n = c_2\sqrt{\frac{\log n+\log d^*}{n}d^*}D\log^{\frac{1}{\alpha}} n$, if $\overline{R}_n(\delta_n,\overline{G}^*_{\mathcal{A}}) > \delta_n^2/(D\log^{\frac{1}{\alpha}} n)$, then

$$\overline{R}_n(\delta_n,\overline{G}^*_{\mathcal{A}}) \lesssim \frac{1}{\sqrt{n}}\rho^{\frac{1}{2}}\overline{R}_n(\delta_n,\overline{G}^*_{\mathcal{A}})^{\frac{1}{2}}\sqrt{d^*(\log n+\log d^*)},$$

which means

$$\overline{R}_n(\delta_n,\overline{G}^*_{\mathcal{A}}) \lesssim \frac{\rho d^*}{n}\log(nd^*).$$

Therefore for a large enough c_2, we have $\overline{R}_n(\delta_n,\overline{G}^*) \le \delta_n^2/(D\log^{\frac{1}{\alpha}} n)$ and $\left(\frac{n\delta_n^2}{D^2\log^{\frac{2}{\alpha}} n}\right)^{\min\{\alpha,1\}} \ge \log(\log\frac{D}{\delta_n})$, the desired conclusion then follows from Theorem 1.

B.2.2. PROOF OF COROLLARY 4

$$\begin{aligned}
&D_{\mathrm{TV}}\left(\int_{\mathcal{Z}} p_{\hat\theta}(\cdot|z)\frac{1}{n}\sum_{i=1}^n q_{\hat\phi}(z|x_i)dz, p_{\mathcal{D}}(\cdot)\right)\\
&=\frac{1}{2}\int_{\mathcal{X}}\left|\int_{\mathcal{Z}} p_{\hat\theta}(x|z)\frac{1}{n}\sum_{i=1}^n q_{\hat\phi}(z|x_i)dz - p_{\mathcal{D}}(x)\right|dx\\
&\le\frac{1}{2}\int_{\mathcal{X}}\left|\int_{\mathcal{Z}} p_{\hat\theta}(x|z)\frac{1}{n}\sum_{i=1}^n q_{\hat\phi}(z|x_i)dz - \int_{\mathcal{Z}} p_{\hat\theta}(x|z)\mathbb{E}_{p_{\mathcal{D}}(x)}q_{\hat\phi}(z|x)dz\right|dx\\
&+\frac{1}{2}\int_{\mathcal{X}}\left|\int_{\mathcal{Z}} p_{\hat\theta}(x|z)\mathbb{E}_{p_{\mathcal{D}}(x)}q_{\hat\phi}(z|x)dz - \int_{\mathcal{Z}} p_{\hat\theta}(x|z)\pi_{\hat\beta}(z)dz\right|dx\\
&+D_{\mathrm{TV}}(p_{\hat\theta,\hat\beta}(\cdot),p_{\mathcal{D}}(\cdot))\\
&\le\sup_{\phi,z}\left|\frac{1}{n}\sum_{i=1}^n q_\phi(z|x_i)-\mathbb{E}_{p_{\mathcal{D}}(x)}q_\phi(z|x)\right| + D_{\mathrm{TV}}\left(\int_{\mathcal{X}} q_{\hat\phi}(\cdot|x)p_{\mathcal{D}}(x)dx, \pi_{\hat\beta}(\cdot)\right)\\
&+D_{\mathrm{TV}}(p_{\hat\theta,\hat\beta}(\cdot),p_{\mathcal{D}}(\cdot)).
\end{aligned}$$

By the fact that $\int_{\mathcal{X}} p_{\hat\theta,\hat\beta}(z|x)p_{\hat\theta,\hat\beta}(x)dx = \pi_{\hat\beta}(z)$, it holds that

$$\begin{aligned}
&D_{\mathrm{TV}}\left(\int_{\mathcal{X}} q_{\hat\phi}(\cdot|x)p_{\mathcal{D}}(x)dx, \pi_{\hat\beta}(\cdot)\right)\\
&=\frac{1}{2}\int_{\mathcal{Z}}\left|\int_{\mathcal{X}} q_{\hat\phi}(z|x)p_{\mathcal{D}}(x)dx - \pi_{\hat\beta}(z)\right|dz\\
&\le\frac{1}{2}\int_{\mathcal{Z}}\left|\int_{\mathcal{X}} q_{\hat\phi}(z|x)p_{\mathcal{D}}(x)dx - \int_{\mathcal{X}} p_{\hat\theta,\hat\beta}(z|x)p_{\mathcal{D}}(x)dx\right|dz\\
&+\frac{1}{2}\int_{\mathcal{Z}}\left|\int_{\mathcal{X}} p_{\hat\theta,\hat\beta}(z|x)p_{\mathcal{D}}(x)dx - \int_{\mathcal{X}} p_{\hat\theta,\hat\beta}(z|x)p_{\hat\theta,\hat\beta}(x)dx\right|dz.
\end{aligned} \tag{6}$$

For the first term in equation (6), we can further upper bound,

$$
\begin{aligned}
&\frac{1}{2}\int_{\mathcal{Z}}\left|\int_{\mathcal{X}} q_{\hat{\phi}}(z|x)p_{\mathcal{D}}(x)dx-\int_{\mathcal{X}} p_{\hat{\theta},\hat{\beta}}(z|x)p_{\mathcal{D}}(x)dx\right|dz\\
&\leq \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{Z}}\left|q_{\hat{\phi}}(z|x)-p_{\hat{\theta},\hat{\beta}}(z|x)\right|dzp_{\mathcal{D}}(x)dx\\
&=\int_{\mathcal{X}} D_{\mathrm{TV}}(q_{\hat{\phi}}(\cdot|x),p_{\hat{\theta},\hat{\beta}}(\cdot|x))p_{\mathcal{D}}(x)dx,
\end{aligned}
$$

and for the second term,

$$
\begin{aligned}
&\frac{1}{2}\int_{\mathcal{Z}}\left|\int_{\mathcal{X}} p_{\hat{\theta},\hat{\beta}}(z|x)p_{\mathcal{D}}(x)dx-\int_{\mathcal{X}} p_{\hat{\theta},\hat{\beta}}(z|x)p_{\hat{\theta}}(x)dx\right|dz\\
&\leq \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{Z}}\left|p_{\mathcal{D}}(x)-p_{\hat{\theta},\hat{\beta}}(x)\right|p_{\hat{\theta},\hat{\beta}}(z|x)dzdx\\
&=D_{\mathrm{TV}}(p_{\mathcal{D}}(\cdot),p_{\hat{\theta},\hat{\beta}}(\cdot)).
\end{aligned}
$$

Now we define $Z_n=\sup_{\phi,z}\left|\frac{1}{n}\sum_{i=1}^n q_\phi(z|x_i)-\mathbb{E}_{p_{\mathcal{D}}(x)}q_\phi(z|x)\right|$, then by $\|q_\phi(z|x)-q_{\phi'}(z'|x)\|\leq c(\|\phi-\phi'\|_2+\|z-z'\|_2)$ and the compactness of parameter space and latent space, using Dudley inequality (see, for example, (8.13) of Vershynin (2018)) and Talagrand concentration inequality (see, for example, 3.27 of Wainwright (2019)), we can get that it holds with probability larger than $1-\exp(c\log n)$ that $Z_n\lesssim\sqrt{\frac{d_\phi\log(d_\phi n)}{n}}$. Then, the desired conclusion follows from Theorem 3 and Pinsker inequality (see, for example, Theorem 2.16 of Massart (2007)).

B.3. Gaussian Encoders and Decoders

We use $p_{\theta,\beta}(x)$ to denote the marginal $\int p_\theta(x|z)\pi_\beta(z)dz$ and $p_{\theta,\beta}(z|x)$ to denote the posterior $\frac{p_\theta(x|z)\pi_\beta(z)}{\int p_\theta(x|z)\pi_\beta(z)dz}$.

B.3.1. Proof of Theorem 7

To begin with, we make the following definition.

Definition 15 *$\mathcal{F}_d^D(L,W,U,b,V)$ is defined as the set of following feedfoward ReLU neural networks (feedfoward neural network with ReLU activation $\sigma(x)=\max(x,0)$): 1. The information of each layer can only come from the previous one layer. 2. U is a $(L-2)$-dimensional vector, $U=(u_1,\cdots,u_{L-2})$. The network has d input units, D output units, L layers, u_{l-1} computation units in layer l $(2\leq l\leq L-1)$ and W weights (parameters). 2. There exists a constant $V\geq 2$ such that in each layer, the absolute value of each weight unit is upper bounded by V. 3. The output unit has the Hard Tanh $h_b(x)=\max(-b,\min(b,x))$ as its activation function. In particular, we use $\mathcal{F}_d^D(L,W,U,b)$ to denote $\mathcal{F}_d^D(L,W,U,b,V)$ with $V=+\infty$.*

Then we consider the following ReLU neural networks: $Q_{\phi_1}(x)\in\mathcal{F}_{d_x}^{d_z}(L_1,U_1,W_1,b_8(\log\frac{1}{\sigma_1})^{\frac{1}{2}})$, $G_{\phi_2}^d(z)\in\mathcal{F}_{d_z}^{d_xd_z}(L_2,U_2,W_2,b_9(\log\frac{1}{\sigma_1})^{\frac{1}{\alpha}})$ and $G_{\theta_1}(z),G_{\phi_3}(z)\in\mathcal{F}_{d_z}^{d_x}(L,U,W,b(\log\frac{1}{\sigma_1})^{\frac{1}{\alpha}},V)$, where $G_{\phi_2}^d(z)$ is rescaled to be a $d_x\times d_z$ matrix. Since there is no boundary towards the support of the data x and the latent variable z, we define compact sets of z and ϵ: $B_z = [-\eta \log^{\frac{1}{2}} \frac{1}{\sigma^*}, \eta \log^{\frac{1}{2}} \frac{1}{\sigma^*}]^{d_z}$ and $B_\epsilon = [-\gamma \log^{\frac{1}{2}} \frac{1}{\sigma^*}, \gamma \log^{\frac{1}{2}} \frac{1}{\sigma^*}]^{d_x}$. And let $B_x = \{x = G_{\mathcal{D}}(z) + \sigma^* \epsilon \,|\, z \in B_z,\ \epsilon \in B_\epsilon\}$. Then we define $\overline{B}_z = [-\overline{\eta} \log^{\frac{1}{2}} \frac{1}{\sigma^*}, \overline{\eta} \log^{\frac{1}{2}} \frac{1}{\sigma^*}]^{d_z}$ so that $Q_{\mathcal{D}}(B_x) \subseteq \overline{B}_z$. Next We define following numbers to characterize the expressivity of families of Q_{ϕ_1}, G_{ϕ_3} and $G^d_{\phi_2}$,

$$\begin{aligned} \epsilon_0 &:= \min_{Q_{\phi_1}} \max_{x \in B_x} \left\| Q_{\phi_1}(x) - Q_{\mathcal{D}}(x) \right\|_2 ; \\ \epsilon_1 &:= \min_{G_{\phi_3}} \max_{z \in \overline{B}_z} \| G_{\phi_3}(z) - G_{\mathcal{D}}(z) \|_2 ; \\ \epsilon_2 &:= \min_{G^d_{\phi_2}} \max_{z \in \overline{B}_z} \| G^d_{\phi_2}(z) - \nabla G_{\mathcal{D}}(z) \|_F . \end{aligned} \tag{7}$$

Here we omit families of $(Q_{\phi_1}, G_{\phi_3}, G^d_{\phi_2})$ and the dependency of $(\epsilon_0, \epsilon_1, \epsilon_2)$ on $(\eta, \gamma, \overline{\eta}, G_{\mathcal{D}}(z), Q_{\mathcal{D}}(x))$ and families of $(Q_{\phi_1}, G_{\phi_3}, G^d_{\phi_2})$ for ease of notation.

Remark 16 *The decoder is using a gaussian to approximate the posterior, so we want $p_{\mathcal{D}}(z|x) = \frac{p_{\mathcal{D}}(x|z)\pi_{\mathcal{D}}(z)}{p_{\mathcal{D}}(x)}$ to be well approximated by a gaussian either. When σ^* is small, the above assumptions on $G_{\mathcal{D}}(z)$ can guarantee that the space of z that are likely under $p_{\mathcal{D}}(z|x)$ is z being close to $Q_{\mathcal{D}}(x)$. Also, we have $p_{\mathcal{D}}(z|x) \propto \pi_{\mathcal{D}}(z) \exp\left(-\frac{1}{2\sigma^{*2}} \sum_{j=1}^{d_x} (x_j - G^j_{\mathcal{D}}(z))^2\right)$, consider the first order Taylor expansion of $G^j_{\mathcal{D}}(z)$ at $z = Q_{\mathcal{D}}(x)$, $p_{\mathcal{D}}(z|x)$ can be well approximated by a gaussian distribution with mean $Q_{\mathcal{D}}(x) + \Sigma_{\mathcal{D}}(x) \nabla G_{\mathcal{D}}(Q_{\mathcal{D}}(x))^T (x - G_{\mathcal{D}}(Q_{\mathcal{D}}(x)))$ and covariance matrix ${\sigma^*}^2 \Sigma_{\mathcal{D}}(x)$ with $\Sigma_{\mathcal{D}}(x) = (\nabla G_{\mathcal{D}}(Q_{\mathcal{D}}(x))^T \nabla G_{\mathcal{D}}(Q_{\mathcal{D}}(x)))^{-1}$, where $\nabla G_{\mathcal{D}}(Q_{\mathcal{D}}(x)) = \nabla G_{\mathcal{D}}(z)|_{z = Q_{\mathcal{D}}(x)}$.*

So we can make specific choices of $\Sigma_\phi(x)$ and $\mu_\phi(x)$,

$$\begin{aligned} \Sigma_\phi(x) &= \bar{\sigma}^2 \left(G^d_{\phi_2}(Q_{\phi_1}(x))^T G^d_{\phi_2}(Q_{\phi_1}(x)) + \bar{\sigma}^2 I_{d_z} \right)^{-1} ; \\ \tilde{\Sigma}_\phi(x)_{i,j} &= \max(-\bar{b}_7, \min(\bar{b}_7, \frac{1}{\bar{\sigma}^2} \Sigma_\phi(x)_{i,j})); \\ \mu_\phi(x) &= Q_{\phi_1}(x) + \tilde{\Sigma}_\phi(x) G^d_{\phi_2}(Q_{\phi_1}(x))^T (x - G_{\phi_3}(Q_{\phi_1}(x))), \end{aligned} \tag{8}$$

where $\Sigma_\phi(x)_{i,j}$ is the (i,j) element of $\Sigma_\phi(x)$, $\bar{b}_7 = b_7 \left(\log \frac{1}{\sigma_1}\right)^{\frac{4}{\alpha^2}}$ with a large enough constant b_7 and $\bar{\sigma} \in [\sigma_1, 1]$ is a parameter. Here we add $\bar{\sigma}^2 I_{d_z}$ to $\Sigma_\phi(x)$ to guarantee the positive definiteness. For ease of notation, we use Θ_{θ_1} to denote the parameter spaces of G_{θ_1}, and use $\Theta_{\tilde{\phi}}$ to denote the cartesian product of parameter spaces of Q_{ϕ_1}, $G^d_{\phi_2}$ and G_{ϕ_3}, we can then define:

$$\begin{aligned} &\mathcal{F}_{dd} = \left\{ p_\theta(x|z) = \mathcal{N}(G_{\theta_1}(z), \sigma^2 I_{d_x}) \,|\, \theta_1 \in \Theta_{\theta_1},\ \sigma \in [\sigma_1, 1] \right\}; \\ &\mathcal{F}_{ed} = \left\{ q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \Sigma_\phi(x)) \,|\, \phi = (\phi_1, \phi_2, \phi_3, \bar{\sigma}),\ (\phi_1, \phi_2, \phi_3) \in \Theta_{\tilde{\phi}},\ \bar{\sigma} \in [\sigma_1, 1],\ \bar{b}_7 = b_7 \left(\log \frac{1}{\sigma_1} \right)^{\frac{4}{\alpha^2}} \right\}. \end{aligned} \tag{9}$$

We then state the following Lemma 17 to bound the error of ReLU neural networks for approximating $Q_{\mathcal{D}}(x)$ satisfying Assumption B, whose domain is close to a d_z-dimensional submanifold, the proof of Lemma 17 is based on the proof of Theorem 1 of Yarotsky (2017).

Lemma 17 *Consider B_x, $\overline{B}_z$, $(Q_{\phi_1}, G_{\phi_3}, G^d_{\phi_2})$ and $(\epsilon_0, \epsilon_1, \epsilon_2)$ defined above, there exist some constants (c_0, c, c_1, c_2, c_3) that only depend on $(\eta, \gamma, \overline{\eta}, d_z, d_x)$ and (α, k, b_6) in Assumption B, such that if we choose*

$$
\begin{aligned}
&b = b_8 = b_9 = c_0,\ L = L_1 = L_2 = c\log\frac{1}{\sigma_1},\ V = \sigma_1^{-c_3},\ N = \left(\frac{\sigma_1}{\left(\log\frac{1}{\sigma_1}\right)^{\frac{2d_z+2}{\alpha^2}+\frac{d_z+k}{\alpha}+\frac{k}{2}}}\right)^{-\frac{1}{k}};\\
&W = c_1\left(\log\frac{1}{\sigma_1}\right)^{\frac{d_z}{\alpha k}+\frac{d_z}{2}+2}({\sigma_1}^2)^{-\frac{d_z}{k}},\ W_1 = c_1N^{d_z}(\log\frac{1}{\sigma_1})^2,\ W_2 = c_1\left(\log\frac{1}{\sigma_1}\right)^{\frac{d_z}{\alpha(k-1)}+\frac{d_z}{2}+2}(\sigma_1)^{-\frac{d_z}{k-1}};\\
&u = c_2\left(\log\frac{1}{\sigma_1}\right)^{\frac{d_z}{\alpha k}+\frac{d_z}{2}+1}({\sigma_1}^2)^{-\frac{d_z}{k}},\ u_1 = c_2N^{d_z}(\log\frac{1}{\sigma_1}),\ u_2 = c_2\left(\log\frac{1}{\sigma_1}\right)^{\frac{d_z}{\alpha(k-1)}+\frac{d_z}{2}+1}(\sigma_1)^{-\frac{d_z}{k-1}};\\
&U = (\underbrace{u,\cdots,u}_{L-2}),\ U_1 = (\underbrace{u_1,\cdots,u_1}_{L_1-2}),\ U_2 = (\underbrace{u_2,\cdots,u_2}_{L_2-2}),
\end{aligned}
$$

then for any $G_{\mathcal{D}}(z)$ and $Q_{\mathcal{D}}(x)$ satisfying Assumption B and $\sigma^* \geq \sigma_1$, it holds that $\frac{\epsilon_0}{\sigma^*}+\frac{\epsilon_1}{\sigma^{*2}}+\frac{\epsilon_2}{\sigma^*} \leq 1$.

We can then state the following lemma to bound the approximation error and the excess risk of the EBVAE estimator with Gaussian encoder/decoder.

Lemma 18 *Consider $\mathcal{F}_{dd}$, $\mathcal{F}_{ed}$ defined in (9), $\mathcal{F}_{prior}$ satisfying Condition B and the EBVAE estimator $p_{\hat{\theta}}$, $q_{\hat{\phi}}$ and $\pi_{\hat{\beta}}$ defined in (2), suppose Assumption B is satisfied, then there exist some constants $(\eta, \gamma, \overline{\eta}, b_7, c_1, c_2, c_3)$ that only depend on (d_z, d_x) and $(\alpha, k, b_2, b_3, b_5, b_6)$ in Assumption B and Condition B, such that when $\frac{\epsilon_0}{\sigma^*}+\frac{\epsilon_1}{\sigma^{*2}}+\frac{\epsilon_2}{\sigma^*} \leq 1$, it holds with probability at least $1-\frac{1}{n^c}$ that,*

$$
\begin{aligned}
&D_{\mathrm{KL}}(p_{\mathcal{D}}(\cdot)||p_{\hat{\theta},\hat{\beta}}(\cdot)) + \mathbb{E}_{p_{\mathcal{D}}(x)}\left[D_{\mathrm{KL}}(q_{\hat{\phi}}(\cdot|x)||p_{\hat{\theta},\hat{\beta}}(\cdot|x))\right]\\
&\leq c_1{\sigma^*}^2\left(\log\frac{1}{\sigma^*}\right)^{\left(\frac{28+10\alpha+3\alpha^2}{\alpha^2}\right)} + c_2\frac{\log^{\frac{2}{\alpha}}n}{n\sigma_1^2}\left(\log\frac{1}{\sigma_1}\right)^{\frac{2}{\alpha}}\left(\log n + L\log(V\|U\|_1) + \log\frac{1}{\sigma_1}\right)\\
&\times\Big(d_\beta + (W_1+W_2)(L_1+L_2)\log(\|U_1\|_1+\|U_2\|_1) + (W_1+W)(L_1+L)\log(\|U_1\|_1+\|U\|_1)\Big).
\end{aligned}
$$

So, by lemma 17 and lemma 18 and the fact that $k \geq 2$ and $\sigma_1 \leq \sigma^*$, one has

$$
\begin{aligned}
&D_{\mathrm{KL}}(p_{\mathcal{D}}(\cdot)||p_{\hat{\theta},\hat{\beta}}(\cdot)) + \mathbb{E}_{p_{\mathcal{D}}(x)}\left[D_{\mathrm{KL}}(q_{\hat{\phi}}(\cdot|x)||p_{\hat{\theta},\hat{\beta}}(\cdot|x))\right]\\
&\leq c_1{\sigma^*}^2\left(\log\frac{1}{\sigma^*}\right)^{\tilde{\alpha}_1} + c_2\frac{\log^{\frac{2}{\alpha}}n}{n}(d_\beta\sigma_1^{-2} + {\sigma_1}^{-(2+\frac{2d_z}{k})})\left(\log\frac{1}{\sigma_1}\right)^{\tilde{\alpha}_2},
\end{aligned}
$$

where $\tilde{\alpha}_1 = \frac{28+10\alpha+3\alpha^2}{\alpha^2}$ and $\tilde{\alpha}_2 = \frac{2}{\alpha}+\frac{d_z}{\alpha(k-1)}+\frac{d_z}{2}+6$.

B.4. Nonparametric Models

We use the notation $p(x)$ to denote the marginal $\int p(x|z)\pi_\beta(z)dz$ and $p(z|x)$ to denote the posterior $\frac{p(x|z)\pi_\beta(z)}{\int p(x|z)\pi_\beta(z)dz}$. We begin the proof of Theorem 9 with the following two lemmas.

Lemma 19 *When Condition C and Assumption C hold, consider $m(p, q, \pi_\beta, x)$ defined in (1), there exists a constant c such that $\forall p(x|z), p'(x|z) \in \mathcal{F}_{dd}$, $\forall q(z|x), q'(z|x) \in \mathcal{F}_{ed}$ and $\forall \pi_\beta(z), \pi_{\beta'}(z) \in \mathcal{F}_{prior}$, it holds that*

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} m(p, q, \pi_\beta, x_i) - m(p', q', \pi_{\beta'}, x_i))^2}$$

$$\leq c\left(\sup_{\substack{x\in\mathcal{X}\\ z\in\mathcal{Z}}}|\log q(z|x) - \log q'(z|x)| + \sup_{\substack{x\in\mathcal{X}\\ z\in\mathcal{Z}}}|\log p(x|z) - \log p'(x|z)| + \sup_{z\in\mathcal{Z}}\left|\log \pi_\beta(z) - \log \pi_{\beta'}(z)\right|\right).$$

Lemma 20 *If $p(x|z) \propto \exp\left(\sum_{j=1}^{k_1} l_j(\overline{x}_j, z)\right)$, $p'(x|z) \propto exp\left(\sum_{j=1}^{k_1} l'_j(\overline{x}_j, z)\right)$ and $q(z|x) \propto exp\left(\sum_{j=1}^{k_2} f_j(z, \overline{x}_j)\right)$, $q'(z|x) \propto exp\left(\sum_{j=1}^{k_2} f'_j(z, \overline{x}_j)\right)$. Then,*

$$\sup_{x\in\mathcal{X},z\in\mathcal{Z}}|\log q(z|x) - \log q'(z|x)| + \sup_{x\in\mathcal{X},z\in\mathcal{Z}}|\log p(x|z) - \log p'(x|z)|$$

$$\leq 2\left(\sup_{x\in\mathcal{X},z\in\mathcal{Z}}\left|\sum_{j=1}^{k_2} f_j(z, \overline{x}_j) - \sum_{j=1}^{k_2} f'_j(z, \overline{x}_j)\right| + \sup_{x\in\mathcal{X},z\in\mathcal{Z}}\left|\sum_{j=1}^{k_1} l_j(\overline{x}_j, z) - \sum_{j=1}^{k_1} l'_j(\overline{x}_j, z)\right|\right).$$

B.4.1. PROOF OF THEOREM 9

W.l.o.g, we can assume $k_1 = k_2 = k$ and $r_1 = r_2 = 1$. Consider $\overline{G}^*$ and G^* defined in Section 3. By Condition C, we have

$$\begin{aligned}|g(x)| &= |m(p, q, \pi_\beta, x) - m(p^*, q^*, \pi_{\beta^*}, x)|\\ &= \left|\log\frac{p(x)}{p^*_{dd,\beta^*}(x)} + D_{\mathrm{KL}}(q(\cdot|x)||p(\cdot|x)) - D_{\mathrm{KL}}(q^*(\cdot|x)||p^*_{dd,\beta^*}(\cdot|x))\right|\\ &\leq 2(\sup_{\substack{p\in\mathcal{F}_{dd}, q\in\mathcal{F}_{ed}\\ \pi_\beta\in\mathcal{F}_{prior}}}\sup_{\substack{x\in[0,1]^{d_x}\\ z\in[0,1]^{d_z}}}(|\log p(x)| + |\log q(z|x)| + |\log p(z|x)|))\\ &\leq 2C.\end{aligned}$$

Therefore $\overline{G}^*$ is uniformly bounded by 2C.

First we consider $\hat{R}_n(\delta, \overline{G}^*) = \mathbb{E}_\varepsilon[\sup_{\substack{g\in\overline{G}^*\\ \|g\|_n\leq\delta}}|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(x_i)|]$ and

$$d_n(g, g') = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(g(x_i) - g'(x_i))^2}.$$

By Corollary 14.3 of Wainwright (2019), $\hat{R}_n(\hat{\delta}_n, \overline{G}^*) \leq \frac{\hat{\delta}_n^2}{2C}$ is satisfied if

$$\frac{64}{\sqrt{n}}\int_{\frac{\hat{\delta}_n^2}{4C}}^{\hat{\delta}_n}\sqrt{\log \mathbf{N}(\overline{G}^*, d_n, \varepsilon)}d\varepsilon \leq \frac{\hat{\delta}_n^2}{2C}.$$

Furthermore, by the same argument of the proof of Theorem 3, one has

$$\log \mathbf{N}(\overline{G}^*, d_n, \varepsilon) \leq \log \mathbf{N}(G^*, d_n, \varepsilon) + \log \frac{4C}{\varepsilon}.$$

Define set $\mathcal{Q}$ and $\mathcal{G}$ as

$$\mathcal{Q} := \left\{ \sum_{j=1}^{k} l_j(\overline{x}_j, z) \,|\, l_j(\overline{x}_j, z) \in C_1^{\overline{\alpha}}([0,1]^{|\overline{x}_j|+d_z}), |\overline{x}_j| \leq \overline{p} \right\};$$

$$\mathcal{G} := \left\{ \sum_{j=1}^{k} f_j(z, \overline{x}_j) \,|\, f_j(z, \overline{x}_j) \in C_1^{\overline{\alpha}}([0,1]^{|\overline{x}_j|+d_z}), |\overline{x}_j| \leq \overline{p} \right\}.$$

Then by Lemma 19 and Lemma 20, we have for some constant c,

$$\mathbf{N}(G^*, d_n, \varepsilon) \leq c\mathbf{N}(\mathcal{G}, \|.\|_\infty, \frac{\varepsilon}{3}) \cdot \mathbf{N}(\mathcal{Q}, \|.\|_\infty, \frac{\varepsilon}{3}) \cdot \mathbf{N}(\Theta_\beta, \|.\|_2, \frac{\varepsilon}{3}).$$

Since every $G_j(\overline{x}_j, z)$ with $|\overline{x}_j| \leq \overline{p}$ can be seen as a function of $\overline{x}'_j \supseteq \overline{x}_j$ with $|\overline{x}'_j| = \overline{p}$, we can assume $|\overline{x}_j| = \overline{p}$. Since $\log \mathbf{N}(B_1^{\overline{\alpha}}([0,1]^{\overline{p}+d_z}), \|.\|_\infty, \varepsilon) \lesssim (\frac{1}{\varepsilon})^{\frac{\overline{p}+d_z}{\overline{\alpha}}}$ (see equation (5.17) of (Wainwright, 2019)) and $\binom{d_x}{\overline{p}} \leq (\frac{ed_x}{\overline{p}})^{\overline{p}}$, we can get

$$\begin{aligned} &\log \mathbf{N}(\mathcal{G}, \|.\|_\infty, \frac{\varepsilon}{2}) \\ &\leq \log \left(\binom{d_x}{\overline{p}} \mathbf{N}(B_1^{\overline{\alpha}}([0,1]^{\overline{p}+d_z}), \|.\|_\infty, \frac{\varepsilon}{k}) \right)^k \\ &\lesssim k^{1+\frac{d_z+\overline{p}}{\overline{\alpha}}} (\frac{1}{\varepsilon})^{\frac{\overline{p}+d_z}{\overline{\alpha}}} + k\overline{p} \log \frac{ed_x}{\overline{p}}. \end{aligned}$$

Similarly, we also have $\log \mathbf{N}(\mathcal{Q}, \|.\|_\infty, \frac{\varepsilon}{2}) \lesssim k^{1+\frac{d_z+\overline{p}}{\overline{\alpha}}} (\frac{1}{\varepsilon})^{\frac{\overline{p}+d_z}{\overline{\alpha}}} + k\overline{p} \log \frac{ed_x}{\overline{p}}$. Then, combined with the fact that $\log \mathbf{N}(\Theta_\beta, \|.\|_2, \varepsilon) \leq d_\beta \log(\frac{3}{\epsilon})$, we can get $\log \mathbf{N}(\overline{G}^*, d_n, \varepsilon) \lesssim k^{1+\frac{d_z+\overline{p}}{\overline{\alpha}}} (\frac{1}{\varepsilon})^{\frac{\overline{p}+d_z}{\overline{\alpha}}} + k\overline{p} \log \frac{ed_x}{\overline{p}}$. Therefore, $\hat{R}_n(\hat{\delta}_n, \overline{G}^*) \leq \frac{\hat{\delta}_n^2}{2C}$ is satisfied if,

1. when $\overline{p} + dz < 2\overline{\alpha}$

$$\frac{1}{\sqrt{n}} \int_0^{\hat{\delta}_n} \sqrt{k^{1+\frac{d_z+\overline{p}}{\overline{\alpha}}} (\frac{1}{\varepsilon})^{\frac{\overline{p}+d_z}{\overline{\alpha}}} + k\overline{p} \log \frac{ed_x}{\overline{p}}} d\varepsilon \lesssim \hat{\delta}_n^2.$$

   Choose $\hat{\delta}_n \asymp n^{\frac{-\overline{\alpha}}{2\overline{\alpha}+d_z+\overline{p}}} k^{\frac{\overline{\alpha}+d_z+\overline{p}}{2\overline{\alpha}+d_z+\overline{p}}} + \sqrt{\frac{k\overline{p}}{n} \log \frac{ed_x}{\overline{p}}}$.

2. when $\overline{p} + dz > 2\overline{\alpha}$

$$\frac{1}{\sqrt{n}} \int_{\frac{\hat{\delta}_n^2}{4C}}^{\infty} \sqrt{k^{1+\frac{d_z+\overline{p}}{\overline{\alpha}}} (\frac{1}{\varepsilon})^{\frac{\overline{p}+d_z}{\overline{\alpha}}}} d\varepsilon + \frac{1}{\sqrt{n}} \hat{\delta}_n \sqrt{k\overline{p} \log \frac{ed_x}{\overline{p}}} \lesssim \hat{\delta}_n^2.$$

   Choose $\hat{\delta}_n \asymp n^{\frac{-\overline{\alpha}}{2(d_z+\overline{p})}} k^{\frac{\overline{\alpha}+d_z+\overline{p}}{2(d_z+\overline{p})}} + \sqrt{\frac{k\overline{p}}{n} \log \frac{ed_x}{\overline{p}}}$.

3. when $\overline{p} + dz = 2\overline{\alpha}$

$$\frac{1}{\sqrt{n}} \int_{\frac{\hat{\delta}_n^2}{4C}}^{\hat{\delta}_n} \sqrt{k^3 (\frac{1}{\varepsilon})^2} d\varepsilon + \sqrt{\frac{k\overline{p}}{n} \log \frac{ed_x}{\overline{p}}} \lesssim \hat{\delta}_n^2.$$

   Choose $\hat{\delta}_n \asymp n^{-\frac{1}{4}} (\log n)^{\frac{1}{2}} k^{\frac{3}{4}} + \sqrt{\frac{k\overline{p}}{n} \log \frac{ed_x}{\overline{p}}}$.

Moreover, for the above choices of $\hat{\delta}_n$, $n\hat{\delta}_n^2 \gtrsim \log(\log \frac{1}{\hat{\delta}_n})$ is satisfied. Let $\overline{\delta}_n$ be the smallest positive solutions to the inequalities $\overline{R}_n(\overline{\delta}_n^*, \overline{G}^*) \leq \frac{\overline{\delta}_n^2}{2C}$. Then if $n\overline{\delta}_n^2 \lesssim \log(\log \frac{1}{\overline{\delta}_n})$, we have $\overline{\delta}_n \lesssim \hat{\delta}_n$. If $n\overline{\delta}_n^2 \gtrsim \log(\log \frac{1}{\overline{\delta}_n})$, we have with probability larger than 0, $\overline{\delta}_n$ is smaller than the smallest positive solution to $\hat{R}_n(\hat{\delta}_n, \overline{G}^*) \leq \frac{\hat{\delta}_n^2}{2C}$ up to some constant (see for example, Proposition 14.25 of Wainwright (2019)). Since the choice of $\hat{\delta}_n$ is independent of $\{x_i\}_{i=1}^n$, we have $\overline{\delta}_n \lesssim \hat{\delta}_n$. Then combined with Assumption C and Theorem 1, we can get the desired conclusion.

Appendix C. Remaining Proofs

C.1. Main Theoretical Results

C.1.1. Proof of Lemma 11

Firstly we state the following lemma for proving the first statement.

Lemma 21 *When $|\log \frac{p(x)}{p_{\mathcal{D}}(x)}| \leq C$, it holds that*

$$\frac{p_{\mathcal{D}}(x)}{p(x)} \left(\log \frac{p_{\mathcal{D}}(x)}{p(x)} \right)^2 \leq (2 + C) \left(\frac{p_{\mathcal{D}}(x)}{p(x)} \log \frac{p_{\mathcal{D}}(x)}{p(x)} - \frac{p_{\mathcal{D}}(x)}{p(x)} + 1 \right).$$

Since $\left\| \sup_{p \in \mathcal{F}_{dd}, \pi \in \mathcal{F}_{prior}} \left| \log \frac{p(x)}{p_{\mathcal{D}}(x)} \right| \right\|_{\psi_\alpha} \leq D$, if we choose $\rho = D(\log n)^{\frac{1}{\alpha}}$ and define

$$\mathcal{A}_1 = \left\{ \sup_{p \in \mathcal{F}_{dd}, \pi \in \mathcal{F}_{prior}} |\log \frac{p(x)}{p_{\mathcal{D}}(x)}| > \rho \right\},$$

then by Chebyshev's inequality, we have $P(\mathcal{A}_1) \leq \frac{2}{n}$. Using $\mathbf{1}_{\mathcal{A}_1}(x)$ to denote the indicator of $\mathcal{A}_1$, and $\mathcal{A}_1^c$ to denote the complementary set of $\mathcal{A}_1$, we can get

$$\mathbb{E}_{p_{\mathcal{D}}(x)} \left[\left(\log \frac{p(x)}{p_{\mathcal{D}}(x)} \right)^2 \right] = \mathbb{E}_{p_{\mathcal{D}}(x)} \left[\left(\log \frac{p(x)}{p_{\mathcal{D}}(x)} \right)^2 \mathbf{1}_{\mathcal{A}_1^c}(x) \right] + \mathbb{E}_{p_{\mathcal{D}}(x)} \left[\left(\log \frac{p(x)}{p_{\mathcal{D}}(x)} \right)^2 \mathbf{1}_{\mathcal{A}_1}(x) \right]. \tag{10}$$

By Lemma 21, we can upper bound the first term of equation (10),

$$\begin{aligned}
&\mathbb{E}_{p_{\mathcal{D}}(x)}\left[\left(\log\frac{p(x)}{p_{\mathcal{D}}(x)}\right)^2 \mathbf{1}_{\mathcal{A}_1^c}(x)\right]\\
&= \mathbb{E}_{p(x)}\left[\frac{p_{\mathcal{D}}(x)}{p(x)}\left(\log\frac{p_{\mathcal{D}}(x)}{p(x)}\right)^2 \mathbf{1}_{\mathcal{A}_1^c(x)}\right]\\
&\leq (2+\rho)\mathbb{E}_{p(x)}\left[\left(\frac{p_{\mathcal{D}}(x)}{p(x)}\log\frac{p_{\mathcal{D}}(x)}{p(x)} - \frac{p_{\mathcal{D}}(x)}{p(x)} + 1\right)\mathbf{1}_{\mathcal{A}_1^c}(x)\right]\\
&\leq (2+\rho)D_{\mathrm{KL}}(p_{\mathcal{D}}(\cdot)||p(\cdot)),
\end{aligned}$$

and for the second term,

$$\begin{aligned}
&\mathbb{E}_{p_{\mathcal{D}}(x)}\left[\left(\log\frac{p(x)}{p_{\mathcal{D}}(x)}\right)^2 \mathbf{1}_{\mathcal{A}_1}(x)\right]\\
&\leq \mathbb{E}_{p_{\mathcal{D}}(x)}\left[\left(\sup_{p\in\mathcal{F}_{dd},\pi\in\mathcal{F}_{prior}} \log^2\frac{p(x)}{p_{\mathcal{D}}(x)}\right)\mathbf{1}_{\mathcal{A}_1}(x)\right]\\
&= \int_0^{+\infty} P\left(\left(\left(\sup_{p\in\mathcal{F}_{dd},\pi\in\mathcal{F}_{prior}} \log^2\frac{p(x)}{p_{\mathcal{D}}(x)}\right)\mathbf{1}_{\mathcal{A}_1}(x)\right) > t\right)dt\\
&\leq \int_0^{D^2(\log n)^{\frac{2}{\alpha}}} P\left(\left(\left(\sup_{p\in\mathcal{F}_{dd},\pi\in\mathcal{F}_{prior}} \log^2\frac{p(x)}{p_{\mathcal{D}}(x)}\right)\mathbf{1}_{\mathcal{A}_1}(x)\right) > 0\right)dt\\
&+ \int_{D^2(\log n)^{\frac{2}{\alpha}}}^{+\infty} P\left(\sup_{p\in\mathcal{F}_{dd},\pi\in\mathcal{F}_{prior}} \log^2\frac{p(x)}{p_{\mathcal{D}}(x)} > t\right)dt\\
&\leq 2D^2\frac{\log^{\frac{2}{\alpha}} n}{n} + \int_{D^2(\log n)^{\frac{2}{\alpha}}}^{+\infty} P\left(\sup_{p\in\mathcal{F}_{dd},\pi\in\mathcal{F}_{prior}}\left|\log\frac{p(x)}{p_{\mathcal{D}}(x)}\right| > \sqrt{t}\right)dt\\
&\leq 2D^2\frac{\log^{\frac{2}{\alpha}} n}{n} + \int_{D^2(\log n)^{\frac{2}{\alpha}}}^{+\infty} 2\exp\left(-\left(\frac{t^{\frac{1}{2}}}{D}\right)^{\alpha}\right)dt\\
&= 2D^2\frac{\log^{\frac{2}{\alpha}} n}{n} + \frac{4}{\alpha}D^2\int_{\log n}^{\infty}\exp(-x)x^{\frac{2}{\alpha}-1}dx\\
&\lesssim D^2\frac{\log^{\frac{2}{\alpha}} n}{n}.
\end{aligned}$$

Therefore $\mathbb{E}_{p_{\mathcal{D}}(x)}\left[\left(\log\frac{p_{\mathcal{D}}(x)}{p(x)}\right)^2\right] \leq c_1\left((D\log^{\frac{1}{\alpha}} n)D_{\mathrm{KL}}(p_{\mathcal{D}}(\cdot)||p(\cdot)) + \frac{D^2\log^{\frac{2}{\alpha}} n}{n}\right)$.

For the second statement, define $\mathcal{A}_2 = \left\{\sup_{p\in\mathcal{F}_{dd},q\in\mathcal{F}_{ed},\pi\in\mathcal{F}_{prior}} D_{\mathrm{KL}}(q(\cdot|x)||p(\cdot|x)) \geq \rho\right\}$. By the fact that

$$\begin{aligned}
&\mathbb{E}_{p_{\mathcal{D}}(x)}\left[(D_{\mathrm{KL}}(q(\cdot|x)||p(\cdot|x)) - D_{\mathrm{KL}}(q'(\cdot|x)||p'(\cdot|x)))^2\right] \\
&\leq 4\rho\mathbb{E}_{p_{\mathcal{D}}(x)}\left[\left(\sqrt{D_{\mathrm{KL}}(q(\cdot|x)||p(\cdot|x))} - \sqrt{D_{\mathrm{KL}}(q'(\cdot|x)||p'(\cdot|x))}\right)^2 \mathbf{1}_{\mathcal{A}_2^c}(x)\right] \\
&+ 4\mathbb{E}_{p_{\mathcal{D}}(x)}\left[\left(\sup_{p\in\mathcal{F}_{dd}, q\in\mathcal{F}_{ed}, \pi\in\mathcal{F}_{prior}} D_{KL}^2(q(\cdot|x)||p(\cdot|x))\right)\mathbf{1}_{\mathcal{A}_2}(x)\right].
\end{aligned}$$

We can get the desired conclusion using the same argument as the first statement.

C.2. Parametric Models

C.2.1. PROOF OF LEMMA 13

Since $\sup_{g\in\overline{G}^*}|g(x)| = \sup_{g\in\overline{G}^*}|g(x)|$, we have $\left\|\sup_{g\in\overline{G}^*}|g(x)|\right\|_{\psi_\alpha} \leq 2D$.

Choose

$$\begin{aligned}
\rho &= 8\mathbb{E}_{p_{\mathcal{D}}(x)}\max_{1\leq i\leq n}\sup_{g\in\overline{G}^*}|g(x_i)| \\
&\leq K_\alpha\left\|\max_{1\leq i\leq n}\sup_{g\in\overline{G}^*}|g(x_i)|\right\|_{\psi_\alpha}.
\end{aligned}$$

Since $\left\|\max_{1\leq i\leq n}\sup_{g\in\overline{G}^*}|g(x_i)|\right\|_{\psi_\alpha} \leq K_\alpha\left\|\sup_{g\in\overline{G}^*}|g(x)|\right\|_{\psi_\alpha}\log^{\frac{1}{\alpha}} n$ (see for example, equation (13) of Adamczak (2008)), it holds that $\rho \lesssim D\log^{\frac{1}{\alpha}} n$. Define $\mathcal{A} = \left\{\sup_{g\in\overline{G}^*}|g(x)| \leq \rho\right\}$, we have

$$\begin{aligned}
&\mathbb{E}_{p_{\mathcal{D}}(x)}\left[\sup_{\substack{g\in\overline{G}^* \\ \|g\|_2\leq\delta}}\left|\frac{1}{n}\sum_{i=1}^n g(x_i) - \mathbb{E}_{p_{\mathcal{D}}(x)}g(x)\right|\right] \\
&\leq \mathbb{E}_{p_{\mathcal{D}}(x)}\left[\sup_{\substack{g\in\overline{G}^* \\ \|g\|_2\leq\delta}}\left|\frac{1}{n}\sum_{i=1}^n g(x_i)\mathbf{1}_{\mathcal{A}}(x_i) - \mathbb{E}_{p_{\mathcal{D}}(x)}(g(x)\mathbf{1}_{\mathcal{A}}(x))\right|\right] \\
&+ \mathbb{E}_{p_{\mathcal{D}}(x)}\left[\sup_{g\in\overline{G}^*}\left|\frac{1}{n}\sum_{i=1}^n g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i) - \mathbb{E}_{p_{\mathcal{D}}(x)}(g(x)\mathbf{1}_{\mathcal{A}^c}(x))\right|\right],
\end{aligned}$$

where we can further upper bounded the second term,

$$\mathbb{E}_{p_{\mathcal{D}}(x)}\left[\sup_{g\in\overline{G}^*}\left|\frac{1}{n}\sum_{i=1}^n g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i) - \mathbb{E}_{p_{\mathcal{D}}(x)}(g(x)\mathbf{1}_{\mathcal{A}^c}(x))\right|\right] \leq 2\mathbb{E}_{p_{\mathcal{D}}(x)}\left[\sup_{g\in\overline{G}^*}\left|\frac{1}{n}\sum_{i=1}^n g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)\right|\right].$$

Let $\tau = \inf\{j \leq n : \sup\limits_{g \in \overline{G}^*} |\sum_{i=1}^{j} g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)| > t\}$. Under $\tau = j$,

$$\begin{aligned}
&\max_{k\leq n}\sup_{g\in\overline{G}^*}|\sum_{i=1}^{k} g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)| \\
&\leq t + \max_{1\leq i\leq n}\sup_{g\in\overline{G}^*}|g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)| + \max_{j<k\leq n}\sup_{g\in\overline{G}^*}|\sum_{i=j+1}^{k} g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)|
\end{aligned}$$

Since $\{\tau = j\}$ only depend on $x_1, \cdots, x_j$, we have

$$\begin{aligned}
&P(\tau = j, \max_{k\leq n}\sup_{g\in\overline{G}^*}|\sum_{i=1}^{k} g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)| > 3t+s) \\
&\leq P(\tau = j, \max_{1\leq i\leq n}\sup_{g\in\overline{G}^*}|g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)| > s) + P(\tau = j)P(\max_{j<k\leq n}\sup_{g\in\overline{G}^*}|\sum_{i=j+1}^{k} g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)| > 2t) \\
&\leq P(\tau = j, \max_{1\leq i\leq n}\sup_{g\in\overline{G}^*}|g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)| > s) + P(\tau = j)P(\max_{k\leq n}\sup_{g\in\overline{G}^*}|\sum_{i=1}^{k} g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)| > t)
\end{aligned}$$

Where the last inequality is due to the fact that for any $1 \leq j < n$, $\sup\limits_{g\in\overline{G}^*}|\sum_{i=j+1}^{k} g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)| \leq 2\max\limits_{k\leq n}\sup\limits_{g\in\overline{G}^*}|\sum_{i=1}^{k} g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)|$. A summation over $j = 1, \cdots, n$ yields

$$\begin{aligned}
P(\max_{k\leq n}\sup_{g\in\overline{G}^*}|\sum_{i=1}^{k} g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)| > 3t+s) &\leq P(\max_{1\leq i\leq n}\sup_{g\in\overline{G}^*}|g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)| > s) \\
&\quad + P^2(\max_{k\leq n}\sup_{g\in\overline{G}^*}|\sum_{i=1}^{k} g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)| > t)
\end{aligned}$$

So,

$$\begin{aligned}
&\mathbb{E}_{p_{\mathcal{D}}(x)}\sup_{g\in\overline{G}^*}\left|\frac{1}{n}\sum_{i=1}^{n} g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)\right| \\
&\leq \frac{1}{n}\mathbb{E}_{p_{\mathcal{D}}(x)}\max_{k\leq n}\sup_{g\in\overline{G}^*}\left|\sum_{i=1}^{k} g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)\right| \\
&= \frac{4}{n}\int_0^{+\infty} P(\max_{k\leq n}\sup_{g\in\overline{G}^*}\left|\sum_{i=1}^{k} g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)\right| > 4t)dt \\
&\leq \frac{4}{n}\int_0^{+\infty} P^2(\max_{k\leq n}\sup_{g\in\overline{G}^*}\left|\sum_{i=1}^{k} g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)\right| > t)dt + \frac{4}{n}\int_0^{+\infty} P(\max_{1\leq i\leq n}\sup_{g\in\overline{G}^*}|g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)| > t)dt \\
&\leq \frac{4}{n}P\left(\max_{k\leq n}\sup_{g\in\overline{G}^*}\left|\sum_{i=1}^{k} g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)\right| > 0\right)\mathbb{E}_{p_{\mathcal{D}}(x)}\max_{k\leq n}\sup_{g\in\overline{G}^*}\left|\sum_{i=1}^{k} g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)\right| \\
&+ \frac{4}{n}\mathbb{E}_{p_{\mathcal{D}}(x)}\max_{1\leq i\leq n}\sup_{g\in\overline{G}^*}|g(x_i)|
\end{aligned}$$

By Markov inequality,

$$P\left(\max_{k\leq n}\sup_{g\in\overline{G}^*}\left|\sum_{i=1}^{k} g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)\right|>0\right)$$
$$\leq P(\max_{1\leq i\leq n}\sup_{g\in\overline{G}^*}|g(x_i)|>\rho)\leq\frac{1}{8}$$
$$\mathbb{E}_{p_{\mathcal{D}}(x)}\max_{1\leq i\leq n}\sup_{g\in\overline{G}^*}|g(x_i)|\lesssim D\log^{\frac{1}{\alpha}} n$$

So, we have $\mathbb{E}_{p_{\mathcal{D}}(x)}\sup\limits_{g\in\overline{G}^*}\left|\frac{1}{n}\sum_{i=1}^{n} g(x_i)\mathbf{1}_{\mathcal{A}^c}(x_i)\right|\lesssim D\frac{\log^{\frac{1}{\alpha}} n}{n}$.

C.2.2. PROOF OF LEMMA 14

Since $r_n\leq 2\rho$ and $r_n\leq 2\sqrt{\frac{d^*}{n}\sum_{i=1}^{n} b^2(x_i)}$,

$$\int_0^{r_n}\sqrt{d^*\log\left(\frac{3\sqrt{\frac{d^*}{n}\sum_{i=1}^{n} b^2(x_i)}}{\varepsilon}\right)+\log\frac{2\rho}{\varepsilon}d\varepsilon}$$
$$\leq\int_0^{r_n}\sqrt{d^*\log\left(\frac{3\sqrt{\frac{d^*}{n}\sum_{i=1}^{n} b^2(x_i)}}{\varepsilon}\right)}+\sqrt{log\frac{2\rho}{\varepsilon}}d\varepsilon$$
$$=r_n\int_0^{1}\sqrt{d^*\log\left(\frac{3\sqrt{\frac{d^*}{n}\sum_{i=1}^{n} b^2(x_i)}}{r_n\varepsilon}\right)}+\sqrt{log\frac{2\rho}{r_n\varepsilon}}d\varepsilon$$
$$\leq r_n\sqrt{d^*\log\left(\frac{3\sqrt{\frac{d^*}{n}\sum_{i=1}^{n} b^2(x_i)}}{r_n}\right)}+r_n\sqrt{\log\frac{2\rho}{r_n}}+r_n\left(\sqrt{d^*}+1\right)\int_0^1\sqrt{\log\frac{1}{\varepsilon}}d\varepsilon.$$

By $\log x\leq\frac{1}{e}x$

$$\mathbb{E}_{p_{\mathcal{D}}(x)}\left[r_n\sqrt{\log\frac{3\sqrt{\frac{d^*}{n}\sum_{i=1}^{n} b^2(x_i)}}{r_n}}\right]$$
$$=\mathbb{E}_{p_{\mathcal{D}}(x)}\left[r_n\sqrt{\log\frac{2\rho\sqrt{d^*}}{r_n}+\log\frac{3\sqrt{\frac{1}{n}\sum_{i=1}^{n} b^2(x_i)}}{2\rho}}\right]$$
$$\leq\mathbb{E}_{p_{\mathcal{D}}(x)}\left[r_n\sqrt{\log\frac{2\rho\sqrt{d^*}}{r_n}}+r_n\sqrt{\frac{1}{e}\frac{3\sqrt{\frac{1}{n}\sum_{i=1}^{n} b^2(x_i)}}{2\rho}}\right].$$

Since $b(x) \in \mathcal{L}^2(p_{\mathcal{D}}(x))$, by Cauchy-Schwarz inequality, we have

$$\mathbb{E}_{p_{\mathcal{D}}(x)}\left[r_n\sqrt{\frac{1}{e}\frac{3\sqrt{\frac{1}{n}\sum_{i=1}^n b^2(x_i)}}{2\rho}}\right] \lesssim \sqrt{\mathbb{E}_{p_{\mathcal{D}}(x)}\frac{r_n^2}{\rho}},$$

$$\mathbb{E}_{p_{\mathcal{D}}(x)}\left[r_n\sqrt{\log\frac{2\rho}{r_n}}\right] \leq 2\rho\mathbb{E}_{p_{\mathcal{D}}(x)}\left[\frac{r_n}{2\rho}\sqrt{\log\frac{2\rho}{r_n}+\frac{1}{2}}\right].$$

Let $\left(\frac{r_n}{2\rho}\right)^2 = y(y \leq 1)$, since $\sqrt{-\frac{1}{2}y\log y + \frac{1}{2}y}$ is concave and non-decreasing when $y \leq 1$, by Jensen inequality, we have

$$\mathbb{E}_{p_{\mathcal{D}}(x)}\left[\frac{r_n}{2\rho}\sqrt{\log\frac{2\rho}{r_n}+\frac{1}{2}}\right] \leq \sqrt{-\frac{1}{2}\mathbb{E}_{p_{\mathcal{D}}(x)}\left[\left(\frac{r_n}{2\rho}\right)^2\log\mathbb{E}_{p_{\mathcal{D}}(x)}\left(\frac{r_n}{2\rho}\right)^2\right]+\frac{1}{2}\mathbb{E}_{p_{\mathcal{D}}(x)}\left(\frac{r_n}{2\rho}\right)^2}.$$

Combine with the fact $\int_0^1\sqrt{\log\frac{1}{\varepsilon}}d\varepsilon$ is less than infinity, we can get the desired conclusion.

C.3. Gaussian Encoders and Decoders

C.3.1. PROOF OF LEMMA 17

We first state the following lemmas about error bounds for approximations with deep ReLU networks stated in Yarotsky (2017).

Lemma 22 (Theorem 1 of Yarotsky (2017)) *There is a deep feedforward ReLU network architecture with depth at most $c(\log(1/\epsilon)+1)$, the absolute value of each weight unit at most ϵ^{-c_1} and weights and computation units at most $c_2\epsilon^{-\frac{d}{\alpha}}\log(1/\epsilon)+1$ that is capable of expressing any function belong to $C_1^\alpha([0,1]^d)$ with error ϵ.*

Lemma 23 (Proposition 3 of Yarotsky (2017)) *Given $M > 0$ and $\epsilon \in (0,1)$, there is a feedforward ReLU network η with two input units that implements a function $\widetilde{\mathrm{x}} : \mathbb{R}^2 \to \mathbb{R}$ so that*

1. *For any inputs x, y, if $|x| \leq M$ and $|y| \leq M$, then $|\widetilde{x}(x,y) - xy| \leq \epsilon$;*
2. *if $x = 0$ or $y = 0$, then $\tilde{x}(x,y) = 0$;*
3. *The depth and the number of weights and computation units in η is not greater than $c_1\ln(1/\epsilon)+c_2$ with an absolute constant c_1 and a constant $c_2 = c_2(M)$.*

W.l.o.g, we can assume $\eta = \overline{\eta} = 1$ and $\gamma = 1$. Then we consider $\mathbf{m} = (m_1, \cdots, m_d) \in \{-N, -(N-1)\cdots, 0, 1, \cdots, N\}^{d_z}$, $z_{\mathbf{m}}^i = \frac{\sqrt{\log\frac{1}{\sigma^*}}m_i}{N}$ and $z_{\mathbf{m}} = (z_{\mathbf{m}}^1, \cdots, z_{\mathbf{m}}^{d_z})$. By the Lipschitzness of $G_{\mathcal{D}}(z)$ and $Q_{\mathcal{D}}(x)$, there exists a constant c_1 such that for any $(z, z') \in B_z = \left[-\sqrt{\log\frac{1}{\sigma^*}}, \sqrt{\log\frac{1}{\sigma^*}}\right]^{d_z}$, it holds that

$$\|G_{\mathcal{D}}(z) - G_{\mathcal{D}}(z')\|_2 \leq c_1 \|z - z'\|_2 \left(\log \frac{1}{\sigma^*}\right)^{\frac{1}{\alpha}};$$
$$\|z - z'\|_2 \leq c_1 \|G_{\mathcal{D}}(z) - G_{\mathcal{D}}(z')\|_2 \left(\log \frac{1}{\sigma^*}\right)^{\frac{2}{\alpha^2}}.$$

Then for $\mathbf{m} \in \{-N, -(N-1) \cdots, 0, 1, \cdots, N\}^{d_z}$, we define

$$\phi_{\mathbf{m}}(x) = \prod_{i=1}^{d_x} \psi \left(\frac{1}{\sigma^* \sqrt{d_x \log \frac{1}{\sigma^*}} + c_1 \sqrt{d_z} \frac{1}{2N} \left(\log \frac{1}{\sigma^*}\right)^{\frac{1}{\alpha}+\frac{1}{2}}} (x_i - G_{\mathcal{D},i}(z_{\mathbf{m}})) \right),$$

where $\psi(x) = \begin{cases} 1, & |x| < 1 \\ 0, & 2 < |x| \\ 2 - |x|, & 1 \leq |x| \leq 2 \end{cases}$ and x_i, $G_{\mathcal{D},i}(z_{\mathbf{m}})$ denote the i-th dimension of x and $G_{\mathcal{D}}(z_{\mathbf{m}})$. For any $z \in B_z$, there exist a $\mathbf{m} \in \{-N, -(N-1) \cdots, 0, 1, \cdots, N\}^{d_z}$, such that

$$\|z - z_{\mathbf{m}}\|_2 \leq \frac{\sqrt{\log \frac{1}{\sigma^*} d_z}}{2N} \quad \|G_{\mathcal{D}}(z) - G_{\mathcal{D}}(z_{\mathbf{m}})\|_2 \leq c_1 \sqrt{d_z} \frac{1}{2N} \left(\log \frac{1}{\sigma^*}\right)^{\frac{1}{\alpha}+\frac{1}{2}}.$$

Therefore for any $x \in B_x = \{x = G_{\mathcal{D}}(z) + \sigma^* \epsilon \,|\, z \in B_z,\, \epsilon \in B_\epsilon\}$, there exists $z_{\mathbf{m}}$, such that $\|x - G_{\mathcal{D}}(z_{\mathbf{m}})\|_2 \leq c_1 \sqrt{d_z} \frac{1}{2N} \left(\log \frac{1}{\sigma^*}\right)^{\frac{1}{\alpha}+\frac{1}{2}} + \sigma^* \sqrt{d_x \log \frac{1}{\sigma^*}}$. It follows that for any $x \in B_x$, $\sum\limits_{\mathbf{m}} \phi_{\mathbf{m}}(x) \geq 1$. Moreover, by the fact that the support of $\phi_{\mathbf{m}}(x)$ is

$$\left\{ x : |x_i - G_{\mathcal{D},i}(z_{\mathbf{m}})| \leq 2(\sigma^* \sqrt{d_x \log \frac{1}{\sigma^*}} + c_1 \sqrt{d_z} \frac{1}{2N} \left(\log \frac{1}{\sigma^*}\right)^{\frac{1}{\alpha}+\frac{1}{2}}), \forall\, 1 \leq i \leq d_x \right\}.$$

We have $\sum\limits_{\mathbf{m}} \phi_{\mathbf{m}}(x) \lesssim \left(\log \frac{1}{\sigma^*}\right)^{\frac{2d_z}{\alpha^2}+\frac{d_z}{\alpha}} + (\sigma^* N)^{d_z} \left(\log \frac{1}{\sigma^*}\right)^{\frac{2d_z}{\alpha^2}}$. Let

$$P_{\mathbf{m},j}(x) = \sum_{\gamma:|\gamma|<k} \left.\frac{D^\gamma Q_{\mathcal{D},j}}{\gamma!}\right|_{x=G_{\mathcal{D}}(z_{\mathbf{m}})} (x - G_{\mathcal{D}}(z_{\mathbf{m}}))^\gamma,$$

with the usual conventions $\gamma! = \prod_{i=1}^{d_x} \gamma^i$, $(x - G_{\mathcal{D}}(z_{\mathbf{m}}))^\gamma = \prod_{i=1}^{d_x} (x_i - G_{\mathcal{D},i}(z_{\mathbf{m}}))^{\gamma_k}$ and $Q_{\mathcal{D},j}$ $(1 \leq j \leq d_z)$ being the jth dimension of $Q_{\mathcal{D}}$. Now define an approximation to $Q_{\mathcal{D}}(x)$ by

$$\tilde{Q}_j(x) = \frac{\sum\limits_{\mathbf{m}} \phi_{\mathbf{m}}(x) P_{\mathbf{m},j}(x)}{\sum\limits_{\mathbf{m}} \phi_{\mathbf{m}}(x)};$$
$$\tilde{Q}(x) = (\tilde{Q}_1(x), \cdots, \tilde{Q}_{d_z}(x)).$$

We have for any $x \in B_x$ and $1 \leq j \leq d_z$,

$$\begin{aligned} |\tilde{Q}_j(x) - Q_{\mathcal{D},j}(x)| &\leq |\sum_{\mathbf{m}} \phi_{\mathbf{m}}(x) P_{\mathbf{m},j}(x) - \sum_{\mathbf{m}} \phi_{\mathbf{m}}(x) Q_{\mathcal{D},j}(x)| \\ &\lesssim \left(\log \frac{1}{\sigma^*}\right)^{\frac{2d_z+2}{\alpha^2}+\frac{d_z+k}{\alpha}+\frac{k}{2}} (1 + (\sigma^* N)^{d_z})(\sigma^* + \frac{1}{N})^k. \end{aligned}$$

Furthermore, by the fact that $x = \max(x, 0) - \max(-x, 0)$, one has for any feedforward ReLU neural work with d input units, depth L, $\|U\|_1$ computation units and W weights, it can be expressed as a feedforward ReLU neural work in which information can only come from the previous one layer with depth L, computation units at most $2L(\|U\|_1 + d)$ and weights at most $4W + 2L(\|U\|_1 + d)$. Choose

$$N \asymp \left(\frac{\sigma_1}{\left(\log \frac{1}{\sigma_1}\right)^{\frac{2d_z+2}{\alpha^2} + \frac{d_z+k}{\alpha} + \frac{k}{2}}} \right)^{-\frac{1}{k}}.$$

By the fact that

$$|x| = \max(x, 0) + \max(-x, 0),$$
$$\max(0, \min(1, x)) = \max(-\max(-x + 1, 0) + 1, 0),$$

and when $x \geq 1$, $f(x) = \frac{1}{x}$ is C^∞, combined with lemma 22, lemma 23 and when $x \in B_x$, $\sum_{\mathbf{m}} \phi_{\mathbf{m}}(x) \geq 1$, we can conclude that there exist ReLU neural networks with depth $O(\log \frac{1}{\sigma_1})$ and weights and computation units at most $O(N^{d_z}(\log \frac{1}{\sigma_1}))$ that approximate $Q_1(x)$ with error at most $\left(\log \frac{1}{\sigma^*}\right)^{\frac{2d_z}{\alpha^2} + \frac{d_z+k}{\alpha} + \frac{k}{2}} (1 + (\sigma^* N)^{d_z})(\sigma^* + \frac{1}{N})^k$ in domain B_x. Since $k \geq 2$, we can choose $W_1, \|U_1\|_1 \asymp N^{d_z}(\log \frac{1}{\sigma_1})^2$, $L_1 \asymp \log \frac{1}{\sigma_1}$ and a large enough constant b_8, such that $\epsilon_0 \leq \frac{1}{3}\sigma^*$. Checking the proof of Lemma 22 and Lemma 23 (Theorem 1 and Proposition 3 of Yarotsky (2017)), we can choose $\|U\|_1, W \asymp \left(\log \frac{1}{\sigma_1}\right)^{\frac{d_z}{\alpha k} + \frac{d_z}{2} + 2} ({\sigma_1}^2)^{-\frac{d_z}{k}}$; $\|U_2\|_1, W_2 \asymp \left(\log \frac{1}{\sigma_1}\right)^{\frac{d_z}{\alpha(k-1)} + \frac{d_z}{2} + 2} (\sigma_1)^{-\frac{d_z}{k-1}}$; $L, L_2 \asymp \log \frac{1}{\sigma_1}$; $V = \left(\frac{1}{\sigma_1}\right)^c$ with a large enough c and (b, b_9) to be large enough constants, such that $\epsilon_1 \leq \frac{1}{3}\sigma^{*2}$ and $\epsilon_2 \leq \frac{1}{3}\sigma^*$.

C.3.2. PROOF OF LEMMA 18

To begin with, we state the following lemma in Anthony and Bartlett (1999) for bounding the covering number of ReLU Neural networks.

Lemma 24 (Theorem 12.2 of Anthony and Bartlett (1999)) *Assume for all $f \in \mathcal{F}$, $\|f\|_\infty \leq M$. Denote the pseudo-dimension of $\mathcal{F}$ as $\mathrm{Pdim}(\mathcal{F})$, then for $n \geq \mathrm{Pdim}(\mathcal{F})$, we have for any ϵ and any $X_1, \ldots, X_n$*

$$\mathbf{N}\left(\epsilon, \mathcal{F}|_{X_1,\ldots,X_n}, \infty\right) \leq \left(\frac{2eM \cdot n}{\epsilon \cdot \mathrm{Pdim}(\mathcal{F})}\right)^{\mathrm{Pdim}(\mathcal{F})}.$$

By the choice of $\mu_\phi(x)$ and $\Sigma_\phi(x)$, we have $\mu_\phi(x) \leq \log^{c_0} \frac{1}{\sigma_1}(c_1\|x\|_2 + c_2)$ and $0 < c_3 \log^{-c_0} \frac{1}{\sigma_1} \sigma_1^2 \leq \lambda_{\min}(\Sigma_\phi(x)) \leq \lambda_{\max}(\Sigma_\phi(x)) \leq 1$ with some constants (c_0, c_1, c_2, c_3). Also,

$$\begin{aligned}
\log p_{\mathcal{D}}(x) &\leq \frac{d_x}{2} \log(\frac{1}{2\pi\sigma^{*2}}); \\
-\log p_{\mathcal{D}}(x) &= \frac{d_x}{2} \log(2\pi\sigma^{*2}) - \log \int \exp\left(-\frac{(x - G_{\mathcal{D}}(z))^T(x - G_{\mathcal{D}}(z))}{2\sigma^{*2}}\right) \pi_{\mathcal{D}}(z) dz \qquad (11) \\
&\leq \frac{d_x}{2} \log(2\pi\sigma^{*2}) + \frac{\|x\|_2^2}{\sigma^{*2}} + \frac{1}{\sigma^{*2}} \int \|G_{\mathcal{D}}(z)\|_2^2 \pi_{\mathcal{D}}(z) dz.
\end{aligned}$$

We then state the following Lemma 25 for bounding the Orlicz norm in Assumption A.

Lemma 25 *Consider $\mathcal{F}_{dd}$ and $\mathcal{F}_{ed}$ defined in equation (9), given Assumption B and Condition B, there exists a constant C_0 that only depend on $(\alpha, k, b_2, b_3, b_5, b_6, d_z, d_x)$ in Assumption B and Condition B, such that*

$$\sup_{\substack{p_\theta \in \mathcal{F}_{dd}, q_\phi \in \mathcal{F}_{ed} \\ \pi_\beta \in \mathcal{F}_{prior}}} \left(|\log p_{\theta,\beta}(x)| + D_{\mathrm{KL}}(q_\phi(\cdot|x) || p_{\theta,\beta}(\cdot|x)) \right) \le \frac{C_0}{\sigma_1^2} \left(\|x\|_2^2 + (\log \frac{1}{\sigma_1})^{\frac{2}{\alpha}} \right).$$

For any $G_{\theta_1}(z) \in \mathcal{F}_{d_z}^{d_x}(L, U, W, b(\log \frac{1}{\sigma_1})^{\frac{1}{\alpha}}, V)$ with $U = (u_1, \cdots, u_{L-2})$, it can be expressed as a fully-connected ReLU neural network with depth L, computation units $\|U\|_1$, and Frobenius norm of weights in each layer at most $V\|U\|_1$. We use θ_1^F to denote the $(d_z+1)u_1 + \sum_{l=1}^{L-3}(u_l + 1)u_{l+1} + d_x(u_{L-2}+1)$ dimensional weights vector of $G_{\theta_1}(z)$ after expressed as a fully-connected ReLU neural network, then it can only has at most W-number of nonzero elements. Consider $\mathcal{B}_x = [-c_1 \log^{\frac{1}{\alpha}} n, c_1 \log^{\frac{1}{\alpha}} n]^{d_x}$ such that $p_{\mathcal{D}}(x \notin \mathcal{B}_x) \le \frac{1}{n^2}$. Next we state a lemma about the lipschitzness of $m(p_\theta, \pi_\beta, q_\phi, x)$ on $\mathcal{B}_x$.

Lemma 26 *Consider $\mathcal{F}_{dd}$ and $\mathcal{F}_{ed}$ defined in (9), given Assumption B and condition B, there exist some constants (c_0, c_1) that only depend on $(\alpha, k, b_2, b_3, b_5, b_6, d_z, d_x)$ in Assumption B and Condition B, such that for any $x \in \mathcal{B}_x$, $(p_\theta, p_{\theta'}) \in \mathcal{F}_{dd}$, $(q_\phi, q_{\phi'}) \in \mathcal{F}_{ed}$ and $(\pi_\beta, \pi_{\beta'}) \in \mathcal{F}_{prior}$,*

$$\begin{aligned}
&\big| m(p_\theta, \pi_\beta, q_\phi, x) - m(p_{\theta'}, \pi_{\beta'}, q_{\phi'}, x) \big| \\
&\le c_2 \log^{c_3} n \frac{1}{\sigma_1^{c_4}} \Big(\sqrt{L} (\|U\|_1 V)^{L-2} \|\theta_1^F - \theta_1^{F'}\|_2 + \|\beta - \beta'\|_2 + \big| \frac{1}{\sigma^2} - \frac{1}{\sigma'^2} \big| \\
&+ (\|U\|_1 V)^{L-1} (\|\mu_\phi(x) - \mu_{\phi'}(x)\|_2 + \|\Sigma_\phi(x) - \Sigma_{\phi'}(x)\|_F) \Big).
\end{aligned}$$

Therefore by

$$\begin{aligned}
&\Sigma_\phi(x) - \Sigma_{\phi'}(x) \\
&= \bar{\sigma}^2 \left(G_{\phi_2}^d(Q_{\phi_1}(x))^T G_{\phi_2}^d(Q_{\phi_1}(x)) + \bar{\sigma}^2 I_{d_z} \right)^{-1} \\
&- \bar{\sigma}^2 \left(G_{\phi_2'}^d(Q_{\phi_1'}(x))^T G_{\phi_2'}^d(Q_{\phi_1'}(x)) + \bar{\sigma'}^2 I_{d_z} \right)^{-1} \\
&+ \bar{\sigma}^2 \left(G_{\phi_2'}^d(Q_{\phi_1'}(x))^T G_{\phi_2'}^d(Q_{\phi_1'}(x)) + \bar{\sigma'}^2 I_{d_z} \right)^{-1} \\
&- \bar{\sigma'}^2 \left(G_{\phi_2'}^d(Q_{\phi_1'}(x))^T G_{\phi_2'}^d(Q_{\phi_1'}(x)) + \bar{\sigma'}^2 I_{d_z} \right)^{-1}.
\end{aligned}$$

And by the fact that

$$\begin{aligned}
\|A^{-1} - A'^{-1}\|_F &= \|A'^{-1} A' A^{-1} - A'^{-1} A A^{-1}\|_F \\
&= \|A'^{-1}(A' - A)A^{-1}\|_F \\
&\le \|A'^{-1}\|_F \|A' - A\|_F \|A^{-1}\|_F.
\end{aligned}$$

We can get that there exists a constant c_0 such that,

$$\|\Sigma_\phi(x) - \Sigma_{\phi'}(x)\|_F \lesssim \frac{\log^{c_0} \frac{1}{\sigma_1}}{\sigma_1^2} \left(\|G_{\phi_2}^d(Q_{\phi_1}(x)) - G_{\phi_2'}^d(Q_{\phi_1'}(x))\|_F + |\bar{\sigma}^2 - (\bar{\sigma}')^2| \right).$$

Moreover,

$$
\begin{aligned}
&\|\mu_\phi(x)-\mu'_\phi(x)\|_2\\
&\le \|Q_{\phi_1}(x)-Q_{\phi'_1}(x)\|_2+\log^{\frac{1}{\alpha}} n\log^{c_0}\frac{1}{\sigma_1}(\|\Sigma_\phi(x)-\Sigma_{\phi'}(x)\|_F+\|G^d_{\phi_2}(Q_{\phi_1}(x))-G^d_{\phi'_2}(Q_{\phi'_1}(x))\|_F)\\
&+\|G_{\phi_3}(Q_{\phi_1}(x))-G_{\phi'_3}(Q_{\phi'_1}(x))\|_2.
\end{aligned}
$$

So we can obtain that under $\mathcal{B}_x$, there exist some constant (c_5, c_6, c_7) such that

$$
\begin{aligned}
&\big|m(p_\theta,\pi_\beta,q_\phi,x)-m(p_{\theta'},\pi_{\beta'},q_{\phi'},x)\big|\\
&\le c_5\log^{c_6} n\frac{1}{\sigma_1^{c_7}}\Big(\|\beta-\beta'\|_2+\sqrt{L}(\|U\|_1V)^{L-1}\|\theta_1^F-\theta_1^{F'}\|_2+(\|U\|_1V)^L\big(\|Q_{\phi_1}(x)-Q_{\phi'_1}(x)\|_2\\
&+\|G_{\phi_3}(Q_{\phi_1}(x))-G_{\phi'_3}(Q_{\phi'_1}(x))\|_2+\|G^d_{\phi_2}(Q_{\phi_1}(x))-G^d_{\phi'_2}(Q_{\phi'_1}(x))\|_F+|\bar\sigma^2-(\bar\sigma')^2|\big)+\left|\frac{1}{\sigma^2}-\frac{1}{\sigma'^2}\right|\Big).
\end{aligned}
$$

Denote $\Theta^F_{\theta_1}$ as the parameter space of θ_1^F, it holds that

$$
\Theta^F_{\theta_1}\subseteq\Big\{\theta_1^F\in\mathbb{R}^{(d_z+1)u^1+\sum_{l=1}^{L-3}(u_l+1)u_{l+1}+d_x(u_{L-2}+1)}\mid\|\theta_1^F\|_0\le W,\ \|\theta_1^F\|_\infty\le V\Big\}.
$$

So one has $\log\mathbf{N}(\Theta^F_{\theta_1},\ell_2,\epsilon)\lesssim W\log\frac{VL\|U\|_1}{\epsilon}$. Recall $\mathcal{B}_x=[-c_1\log^{\frac{1}{\alpha}}n, c_1\log^{\frac{1}{\alpha}}n]^{d_x}$, we have

$$
P\left(\bigcup_{1\le i\le n}\{x_i\in\mathcal{B}_x^c\}\right)\le\frac{1}{n^2}n=\frac{1}{n}.
$$

Moreover, by Lemma 25, for any $x\in\mathcal{B}_x$, it holds that

$$
\begin{aligned}
&\sup_{\substack{p_\theta\in\mathcal{F}_{dd},q_\phi\in\mathcal{F}_{ed}\\ \pi_\beta\in\mathcal{F}_{prior}}}|m(p_\theta,q_\phi,\pi_\beta,x)-\log p_{\mathcal{D}}(x)|\lesssim\frac{(\log\frac{n}{\sigma_1})^{\frac{2}{\alpha}}}{\sigma_1^2};\\
&\sup_{g\in\overline{G}^*}|g(x)|\lesssim\frac{(\log\frac{n}{\sigma_1})^{\frac{2}{\alpha}}}{\sigma_1^2}.
\end{aligned}
$$

Then, by changing the set $\mathcal{A}$ to $\mathcal{B}_x$ in the proof of Lemma 13, we can get

$$
\begin{aligned}
&\mathbb{E}_{p_{\mathcal{D}}(x)}\left[\sup_{\substack{g\in\overline{G}^*\\ \|g\|_2\le\delta}}\left|\frac{1}{n}\sum_{i=1}^n g(x_i)-\mathbb{E}_{p_{\mathcal{D}}(x)}g(x)\right|\right]\\
&\le\mathbb{E}_{p_{\mathcal{D}}(x)}\left[\sup_{\substack{g\in\overline{G}^*\\ \|g\|_2\le\delta}}\left|\frac{1}{n}\sum_{i=1}^n g(x_i)\mathbf{1}_{\mathcal{B}_x}(x_i)-\mathbb{E}_{p_{\mathcal{D}}(x)}(g(x)\mathbf{1}_{\mathcal{B}_x}(x))\right|\right]+C\frac{(\log\frac{n}{\sigma_1})^{\frac{2}{\alpha}}}{n\sigma_1^2}.
\end{aligned}
$$

Therefore by lemma 25, lemma 26 and lemma 24, like the proof of Theorem 3, we should choose

$$\delta_n \asymp \sqrt{\left(W + d_\beta + \text{Pdim}(G^d_{\phi_2}(Q_{\phi_1}(x))) + \text{Pdim}(Q_{\phi_1}(x)) + \text{Pdim}(G_{\phi_3}(Q_{\phi_1}(x)))\right)}$$
$$\sqrt{(\log n + L\log(\|U\|_1 V) + \log \frac{1}{\sigma_1})\frac{1}{n}\frac{(\log \frac{n}{\sigma_1})^{\frac{2}{\alpha}}}{\sigma_1^2}}.$$

Furthermore, for the Hard Tanh function $h(x) = \max(-b, \min(b, x))$, it can be express as $h(x) = \sigma(-\sigma(-x+b)+2b) - b$ with $\sigma(x) = \max(x, 0)$. Then, by Theorem 6 of Bartlett et al. (2019), we have

$$\text{Pdim}(G^d_{\phi_2}(Q_{\phi_1}(x))) = O((W_1+W_2)(L_1+L_2)\log(\|U_1\|_1 + \|U_2\|_1));$$
$$\text{Pdim}(Q_{\phi_1}(x)) = O(W_1 L_1 \log\|U_1\|_1);$$
$$\text{Pdim}(G_{\phi_3}(Q_{\phi_1}(x))) = O((W+W_1)(L+L_1)\log(U + \|U_1\|_1)).$$

We then bound the approximation error in the following lemma.

Lemma 27 *Consider $\mathcal{F}_{dd}$, $\mathcal{F}_{ed}$ defined in (9) and $\mathcal{F}_{prior}$ satisfying Condition B, given Assumption B, there exist some constants $(\eta, \gamma, \overline{\eta}, b_7, c)$ that only depend on (d_z, d_x) and $(\alpha, k, b_2, b_3, b_5, b_6)$ in Assumption B and Condition B, such that when $\frac{\epsilon_0}{\sigma^*} + \frac{\epsilon_1}{\sigma^{*2}} + \frac{\epsilon_2}{\sigma^*} \leq 1$,*

$$\min_{\substack{p_\theta \in \mathcal{F}_{dd}, q_\phi \in \mathcal{F}_{ed} \\ \pi_\beta \in \mathcal{F}_{prior}}} D_{\text{KL}}(p_{\mathcal{D}}(\cdot)||p_{\theta,\beta}(\cdot)) + \mathbb{E}_{p_{\mathcal{D}}(x)}\left[D_{\text{KL}}(q_\phi(\cdot|x)||p_{\theta,\beta}(\cdot|x))\right]$$
$$\leq c\sigma^{*2}\left(\log\frac{1}{\sigma^*}\right)^{\left(\frac{28+10\alpha+3\alpha^2}{\alpha^2}\right)},$$

with $(\epsilon_0, \epsilon_1, \epsilon_2)$ being defined in equation (7).

We can then get the desired conclusion using Theorem 1.

C.4. Nonparametric Models

C.4.1. PROOF OF LEMMA 19 AND LEMMA 20

$$\begin{aligned}&\sqrt{\frac{1}{n}\sum_{i=1}^n m(p,q,\pi_\beta,x_i) - m(p',q',\pi_{\beta'},x_i))^2} \\ &\leq \sup_x \left|\log\frac{p(x)}{p'(x)}\right| + \sup_x \left|D_{\text{KL}}(q(\cdot|x)||p(\cdot|x)) - D_{\text{KL}}(q'(\cdot|x)||p'(\cdot|x))\right|.\end{aligned} \tag{12}$$

For the first term of equation (12), one has

$$\begin{aligned}&p'(x)\sup_z\frac{p(x|z)\pi_\beta(z)}{p'(x|z)\pi_{\beta'}(z)} \\ &= \int_{\mathcal{Z}} p'(x|z)\sup_z\frac{p(x|z)\pi_\beta(z)}{p'(x|z)\pi_{\beta'}(z)}\pi_{\beta'}(z)dz \\ &\geq \int_{\mathcal{Z}} p'(x|z)\frac{p(x|z)\pi_\beta(z)}{p'(x|z)\pi_{\beta'}(z)}\pi_{\beta'}(z)dz \\ &= p(x).\end{aligned}$$

Similarly, it holds that $\frac{p'(x)}{p(x)} \leq \sup_z \frac{p'(x|z)\pi_{\beta'}(z)}{p(x|z)\pi_\beta(z)}$. Therefore when $\frac{p(x)}{p'(x)} \geq 1$,

$$\begin{aligned}\left|\log\frac{p(x)}{p'(x)}\right| = \log\frac{p(x)}{p'(x)} &\leq \log\sup_z \frac{p(x|z)\pi_\beta(z)}{p'(x|z)\pi_{\beta'}(z)}\\ &\leq \sup_z\left|\log p(x|z) - \log p'(x|z)\right| + \sup_z\left|\log \pi_\beta(z) - \log \pi_{\beta'}(z)\right|.\end{aligned}$$

Similarly, when $\frac{p(x)}{p'(x)} \leq 1$,

$$\begin{aligned}\left|\log\frac{p(x)}{p'(x)}\right| &\leq \log\sup_z \frac{p'(x|z)\pi_{\beta'}(z)}{p(x|z)\pi_\beta(z)}\\ &\leq \sup_z\left|\log p(x|z) - \log p'(x|z)\right| + \sup_z\left|\log \pi_\beta(z) - \log \pi_{\beta'}(z)\right|.\end{aligned}$$

For the second term of equation (12),

$$\begin{aligned}&\sup_x\left|D_{\mathrm{KL}}(q(\cdot|x)||p(\cdot|x)) - D_{\mathrm{KL}}(q'(\cdot|x)||p'(\cdot|x))\right|\\ &\leq \sup_x\left|\int_{\mathcal{Z}}\log\frac{q(z|x)}{p(z|x)}q(z|x)dz - \int_{\mathcal{Z}}\log\frac{q(z|x)}{p'(z|x)}q(z|x)dz\right|\\ &+\sup_x\left|\int_{\mathcal{Z}}\log\frac{q(z|x)}{p'(z|x)}q(z|x)dz - \int_{\mathcal{Z}}\log\frac{q'(z|x)}{p'(z|x)}q'(z|x)dz\right|.\end{aligned} \tag{13}$$

Then for the first part of equation (13),

$$\begin{aligned}&\sup_x\left|\int_{\mathcal{Z}}\log\frac{p'(z|x)}{p(z|x)}q(z|x)dz\right|\\ &\leq \sup_{x,z}\left|\log\frac{p'(x|z)p'(x)}{p(x|z)p(x)}\right| + \sup_z\left|\log\pi_\beta(z) - \log\pi_{\beta'}(z)\right|\\ &\leq 2\sup_{x,z}\left|\log p(x|z) - \log p'(x|z)\right| + 2\sup_z\left|\log\pi_\beta(z) - \log\pi_{\beta'}(z)\right|.\end{aligned}$$

For the second part of equation (13),

$$\begin{aligned}&\sup_x\left|\int_{\mathcal{Z}}\log\frac{q(z|x)}{p'(z|x)}\frac{q(z|x)}{p'(z|x)}p'(z|x)dz - \int_{\mathcal{Z}}\log\frac{q'(z|x)}{p'(z|x)}\frac{q'(z|x)}{p'(z|x)}p'(z|x)dz\right|\\ &\leq \sup_x\int_{\mathcal{Z}}\left|\log\frac{q(z|x)}{p'(z|x)}\frac{q(z|x)}{p'(z|x)} - \log\frac{q'(z|x)}{p'(z|x)}\frac{q'(z|x)}{p'(z|x)}\right|p'(z|x)dz\\ &\leq \sup_{\substack{p\in\mathcal{F}_{dd},q\in\mathcal{F}_{ed}\\ \pi_\beta\in\mathcal{F}_{prior}}}\sup_{x,z}(1+|\log p(z|x)|+|\log q(z|x)|)\cdot\sup_x\int_{\mathcal{Z}}\left|\frac{q(z|x)}{p'(z|x)} - \frac{q'(z|x)}{p'(z|x)}\right|p'(z|x)dz\\ &\leq (C+1)\sup_x\int_{\mathcal{Z}}\left|q(z|x) - q'(z|x)\right|dz.\end{aligned}$$

Then by $x \log x \geq x - 1$

$$
\begin{aligned}
&\int_{\mathcal{Z}} \left|q(z|x) - q'(z|x)\right| dz \\
&= \int_{\mathcal{Z}} \left(\frac{q(z|x)}{q'(z|x)} - 1\right) \mathbf{1}(q(z|x) \geq q'(z|x)) q'(z|x) dz \\
&+ \int_{\mathcal{Z}} \left(\frac{q'(z|x)}{q(z|x)} - 1\right) \mathbf{1}(q'(z|x) \geq q(z|x)) q(z|x) dz \\
&\leq \int_{\mathcal{Z}} \left|\frac{q(z|x)}{q'(z|x)} \log \frac{q(z|x)}{q'(z|x)}\right| q'(z|x) dz + \int_{\mathcal{Z}} \left|\frac{q'(z|x)}{q(z|x)} \log \frac{q'(z|x)}{q(z|x)}\right| q(z|x) dz \\
&\leq 2 \sup_{z} \left|\log q(z|x) - \log q'(z|x)\right|.
\end{aligned}
$$

We then get the desired conclusion in Lemma 19.

For Lemma 20, since $p(x|z) \propto exp\left(\sum_{j=1}^{k_1} l_j(\overline{x}_j, z)\right)$, then we can write $\log p(x|z)$ as $\sum_{j=1}^{k_1} l_j(\overline{x}_j, z) - C(x)$, with $C(x) = \log \int_{\mathcal{Z}} \exp(\sum_{j=1}^{k_1} l_j(\overline{x}_j, z)) dz$. By the same argument of the proof of Lemma 19, we have

$$
\begin{aligned}
\frac{\int_{\mathcal{Z}} \exp(\sum_{j=1}^{k_1} l_j(\overline{x}_j, z)) dz}{\int_{\mathcal{Z}} \exp(\sum_{j=1}^{k_1} l'_j(\overline{x}_j, z)) dz} &\leq \sup_{z} \frac{\exp(\sum_{j=1}^{k_1} l_j(\overline{x}_j, z))}{\exp(\sum_{j=1}^{k_1} l'_j(\overline{x}_j, z))} \\
\left|\log \frac{\int_{\mathcal{Z}} \exp(\sum_{j=1}^{k_1} l_j(\overline{x}_j, z)) dz}{\int_{\mathcal{Z}} \exp(\sum_{j=1}^{k_1} l'_j(\overline{x}_j, z)) dz}\right| &\leq \sup_{z} \left|\sum_{j=1}^{k_1} l_j(\overline{x}_j, z) - \sum_{j=1}^{k_1} l'_j(\overline{x}_j, z)\right|.
\end{aligned}
$$

So we have $\sup_{x,z} |\log p(x|z) - \log p'(x|z)| \leq 2 \sup_{x,z} \left|\sum_{j=1}^{k_1} l_j(\overline{x}_j, z) - \sum_{j=1}^{k_1} l'_j(\overline{x}_j, z)\right|$. Similarly for the case of $q(z|x)$.

C.5. Proof of additional lemmas

C.5.1. PROOF OF LEMMA 21

Let $y = \frac{p_{\mathcal{D}}(x)}{p(x)}$, then $y \in [e^{-C}, e^{C}]$.

$$
\begin{aligned}
f(y) &= (2 + C)(y \log y - y + 1) - y(\log y)^2. \\
f'(y) &= (C - \log y) \log y.
\end{aligned}
$$

When $1 \leq y \leq e^{C}$, $f'(y) > 0$. When $e^{-C} \leq y < 1$, $f'(y) < 0$. Since $f(1) = 0$, when $y \in [e^{-C}, e^{C}])$, it holds that $f(y) \geq 0$.

C.5.2. PROOF OF LEMMA 25

For $p_{\theta,\beta}(x) = \int p_\theta(x|z) \pi_\beta(z) dz$, define

$$A_{\theta,\beta} = \{x : p_{\theta,\beta}(x) > 1\}.$$

Then for any $x \in A_{\theta,\beta}$,

$$\begin{aligned}
|\log p_{\theta,\beta}(x)| = \log p_{\theta,\beta}(x) &= \log \int p_\theta(x|z)\pi_\beta(z)dz \\
&= \log \int \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{d_x}{2}} \exp\left(-\frac{\sum_{j=1}^{d_x}(x_j - G_{\theta,j}(z))^2}{2\sigma^2}\right)\pi_\beta(z)dz \\
&\le \log \int \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{d_x}{2}} \pi_\beta(z)dz \\
&\le -\frac{d_x}{2}\log(2\pi\sigma_1^2),
\end{aligned}$$

$$\begin{aligned}
D_{\mathrm{KL}}(q_\phi(\cdot|x)||p_{\theta,\beta}(\cdot|x)) &= \int \log \frac{q_\phi(z|x)p_{\theta,\beta}(x)}{p_\theta(x|z)\pi_\beta(z)} q_\phi(z|x)dz \\
&= \int \log q_\phi(z|x)q_\phi(z|x)dz + \log p_{\theta,\beta}(x) \\
&\quad - \int \log p_\theta(x|z)q_\phi(z|x)dz - \int \log \pi_\beta(z)q_\phi(z|x)dz.
\end{aligned}$$

For any $x \in A^c_{\theta,\beta}$,

$$\begin{aligned}
&|\log p_{\theta,\beta}(x)| + D_{\mathrm{KL}}(q_\phi(\cdot|x)||p_{\theta,\beta}(\cdot|x)) \\
&= -\int \log p_\theta(x|z)q_\phi(z|x)dz + D_{\mathrm{KL}}(q_\phi(\cdot|x)||\pi_\beta(\cdot)) \\
&= \int \log q_\phi(z|x)q_\phi(z|x)dz - \int \log p_\theta(x|z)q_\phi(z|x)dz - \int \log \pi_\beta(z)q_\phi(z|x)dz,
\end{aligned}$$

$$\int \log q_\phi(z|x)q_\phi(z|x)dz = -\frac{1}{2}\log|\Sigma_\phi(x)| - \frac{d_z}{2}(1 + \log(2\pi)),$$

$$\begin{aligned}
-\int \log p_\theta(x|z)q_\phi(z|x)dz &= \frac{d_x}{2}\log(2\pi\sigma^2) + \int \frac{\sum_{j=1}^{d_x}(x_j - G_{\theta,j}(z))^2}{2\sigma^2} q_\phi(z|x)dz \\
&\le \frac{d_x}{2}\log(2\pi) + \frac{1}{\sigma_1^2}\sum_{j=1}^{d_x}\left(x_j^2 + \int G^2_{\theta,j}(z)q_\phi(z|x)dz\right).
\end{aligned}$$

By the definition of $\mathcal{F}_{dd}$ and $\mathcal{F}_{ed}$ in (9) and Condition B, there exists a constant c such that

$$\begin{aligned}
\int \|G_\theta(z)\|_2^2 q_\phi(z|x)dz &\le c(\log\frac{1}{\sigma_1})^{\frac{2}{\alpha}}; \\
\left|\int \log \pi_\beta(z)q_\phi(z|x)dz\right| &\le c\log\frac{1}{\sigma_1}.
\end{aligned}$$

Then by the fact that when $\sigma_1 \le 1$, $\log\frac{1}{\sigma_1^2} \le \frac{1}{\sigma_1^2}$, we can get the desired conclusion.

C.5.3. PROOF OF LEMMA 26

We begin the proof with the following lemma about the Lipschitzness of ReLU neural networks (w.r.t the parameter).

Lemma 28 *If $G_{\theta_1}(z)$ is ReLU neural network in which the information of each layer can only come from the previous one layer. Also, it has L layers, use Hard Tanh as the activation function for the output and there exists a constant $V \geq 2$ such that in each layer, the units $\overline{\omega}$ has $\|\overline{\omega}\|_2 \leq V$, then, for any $G_{\theta_1}(z)$ and $G_{\theta_1'}(z)$,*

$$\|G_{\theta_1}(z) - G_{\theta_1'}(z)\|_2 \leq V^{L-2}(2 + \|z\|_2)\sqrt{2L}\|\theta_1 - \theta_1'\|_2.$$

We then return to our proof of Lemma 26. For any $x \in \mathcal{B}_x$,

$$\begin{aligned}
&|m(p_\theta, q_\phi, \pi_\beta, x) - m(p_{\theta'}, q_{\phi'}, \pi_{\beta'}, x)| \\
&\leq \left|\int \log \frac{p_\theta(x|z)}{p_{\theta'}(x|z)} q_\phi(z|x) dz\right| + \left|\int \log p_{\theta'}(x|z)(q_\phi(z|x) - q_{\phi'}(z|x)) dz\right| \\
&+ \left|\int \log q_\phi(z|x) q_\phi(z|x) dz - \int \log q_{\phi'}(z|x) q_{\phi'}(z|x) dz\right| \\
&+ \left|\int \log \pi_\beta(z) q_\phi(z|x) dz - \int \log \pi_{\beta'}(z) q_{\phi'}(z|x) dz\right|.
\end{aligned} \tag{14}$$

For the first term of equation (14), since in each layer of $G_{\theta_1}(z)$, the weights $\overline{w}$ has $\|\overline{w}\|_F \leq V\|U\|_1$. Then by Lemma 28 and the boundedness of $\mu_\phi(x)$ and $\Sigma_\phi(x)$, we have

$$\begin{aligned}
&\left|\int \log \frac{p_\theta(x|z)}{p_{\theta'}(x|z)} q_\phi(z|x) dz\right| \\
&\leq \int \frac{1}{\sigma^2} \sum_{j=1}^{d_x} \left|\frac{G^2_{\theta_1,j}(z) - G^2_{\theta_1',j}(z)}{2} + x_j(G_{\theta_1',j}(z) - G_{\theta_1,j}(z))\right| q_\phi(z|x) dz \\
&+ \frac{d_x}{2}\left|\log \frac{1}{\sigma^2} - \log \frac{1}{\sigma'^2}\right| + \left|\frac{1}{\sigma^2} - \frac{1}{\sigma'^2}\right| \int \frac{\sum_{j=1}^{d_x}(x_j - G_{\theta_1',j}(z))^2}{2} q_\phi(z|x) dz \\
&\leq \frac{1}{\sigma_1^2} \int \left(\|G_{\theta_1,j}(z) - G_{\theta_1',j}(z)\|_2 \left(\frac{1}{2}\|G_{\theta_1,j}(z) + G_{\theta_1',j}(z)\|_2 + \|x\|_2\right)\right) q_\phi(z|x) dz + c(\log \frac{n}{\sigma_1})^{\frac{2}{\alpha}} \left|\frac{1}{\sigma^2} - \frac{1}{\sigma'^2}\right| \\
&\lesssim \frac{\sqrt{L}(\log \frac{n}{\sigma_1})^{\frac{2}{\alpha}}}{\sigma_1^2} (\|U\|_1 V)^{L-2} \|\theta_1 - \theta_1'\|_2 + (\log \frac{n}{\sigma_1})^{\frac{2}{\alpha}} \left|\frac{1}{\sigma^2} - \frac{1}{\sigma'^2}\right|.
\end{aligned}$$

For the second term of equation (14),

$$\begin{aligned}
&\left|\int \log p_{\theta'}(x|z)(q_\phi(z|x) - q_{\phi'}(z|x)) dz\right| \\
&= \frac{1}{\sigma'^2} \left|\int -\frac{\sum_{j=1}^{d_x}(x_j - G_{\theta_1',j}(z))^2}{2} (q_\phi(z|x) - q_{\phi'}(z|x)) dz\right| \\
&\leq \frac{1}{\sigma'^2} \left|\int -\frac{\sum_{j=1}^{d_x} G^2_{\theta_1',j}(z) - 2x_j G_{\theta_1',j}(z)}{2} (q_\phi(z|x) - q_{\phi'}(z|x)) dz\right|.
\end{aligned}$$

By the fact that $\|G_{\theta_1}(z) - G_{\theta_1}(z')\|_2 \le (\|U\|_1 V)^{L-1}\|z - z'\|_2$, it holds that

$$\begin{aligned}
&\left|\int \|G_{\theta_1'}(z)\|_2^2 (q_\phi(z|x) - q_{\phi'}(z|x))dz\right| \\
&\le \inf_{\gamma_x \in \Pi\left(q_\phi(\cdot|x), q_{\phi'}(\cdot|x)\right)} \int_{\mathbb{R}^{d_z}\times\mathbb{R}^{d_z}} \left|\|G_{\theta_1'}(z)\|_2^2 - \|G_{\theta_1'}(z_0)\|_2^2\right| d\gamma_x \\
&\le \inf_{\gamma_x \in \Pi\left(q_\phi(\cdot|x), q_{\phi'}(\cdot|x)\right)} \int_{\mathbb{R}^{d_z}\times\mathbb{R}^{d_z}} \|G_{\theta_1'}(z) - G_{\theta_1'}(z_0)\|_2 \|G_{\theta_1'}(z) + G_{\theta_1'}(z_0)\|_2 d\gamma_x \\
&\le c(\|U\|_1 V)^{L-1}(\log\frac{1}{\sigma_1})^{\frac{1}{\alpha}} W_2(q_\phi(\cdot|x), q_{\phi'}(\cdot|x)),
\end{aligned}$$

where $W_2(\mu_0, \mu_1)$ denotes the Wasserstein-2 distance defined as (Santambrogio, 2015):

$$W_2^2(\mu_0, \mu_1) := \inf_{Y_0\sim\mu_0; Y_1\sim\mu_1} \mathbb{E}\left(\|Y_0 - Y_1\|^2\right) = \inf_{\gamma\in\Pi(\mu_0,\mu_1)} \int_{\mathbb{R}^{d_z}\times\mathbb{R}^{d_z}} \|y_0 - y_1\|_2^2 \, d\gamma(y_0, y_1).$$

Similarly, we can get

$$\begin{aligned}
&\left|\int \sum_{j=1}^{d_x} x_j G_{\theta_1'}(z)(q_\phi(z|x) - q_{\phi'}(z|x))dz\right| \\
&\le c \log^{\frac{1}{\alpha}} n(\|U\|_1 V)^{L-1} W_2(q_\phi(\cdot|x), q_{\phi'}(\cdot|x)).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left|\int \log p_{\theta'}(x|z)(q_\phi(z|x) - q_{\phi'}(z|x))dz\right| \\
&\lesssim \frac{(\|U\|_1 V)^{L-1}}{\sigma_1^2}(\log\frac{n}{\sigma_1})^{\frac{1}{\alpha}} W_2(q_\phi(\cdot|x), q_{\phi'}(\cdot|x)).
\end{aligned}$$

Furthermore, by Givens and Shortt (1984), we have

$$\begin{aligned}
W_2(q_\phi(\cdot|x), q_{\phi'}(\cdot|x)) &= \left\|\mu_\phi(x) - \mu_{\phi'}(x)\right\|_2^2 + \operatorname{Tr}\left(\Sigma_\phi(x) + \Sigma_{\phi'}(x) - 2\left(\Sigma_\phi^{\frac{1}{2}}(x)\Sigma_{\phi'}(x)\Sigma_\phi^{\frac{1}{2}}(x)\right)^{\frac{1}{2}}\right) \\
&= \left\|\mu_\phi(x) - \mu_{\phi'}(x)\right\|_2^2 + \|\Sigma_\phi^{\frac{1}{2}}(x) - \Sigma_{\phi'}^{\frac{1}{2}}(x)U(x,\phi,\phi')\|_F^2,
\end{aligned}$$

where $U(x,\phi,\phi') = \Sigma_{\phi'}^{-\frac{1}{2}}(x)\Sigma_\phi^{-\frac{1}{2}}(x)\left(\Sigma_\phi^{\frac{1}{2}}(x)\Sigma_{\phi'}(x)\Sigma_\phi^{\frac{1}{2}}(x)\right)^{\frac{1}{2}}$. Then let $\Sigma_\phi(x) = US^2U^T$ and $\Sigma_{\phi'}(x) = VS_1^2V^T$ be the eigenvalue decomposition of $\Sigma_\phi(x)$ and $\Sigma_{\phi'}(x)$, we have

$$U(x,\phi,\phi') = US^{-1}U^TVS_1^{-1}V^T\left(USU^TVS_1^2V^TUSU^T\right)^{\frac{1}{2}}.$$

By Davis-Kahan theorem (Davis and Kahan, 1970) and the boundedness of the eigenvalues of $\Sigma_\phi(x)$, it holds with a constant c_4 that

$$\begin{aligned}
\|I - U^TV\|_F &\lesssim \sigma_1^{-c_4}\|\Sigma_\phi(x) - \Sigma_{\phi'}(x)\|_F; \\
\|I - V^TU\|_F &\lesssim \sigma_1^{-c_4}\|\Sigma_\phi(x) - \Sigma_{\phi'}(x)\|_F.
\end{aligned}$$

Then combine all these facts, we have

$$\begin{aligned}&\left|\int \log p_{\theta'}(x|z)(q_\phi(z|x) - q_{\phi'}(z|x))dz\right| \\ &\lesssim (\|U\|_1 V)^{L-1}\frac{\log^{c_3} n}{\sigma_1^{c_4}}(\|\mu_\phi(x) - \mu_{\phi'}(x)\|_2 + \|\Sigma_\phi(x) - \Sigma_{\phi'}(x)\|_F).\end{aligned}$$

For the third term of equation (14),

$$\begin{aligned}&\left|\int \log q_\phi(z|x)q_\phi(z|x)dz - \int \log q_{\phi'}(z|x)q_{\phi'}(z|x)dz\right| \\ &= \left|\frac{1}{2}\log|\Sigma_\phi(x)| - \frac{1}{2}\log|\Sigma_{\phi'}(x)|\right| \\ &\lesssim \sigma_1^{-c_4}\|\Sigma_\phi(x) - \Sigma_{\phi'}(x)\|_F.\end{aligned}$$

For the last term of equation (14), by Condition B, we can get

$$\begin{aligned}|\log \pi_\beta(z) - \log \pi_\beta(z_0)| &= \nabla_z \log \pi_\beta(cz + (1-c)z_0)(z - z_0) \\ &\leq (c_1(\|z\|_2 + \|z_0\|_2) + c_2)\|z - z_0\|_2.\end{aligned}$$

Therefore,

$$\begin{aligned}|\log \pi_\beta(z)q_\phi(z|x)dz - \log \pi_\beta(z)q_{\phi'}(z|x)dz| &\leq \inf_{\gamma_x \in \Pi\left(q_\phi(z|x), q_{\phi'}(z|x)\right)} \int_{\mathbb{R}^d\times\mathbb{R}^d} (c_1(\|z\|_2 + \|z_0\|_2) + c_2)\|z - z_0\|_2 d\gamma_x \\ &\leq \left(2c_2^2 + 4c_1^2(\int \|z\|_2^2 q_\phi(z|x)dz + \int \|z\|_2^2 q_{\phi'}(z|x)dz\right) \\ &\quad\times \left(\inf_{\gamma_x \in \Pi\left(q_\phi(z|x), q_{\phi'}(z|x)\right)} \int_{\mathbb{R}^d\times\mathbb{R}^d} |z - z_0|_2^2 d\gamma_x\right)^{\frac{1}{2}}.\end{aligned}$$

Then use the same strategy for bounding the second term of equation (14) and the fact that $|\log \pi_\beta(z) - \log \pi_{\beta'}(z)| \leq (c_3\|z\|_2^{c_5} + c_4)\|\beta - \beta'\|_2$, we have

$$\begin{aligned}&\left|\int \log \pi_\beta(z)q_\phi(z|x)dz - \int \log \pi_{\beta'}(z)q_{\phi'}(z|x)dz\right| \\ &\lesssim \frac{\log^{c_3} n}{\sigma_1^{c_4}}(\|\beta - \beta'\|_2 + \|\mu_\phi(x) - \mu_{\phi'}(x)\|_2 + \|\Sigma_\phi(x) - \Sigma_{\phi'}(x)\|_F).\end{aligned}$$

We can then get the desired conclusion.

C.5.4. PROOF OF LEMMA 27

Assume $Q_1(x)$, $G_1(z)$ and $G_1^d(z)$ achieve the rate ϵ_0, ϵ_1 and ϵ_2, that is

$$\begin{aligned}&\max_{x\in B_x}\|Q_1(x) - Q_{\mathcal{D}}(x)\|_2 = \epsilon_0; \\ &\max_{z\in \overline{B}_z}\|G_1(z) - G_{\mathcal{D}}(z)\|_2 = \epsilon_1; \\ &\max_{z\in \overline{B}_z}\|G_1^d(z) - \nabla G_{\mathcal{D}}(z)\|_F = \epsilon_2.\end{aligned} \tag{15}$$

By Assumption B, we have $\sup_{\|z\|_2\leq r}\|\nabla G_{\mathcal{D}}(z)\|_F \leq c_5 r^{\frac{2}{\alpha}}+c_6$ and $\sup_{\|z\|_2\leq r}\|\nabla Q_{\mathcal{D}}(x)|_{x=G_{\mathcal{D}}(z)}\|_F \leq c_5 r^{\frac{4}{\alpha^2}}+c_6$. Moreover, by the fact that $z = Q_{\mathcal{D}}(G_{\mathcal{D}}(z))$, we have $I_{d_z} = \nabla Q_{\mathcal{D}}(x)|_{x=G_{\mathcal{D}}(z)}\nabla G_{\mathcal{D}}(z)$. Then, for a fixed $z \in \overline{B}_z$, let $\nabla Q_{\mathcal{D}}(x)^T|_{x=G_{\mathcal{D}}(z)} = U_1 S_1 V_1^T$ and $\nabla G_{\mathcal{D}}(z) = U_2 S_2 V_2^T$ be the singular value decomposition of $\nabla Q_{\mathcal{D}}(x)^T|_{x=G_{\mathcal{D}}(z)}$ and $\nabla G_{\mathcal{D}}(z)$, where $U_1, U_2 \in \mathbb{O}(d_x, d_z)$, $V_1, V_2 \in \mathbb{O}(d_z)$. Then it holds that $S_2^{-1} = V_2^T V_1 S_1 U_1^T U_2$. We can thus obtain that when $\|z\|_2 \leq r$, $\lambda_{\min}(\nabla G_{\mathcal{D}}(z)^T\nabla G_{\mathcal{D}}(z)) \geq \frac{1}{a\left(1+r^{\frac{8}{\alpha^2}}\right)}$ for some constant $a > 0$. Given this fact, we define

$$\begin{aligned}
&\Sigma_1(x) = (G_1^d(Q_1(x))^T G_1^d(Q_1(x)) + \sigma^{*2} I_{d_z})^{-1};\\
&\tilde{\Sigma}_1(x)_{i,j} = \max(-\bar{b}_7, \min(\bar{b}_7, \Sigma_1(x)_{i,j})) \quad (1 \leq i,j \leq d_x,\ \bar{b}_7 = b_7(\log\frac{1}{\sigma^*})^{\frac{4}{\alpha^2}});\\
&\mu_1(x) = Q_1(x) + \tilde{\Sigma}_1(x) G_1^d(Q_1(x))^T(x - G_1(Q_1(x)));\\
&\epsilon_2' = \max_{x\in B_x}\|\mu_1(x) - (Q_1(x) + \Sigma(x)\nabla G_{\mathcal{D}}(Q_1(x))^T(x - G_{\mathcal{D}}(Q_1(x))))\|_2;\\
&\epsilon_3' = \max_{x\in B_x}\|\Sigma_1(x)^{-1} - \Sigma(x)^{-1}\|_F,
\end{aligned} \tag{16}$$

in which $\Sigma(x)$ is defined as $(\nabla G_{\mathcal{D}}(Q_1(x))^T\nabla G_{\mathcal{D}}(Q_1(x)))^{-1}$ with $\nabla G_{\mathcal{D}}(Q_1(x)) = \nabla G_{\mathcal{D}}(z)|_{z=Q_1(x)}$.

Define

$$\begin{aligned}
&p(x|z) \sim \mathcal{N}\left(G_1(z), \sigma^{*2} I\right), \quad p(x) = \int p(x|z)\pi_{\mathcal{D}}(z)dz;\\
&q(z|x) \sim \mathcal{N}\left(\mu_1(x), \sigma^{*2}\Sigma_1(x)\right);\\
&p(z|x) = \frac{p(x|z)\pi_{\mathcal{D}}(z)}{p(x)}.
\end{aligned} \tag{17}$$

Consider $z = Q_1(x) + (z^0 - Q_1(x))\sigma^*$, define

$$\begin{aligned}
&q^0(z^0|x) \sim \mathcal{N}\left(Q_1(x) + \frac{\mu_1(x) - Q_1(x)}{\sigma^*}, \Sigma_1(x)\right);\\
&p^0(z^0|x) = \sigma^{*d_z} p\left(z = Q_1(x) + (z^0 - Q_1(x))\sigma^* | x\right).
\end{aligned}$$

Since D_{KL} is invariant to affine transformations, we have

$$D_{\mathrm{KL}}(q^0(\cdot|x)||p^0(\cdot|x)) = D_{\mathrm{KL}}(q(\cdot|x)||p(\cdot|x)).$$

Recall that $B_z = [-\eta\log^{\frac{1}{2}}\frac{1}{\sigma^*}, \eta\log^{\frac{1}{2}}\frac{1}{\sigma^*}]^{d_z}$ and $B_\epsilon = [-\gamma\log^{\frac{1}{2}}\frac{1}{\sigma^*}, \gamma\log^{\frac{1}{2}}\frac{1}{\sigma^*}]^{d_x}$, then by Lemma 25 and the assumption that $\|G_{\mathcal{D}}(z)\|_2 \leq c_3\|z\|_2^{\frac{2}{\alpha}} + c_4$, we have for sufficient large η and γ, it holds that

$$\begin{aligned}
&\mathbb{E}_{p_{\mathcal{D}}(x)}\int \log\frac{q^0(z^0|x)}{p^0(z^0|x)} q^0(z^0|x)dz^0\\
&= \int\int\int log\frac{q^0(z^0|G_{\mathcal{D}}(z) + \epsilon\sigma^*)}{p^0(z^0|G_{\mathcal{D}}(z) + \epsilon\sigma^*)} q^0(z^0|G_{\mathcal{D}}(z) + \epsilon\sigma^*)dz^0\pi_{\mathcal{D}}(z)p(\epsilon)dzd\epsilon\\
&\leq \int_{B_\epsilon}\int_{B_z}\int log\frac{q^0(z^0|G_{\mathcal{D}}(z) + \epsilon\sigma^*)}{p^0(z^0|G_{\mathcal{D}}(z) + \epsilon\sigma^*)} q^0(z^0|G_{\mathcal{D}}(z) + \epsilon\sigma^*)dz^0\pi_{\mathcal{D}}(z)p(\epsilon)dzd\epsilon + \sigma^{*2}.
\end{aligned}$$

Define $x = G_{\mathcal{D}}(z) + \epsilon\sigma^*$, $\overline{x} = G_{\mathcal{D}}(z)$ and $r = Q_1(x) + (z^0 - Q_1(x))\sigma^*$. Then,

$$
\begin{aligned}
&p^0(z^0|x) = p(r|x){\sigma^*}^{d_z} = \frac{p(x|r)\pi_{\mathcal{D}}(r){\sigma^*}^{d_z}}{p(x)};\\
&\pi_{\mathcal{D}}(r) = \pi_{\mathcal{D}}(Q_1(x)) + \sigma^*\nabla\pi_{\mathcal{D}}(a_{z^0})^T(z^0 - Q_1(x)),\\
&a_{z^0} = Q_1(x) + c\sigma^*(z^0 - Q_1(x)) \quad (c \in [0,1]);
\end{aligned}
$$

and

$$
\begin{aligned}
p(x|r) &= \left(\frac{1}{2\pi{\sigma^*}^2}\right)^{\frac{d_x}{2}} \exp\left(-\frac{(x - G_1(r))^T(x - G_1(r))}{2{\sigma^*}^2}\right)\\
&= \left(\frac{1}{2\pi{\sigma^*}^2}\right)^{\frac{d_x}{2}} \exp\left(-\frac{(x - G_{\mathcal{D}}(r))^T(x - G_{\mathcal{D}}(r))}{2{\sigma^*}^2}\right)\\
&\exp\left(-\frac{\|G_{\mathcal{D}}(r) - G_1(r)\|_2^2}{2{\sigma^*}^2}\right)\exp\left(-\frac{(G_{\mathcal{D}}(r) - G_1(r))^T(x - G_{\mathcal{D}}(r))}{{\sigma^*}^2}\right).
\end{aligned}
$$

Let $D = -\frac{\|G_{\mathcal{D}}(r) - G_1(r)\|_2^2}{2{\sigma^*}^2}$ and $E = -\frac{(G_{\mathcal{D}}(r) - G_1(r))^T(x - G_{\mathcal{D}}(r))}{{\sigma^*}^2}$.

$$
G_{\mathcal{D}}(r) = G_{\mathcal{D}}(Q_1(x)) + \nabla G_{\mathcal{D}}(Q_1(x))(z^0 - Q_1(x))\sigma^* + R_n(x, z^0).
$$

Since for any $z \in B_z$ and $\epsilon \in B_\epsilon$, it holds that $\overline{x} = G_{\mathcal{D}}(z)$ and $x = \overline{x} + \sigma^*\epsilon$ belong to B_x. Then,

$$
\begin{aligned}
\|x - G_{\mathcal{D}}(Q_1(x))\|_2 &= \|x - \overline{x} + \overline{x} - G_{\mathcal{D}}(Q_{\mathcal{D}}(\overline{x})) + G_{\mathcal{D}}(Q_{\mathcal{D}}(\overline{x})) - G_{\mathcal{D}}(Q_1(x))\|_2\\
&\lesssim \epsilon_0\left(\log\frac{1}{\sigma^*}\right)^{\frac{1}{\alpha}} + \sigma^*\left(\log\frac{1}{\sigma^*}\right)^{\frac{2}{\alpha^2}+\frac{1}{\alpha}+\frac{1}{2}}\\
&\lesssim \sigma^*\left(\log\frac{1}{\sigma^*}\right)^{\frac{2}{\alpha^2}+\frac{1}{\alpha}+\frac{1}{2}}.
\end{aligned}
$$

Define $\Sigma(x) = \left(\nabla G_{\mathcal{D}}(Q_1(x))^T\nabla G_{\mathcal{D}}(Q_1(x))\right)^{-1}$, we have

$$
\begin{aligned}
p(x|r) = &\left((2\pi)^{d_z}|\Sigma(x)|\right)^{\frac{1}{2}}\mathcal{N}(z_0, Q_1(x) + \frac{\Sigma(x)}{\sigma^*}\nabla G_{\mathcal{D}}(Q_1(x))^T(x - G_{\mathcal{D}}(Q_1(x))), \Sigma(x))\\
&\times\left(\frac{1}{2\pi{\sigma^*}^2}\right)^{\frac{d_x}{2}}\exp\left(-\frac{(x - G_{\mathcal{D}}(Q_1(x)))^T(x - G_{\mathcal{D}}(Q_1(x)))}{2{\sigma^*}^2}\right)\\
&\times\exp\left(\frac{(x - G_{\mathcal{D}}(Q_1(x)))^T\nabla G_{\mathcal{D}}(Q_1(x))\Sigma(x)\nabla G_{\mathcal{D}}(Q_1(x))^T(x - G_{\mathcal{D}}(Q_1(x)))}{2{\sigma^*}^2}\right)\\
&\times\exp\left(\frac{(x - G_{\mathcal{D}}(Q_1(x)) - \sigma^*\nabla G_{\mathcal{D}}(Q_1(x))(z^0 - Q_1(x))^T}{{\sigma^*}^2}R_n(x, z^0)\right)\\
&\times\exp\left(-\frac{R_n(x, z^0)^TR_n(x, z^0)}{2{\sigma^*}^2}\right)\exp(D + E),
\end{aligned}
$$

where $\mathcal{N}(z_0, Q_1(x) + \frac{\Sigma(x)}{\sigma^*}\nabla G_{\mathcal{D}}(Q_1(x))^T(x - G_{\mathcal{D}}(Q_1(x))), \Sigma(x))$ is the corresponding normal density with variable z_0, mean vector $Q_1(x) + \frac{\Sigma(x)}{\sigma^*}\nabla G_{\mathcal{D}}(Q_1(x))^T(x - G_{\mathcal{D}}(Q_1(x)))$ and covariance matrix $\Sigma(x)$. Set

$$\begin{aligned} B =& \frac{(x - G_{\mathcal{D}}(Q_1(x)))^T \nabla G_{\mathcal{D}}(Q_1(x))\Sigma(x)\nabla G_{\mathcal{D}}(Q_1(x))^T(x - G_{\mathcal{D}}(Q_1(x)))}{2\sigma^{*2}}; \\ &- \frac{(x - G_{\mathcal{D}}(Q_1(x)))^T(x - G_{\mathcal{D}}(Q_1(x)))}{2\sigma^{*2}}. \\ C =& \frac{(x - G_{\mathcal{D}}(Q_1(x)) - \sigma^* \nabla G_{\mathcal{D}}(Q_1(x))(z^0 - Q_1(x))^T}{\sigma^{*2}} R_n(x, z^0) - \frac{R_n(x, z^0)^T R_n(x, z^0)}{2\sigma^{*2}} \end{aligned}$$

Then,

$$\begin{aligned} p(x|r) = &\left((2\pi)^{d_z}|\Sigma(x)|\right)^{\frac{1}{2}} \mathcal{N}(z_0, Q_1(x) + \frac{\Sigma(x)}{\sigma^*}\nabla G_{\mathcal{D}}(Q_1(x))^T(x - G_{\mathcal{D}}(Q_1(x))), \Sigma(x)) \\ &\left(\frac{1}{2\pi\sigma^{*2}}\right)^{\frac{d_x}{2}} \exp(B)\exp(C + D + E). \end{aligned}$$

We then bound $p(x) = \int p(x|r)\pi_{\mathcal{D}}(r)dr$ with the following lemma.

Lemma 29 *Given above notations in Section C.5.4 and* $\alpha_1 = \frac{4}{\alpha^2} + \frac{1}{\alpha} + \frac{1}{2}$*, there exist some constants* $(c_0, c_1, c_2, c_3, c_4)$*, such that for any* $x \in B_x$*, it holds that*

$$\begin{aligned} &\exp\left(-c_1\left(\sigma^*\left(\log\frac{1}{\sigma^*}\right)^{\frac{2}{\alpha}+3\alpha_1} + \epsilon_0\left(\log\frac{1}{\sigma^*}\right)^{\frac{2}{\alpha}+2\alpha_1}\right)\right)\exp\left(-c_2\frac{\epsilon_1}{\sigma^*}\left(\log\frac{1}{\sigma^*}\right)^{\frac{1}{\alpha}+\alpha_1}\right) - c_0\sigma^*\left(\log\frac{1}{\sigma^*}\right)^{\frac{1}{2}+\alpha_1} \\ \leq& \frac{p(x)}{\sigma^{*d_z}\left(\frac{1}{2\pi\sigma^{*2}}\right)^{\frac{d_x}{2}}\left((2\pi)^{d_z}|\Sigma(x)|\right)^{\frac{1}{2}}\pi_{\mathcal{D}}(Q_1(x))exp(B)} \leq \\ &\exp\left(c_1\left(\sigma^*\left(\log\frac{1}{\sigma^*}\right)^{\frac{2}{\alpha}+3\alpha_1} + \epsilon_0\left(\log\frac{1}{\sigma^*}\right)^{\frac{2}{\alpha}+2\alpha_1}\right)\right)\exp\left(c_2\frac{\epsilon_1}{\sigma^*}\left(\log\frac{1}{\sigma^*}\right)^{\frac{1}{\alpha}+\alpha_1}\right) + c_0\sigma^*\left(\log\frac{1}{\sigma^*}\right)^{\frac{1}{2}+\alpha_1}. \end{aligned}$$

And there exists $A_{z^0} = [-c_3\left(\log\frac{1}{\sigma^*}\right)^{\alpha_1}, c_3\left(\log\frac{1}{\sigma^*}\right)^{\alpha_1}]^{d_z}$*, such that for any* $x \in B_x$*, it holds that*

$$\begin{aligned} &\int_{A^c_{z^0}} p^0(z^0|x)dz^0 \leq \sigma^{*2}; \\ &\int_{A^c_{z^0}} q^0(z^0|x)dz^0 \leq \sigma^{*2}; \\ &\int_{A^c_{z^0}} \log\frac{q^0(z^0|x)}{p^0(z^0|x)} q^0(z^0|x)dz^0 \leq \sigma^{*2}. \end{aligned}$$

Given Lemma 29, we have

$$\begin{aligned} D_{\mathrm{KL}}(q^0(\cdot|x)||p^0(\cdot|x)) &\leq \int_{A_{z^0}}\left(\frac{q^0(z^0|x)}{p^0(z^0|x)} - 1\right)^2 p^0(z^0|x)dz^0 + 2\sigma^{*2} \\ &= \int_{A_{z^0}}\left(\exp\left(\log\frac{q^0(z^0|x)}{p^0(z^0|x)}\right) - 1\right)^2 p^0(z^0|x)dz^0 + 2\sigma^{*2}. \end{aligned}$$

Since

$$\begin{aligned}
&\left|\log \frac{q^0(z^0|x)}{p^0(z^0|x)}\right| \\
\le &\left|\log \frac{\mathcal{N}\left(z_0, Q_1(x) + \frac{\mu_1(x)-Q_1(x)}{\sigma^*}, \Sigma_1(x)\right)}{\mathcal{N}\left(z_0, Q_1(x) + \frac{\Sigma(x)}{\sigma^*}\nabla G_{\mathcal{D}}(Q_1(x))^T(x - G_{\mathcal{D}}(Q_1(x))), \Sigma(x)\right)}\right| \\
+ &\left|\log \frac{\mathcal{N}\left(z_0, Q_1(x) + \frac{\Sigma(x)}{\sigma^*}\nabla G_{\mathcal{D}}(Q_1(x))^T(x - G_{\mathcal{D}}(Q_1(x))), \Sigma(x)\right)}{p^0(z^0|x)}\right|.
\end{aligned} \tag{18}$$

Then under B_x, for the first term of equation (18), by the fact that when $\|z\|_2 \le \sqrt{\log \frac{1}{\sigma^*}}$, it holds that $\lambda_{\min}(\nabla G_{\mathcal{D}}(z)^T \nabla G_{\mathcal{D}}(z)) \gtrsim (\log \frac{1}{\sigma^*})^{-\frac{4}{\alpha^2}}$ and $\lambda_{\max}(\nabla G_{\mathcal{D}}(z)^T \nabla G_{\mathcal{D}}(z)) \lesssim \left(\log \frac{1}{\sigma^*}\right)^{\frac{2}{\alpha}}$, we can obtain

$$\sup_{x \in B_x} |\log|\Sigma_1(x)| - \log|\Sigma(x)|| \lesssim \epsilon_3 \left(\log \frac{1}{\sigma^*}\right)^{\frac{4}{\alpha^2}}.$$

Recall the definition of ϵ_2' and ϵ_3' in equation (16). Combined with the fact that

$$\sup_{x \in B_x} \left\| \frac{1}{\sigma^*}(\mu_1(x) - Q_1(x) - \Sigma(x)\nabla G_{\mathcal{D}}(Q_1(x))^T(x - G_{\mathcal{D}}(Q_1(x)))) \right\|_2 \lesssim \frac{\epsilon_2'}{\sigma^*},$$

we can get

$$\begin{aligned}
&\sup_{\substack{x \in B_x \\ z_0 \in A_{z^0}}} \left|\log \frac{\mathcal{N}\left(z_0, Q_1(x) + \frac{\mu_1(x)-Q_1(x)}{\sigma^*}, \Sigma_1(x)\right)}{\mathcal{N}\left(z_0, Q_1(x) + \frac{\Sigma(x)}{\sigma^*}\nabla G_{\mathcal{D}}(Q_1(x))^T(x - G_{\mathcal{D}}(Q_1(x))), \Sigma(x)\right)}\right| \\
&\lesssim \epsilon_3' \left(\log \frac{1}{\sigma^*}\right)^{2\alpha_1} + \frac{\epsilon_2'}{\sigma^*}\left(\log \frac{1}{\sigma^*}\right)^{\alpha_1 + \frac{2}{\alpha}}.
\end{aligned}$$

For the second term of (18), since when $x \in B_x$, $Q_1(x) \in \overline{B}_z$, then for $z^0 \in A_{z^0}$ and $x \in B_x$, we have $r = Q_1(x) + \sigma^*(z_0 - Q_1(x)) \in \overline{B}_z$ given large enough $\overline{\eta}$. And for $x \in B_x$ and $r \in \overline{B}_z$, we have

$$\begin{aligned}
|C| &= \left| \frac{(x - G_{\mathcal{D}}(Q_1(x)) - \sigma^* \nabla G_{\mathcal{D}}(Q_1(x))(z^0 - Q_1(x))^T}{\sigma^{*2}} R_n(x, z^0) - \frac{R_n(x, z^0)^T R_n(x, z^0)}{2\sigma^{*2}} \right| \\
&\lesssim \sigma^* \left(\log \frac{1}{\sigma^*}\right)^{\frac{2}{\alpha} + 3\alpha_1} + \epsilon_0 \left(\log \frac{1}{\sigma^*}\right)^{\frac{2}{\alpha} + 2\alpha_1}; \\
|D| &= \frac{\|G_{\mathcal{D}}(r) - G_1(r)\|_2^2}{2\sigma^{*2}} \le \frac{\epsilon_1^2}{2\sigma^{*2}}; \\
|E| &= \left| \frac{(G_{\mathcal{D}}(r) - G_1(r))^T (x - G_{\mathcal{D}}(r))}{2\sigma^{*2}} \right| \lesssim \frac{\epsilon_1}{\sigma^*}\left(\log \frac{1}{\sigma^*}\right)^{\frac{1}{\alpha} + \alpha_1}.
\end{aligned}$$

By the assumption that $\|\nabla \log \pi_{\mathcal{D}}(z)\|_2 \le c_1 \|z\|_2 + c_2$, we have for any $x \in B_x$, $z^0 \in A_{z^0}$ and

$a_{z^0} = Q_1(x) + c\sigma^*(z^0 - Q_1(x))$ $(c \in [0,1])$, it holds that

$$\frac{\|\nabla \pi_{\mathcal{D}}(a_{z^0})\|_2}{\pi_{\mathcal{D}}(Q_1(x))} \lesssim \sqrt{\log \frac{1}{\sigma^*}}.$$

Then combined with Lemma 29, and the fact that $x\exp(x) \leq ex$ when $x \leq 1$. We can get that for any $z^0 \in A_{z^0}$ and $x \in B_x$,

$$\begin{aligned} &\left| \log \frac{\mathcal{N}\left(z_0, Q_1(x) + \frac{\Sigma(x)}{\sigma^*}\nabla G_{\mathcal{D}}(Q_1(x))^T(x - G_{\mathcal{D}}(Q_1(x))), \Sigma(x)\right)}{p^0(z^0|x)} \right| \\ &\lesssim \sigma^* \left(\log \frac{1}{\sigma^*}\right)^{\frac{2}{\alpha}+3\alpha_1} + \epsilon_0 \left(\log \frac{1}{\sigma^*}\right)^{\frac{2}{\alpha}+2\alpha_1} + \frac{\epsilon_1}{\sigma^*} \left(\log \frac{1}{\sigma^*}\right)^{\frac{1}{\alpha}+\alpha_1}. \end{aligned}$$

So finally, we have

$$\begin{aligned} D_{\mathrm{KL}}(q^0(\cdot|x)||p^0(\cdot|x)) \lesssim \sigma^{*2} \left(\log \frac{1}{\sigma^*}\right)^{\frac{4}{\alpha}+6\alpha_1} + \frac{\epsilon_1^2}{\sigma^{*2}} \left(\log \frac{1}{\sigma^*}\right)^{\frac{2}{\alpha}+2\alpha_1} + \frac{\epsilon_2'^2}{\sigma^{*2}} \left(\log \frac{1}{\sigma^*}\right)^{\frac{4}{\alpha}+2\alpha_1} \\ + \epsilon_0^2 \left(\log \frac{1}{\sigma^*}\right)^{\frac{4}{\alpha}+4\alpha_1} + \epsilon_3'^2 \left(\log \frac{1}{\sigma^*}\right)^{4\alpha_1}. \end{aligned}$$

Also, by Assumption C, we can choose a large enough b_7 such that

$$\bar{b}_7 = b_7 (\log \frac{1}{\sigma_1})^{\frac{4}{\alpha^2}} \geq \max_{1 \leq i,j \leq d_z} \sup_{x \in B_x} |\Sigma(x)_{i,j}|,$$

with $\Sigma(x) = \left(\nabla G_{\mathcal{D}}\left(Q_1(x)\right)^T \nabla G_{\mathcal{D}}\left(Q_1(x)\right)\right)^{-1}$. So by the definition of $\mu_1(x)$ and $\Sigma_1(x)$ in eqaution (16), we have,

$$\begin{aligned} \epsilon_2' &\lesssim \epsilon_1 \left(\log \frac{1}{\sigma^*}\right)^{\frac{2}{\alpha^2}} + \sigma^* \epsilon_2 \left(\log \frac{1}{\sigma^*}\right)^{\frac{10}{\alpha^2}+\frac{3}{\alpha}+\frac{1}{2}}; \\ \epsilon_3' &\lesssim \epsilon_2 \left(\log \frac{1}{\sigma^*}\right)^{\frac{1}{\alpha}} + \sigma^{*2}. \end{aligned}$$

We then bound $D_{\mathrm{KL}}(p_{\mathcal{D}}(\cdot)||p(\cdot))$ with the following lemma.

Lemma 30 *Given Assumption B and Condition B, there exists a constant c, such that*

$$D_{\mathrm{KL}}(p_{\mathcal{D}}(\cdot)||p(\cdot)) \leq c\left(\frac{\epsilon_1^2}{\sigma^{*2}}(\log \frac{1}{\sigma^*})^{\frac{4}{\alpha^2}+1+\frac{2}{\alpha}} + \sigma^{*2}\right),$$

where $p(x)$ is defined in equation (17).

We can then get the desired conclusion.

C.5.5. PROOF FOR LEMMA 28

We first consider the case that the activation function of the output is an identity function. when $L = 2$,

$$G^2_\theta(z) = w_1 z + b_1.$$

Then we have

$$\begin{aligned}
&\|G^2_{\theta_1}(z)\|_2 \le \|b_1\|_2 + \|w_1\|_2\|z\|_2 \le V(1 + \|z\|_2);\\
&\|G^2_{\theta_1}(z) - G^2_{\theta_1'}(z)\|_2 \le \|b_1 - b_1'\|_2 + \|w_1 z - w_1' z\|_2\\
&\le (2 + \|z\|_2)(\|w_1 - w_1'\|_2 + \|b_1 - b_1'\|_2)).
\end{aligned}$$

If it's hold for k-depth ReLU neural network that,

$$\begin{aligned}
&\|G^k_{\theta_1}(z)\|_2 \le V^{k-1}(1 + \|z\|_2) + \sum_{j=1}^{k-2} V^j;\\
&\|G^k_{\theta_1}(z) - G^k_{\theta_1'}(z)\|_2 \le V^{k-2}(2 + \|z\|_2)\sum_{j=1}^{k-1}(\|w_j - w_j'\|_2 + \|b_j - b_j'\|_2).
\end{aligned}$$

Then,

$$\begin{aligned}
\|G^{k+1}_{\theta_1}(z)\|_2 &\le \|w_{k+1}\sigma(G^k_{\theta_1}(z))\|_2 + \|b_{k+1}\|_2\\
&\le \|b_{k+1}\|_2 + \|w_{k+1}\|_F\left(V^{k-1}(1 + \|z\|_2) + \sum_{j=1}^{k-2} V^j\right)\\
&\le V^k(1 + \|z\|_2) + \sum_{j=1}^{k-1} V^j\\
&\le V^k(2 + \|z\|_2);
\end{aligned}$$

$$\begin{aligned}
&\|G^{k+1}_{\theta_1}(z) - G^{k+1}_{\theta_1'}(z)\|_2\\
&\le \|w_{k+1}\sigma(G^k_{\theta_1}(z)) - w_{k+1}'\sigma(G^k_{\theta_1'}(z))\|_2 + \|b_{k+1} - b_{k+1}'\|_2\\
&\le \|w_{k+1} - w_{k+1}'\|_F\|G^k_{\theta_1}(z)\|_2 + \|b_{k+1} - b_{k+1}'\|_2\\
&+ \|w_{k+1}'\|_F\left(V^{k-2}(2 + \|z\|_2)\sum_{j=1}^{k-1}(\|w_j - w_j'\|_2 + \|b_j - b_j'\|_2)\right)\\
&\le V^{k-1}(2 + \|z\|_2)\sum_{j=1}^{k}(\|w_j - w_j'\|_2 + \|b_j - b_j'\|_2)\\
&\le V^{k-1}(2 + \|z\|_2)\sqrt{2k}\|\theta - \theta'\|_2.
\end{aligned}$$

Furthermore, by the fact that for $h(x) = \max(-b_1, \min(b_1, x))$, it holds that

$$\begin{aligned}
&|h(x)| \le |x|;\\
&|h(x) - h(x')| \le |x - x'|,
\end{aligned}$$

the desired conclusion also holds for the case that the activation function of the output is $h(x)$.

C.5.6. PROOF OF LEMMA 29

Set $r = Q_1(x) + \sigma^*(z^0 - Q_1(x))$ then

$$\begin{aligned} p(x) &= \int p(x|z=r)\pi_{\mathcal{D}}(z=r)dr \\ &= \sigma^{*d_z} \int p(x|r)(\pi_{\mathcal{D}}(Q_1(x)) + \sigma^* \nabla \pi_{\mathcal{D}}(a_{z^0})^T (z^0 - Q_1(x)))dz^0. \end{aligned}$$

By the assumptions on $\pi_{\mathcal{D}}(z)$ and the fact that $\|x - G_{\mathcal{D}}(Q_1(x))\|_2 \lesssim \sigma^* \left(\log \frac{1}{\sigma^*}\right)^{\frac{2}{\alpha^2}+\frac{1}{\alpha}+\frac{1}{2}}$, there exist some constants (c_0, c_1) such that for any $x \in B_x$,

$$\begin{aligned} \frac{1}{\pi_{\mathcal{D}}(Q_1(x))} &\le \exp(c_0 \log \frac{1}{\sigma^*}); \\ \frac{1}{\exp(B)} &\le \exp\left(c_1 \left(\log \frac{1}{\sigma^*}\right)^{\frac{4}{\alpha^2}+1+\frac{2}{\alpha}}\right), \end{aligned}$$

where recall that

$$\begin{aligned} B =& \frac{(x - G_{\mathcal{D}}(Q_1(x)))^T \nabla G_{\mathcal{D}}(Q_1(x)) \Sigma(x) \nabla G_{\mathcal{D}}(Q_1(x))^T (x - G_{\mathcal{D}}(Q_1(x)))}{2\sigma^{*2}} \\ &- \frac{(x - G_{\mathcal{D}}(Q_1(x)))^T (x - G_{\mathcal{D}}(Q_1(x)))}{2\sigma^{*2}}. \end{aligned}$$

Let

$$\begin{aligned} B_z^1 &= \left[-c_2 \left(\log \frac{1}{\sigma^*}\right)^{\frac{2}{\alpha^2}+\frac{1}{\alpha}+\frac{1}{2}}, c_2 \left(\log \frac{1}{\sigma^*}\right)^{\frac{2}{\alpha^2}+\frac{1}{\alpha}+\frac{1}{2}}\right]^{d_z}; \\ A_{z^0}^1 &= \left[-c_3 \left(\log \frac{1}{\sigma^*}\right)^{\alpha_1}, c_3 \left(\log \frac{1}{\sigma^*}\right)^{\alpha_1}\right]^{d_z}; \\ A_{z^0}^2 &= \left\{z^0 \,|\, r = Q_1(x) + \sigma^*(z^0 - Q_1(x)) \in B_z^1\right\}. \end{aligned}$$

For sufficiently large c_2, we have $\forall x \in B_x$,

$$\begin{aligned} &\int_{(B_z^1)^c} p(x|z=r)\pi_{\mathcal{D}}(z=r)dr \\ &\le \left(\frac{1}{2\pi\sigma^{*2}}\right)^{\frac{d_x}{2}} \sigma^{*d_z} \pi_{\mathcal{D}}(z \in (B_z^1)^c) \exp\left(d_z \log \frac{1}{\sigma^*}\right) \\ &\le \sigma^{*2} \left(\frac{1}{2\pi\sigma^{*2}}\right)^{\frac{d_x}{2}} \sigma^{*d_z} \pi_{\mathcal{D}}(Q_1(X)) \exp(B). \end{aligned}$$

Under $z_0 \in A_{z^0}^{1,c} \bigcap A_{z^0}^2$, for any $x \in B_x$, we have

$$\begin{aligned} G_{\mathcal{D}}(r) &= G_{\mathcal{D}}(Q_1(x)) + \nabla G_{\mathcal{D}}(b_{z^0})(z^0 - Q_1(x))\sigma^*, \\ b_{z^0} &= Q_1(x) + c\sigma^*(z^0 - Q_1(x)) \quad c \in [0,1]. \end{aligned}$$

So there exists a constant a, such that

$$(z^0 - Q_1(x))^T \nabla G_{\mathcal{D}}(b_{z^0})^T \nabla G_{\mathcal{D}}(b_{z^0})(z^0 - Q_1(x)) \geq a \left(\log \frac{1}{\sigma^*}\right)^{-\frac{4}{\alpha^2}} \|z^0 - Q_1(x)\|_2^2.$$

Then, by the fact that

$$\begin{aligned}
&(x - G_1(r))^T(x - G_1(r)) \\
=&(x - G_{\mathcal{D}}(Q_1(x)) - \nabla G_{\mathcal{D}}(b_{z^0})(z^0 - Q_1(x))\sigma^* + (G_{\mathcal{D}}(r) - G_1(r)))^T \\
&(x - G_{\mathcal{D}}(Q_1(x)) - \nabla G_{\mathcal{D}}(b_{z^0})(z^0 - Q_1(x))\sigma^* + (G_{\mathcal{D}}(r) - G_1(r)).
\end{aligned}$$

We have for large enough $A^1_{z_0}$, there exists a constant c_4 such that for any $x \in B_x$,

$$\begin{aligned}
&\sigma^{*d_z} \int_{A^{1,c}_{z^0} \bigcap A^2_{z^0}} p(x|r)\pi(r) dz^0 \\
\leq& \sup_{A^{1,c}_{z^0} \bigcap A^2_{z^0}} p(x|r = Q_1(x) + \sigma^*(z^0 - Q_1(x))) \\
\leq& c_4 \sigma^{*2} \sigma^{*d_z} \left(\frac{1}{2\pi\sigma^{*2}}\right)^{\frac{d_x}{2}} \pi_{\mathcal{D}}(Q_1(x)) \exp(B).
\end{aligned}$$

Then, we only need to bound

$$\sigma^{*d_z} \int_{A^1_{z^0}} p(x|r)(\pi(Q_1(x)) + \sigma^* \nabla \pi(a_{z^0})^T(z^0 - Q_1(x)))dz^0.$$

We first bound

$$\sigma^{*d_z} \int_{A^1_{z^0}} p(x|r)\sigma^* \nabla \pi(a_{z^0})^T(z^0 - Q_1(x))dz^0.$$

By the assumption that $\|\nabla \log \pi_{\mathcal{D}}(z)\|_2 \leq c_1\|z\|_2 + c_2$, we have for any $x \in B_x$, $z^0 \in A^1_{z^0}$ and $a_{z^0} = Q_1(x) + c\sigma^*(z^0 - Q_1(x)) \quad (c \in [0,1])$,

$$\begin{aligned}
\frac{\|\nabla \pi_{\mathcal{D}}(a_{z^0})\|_2}{\pi_{\mathcal{D}}(Q_1(x))} &= \|\nabla \log \pi_{\mathcal{D}}(a_{z^0})\|_2 \exp(\log \pi_{\mathcal{D}}(a_{z^0}) - \log \pi_{\mathcal{D}}(Q_1(x))) \\
&\lesssim \sqrt{\log \frac{1}{\sigma^*}}.
\end{aligned}$$

And using the fact that

$$\begin{aligned}
p(x|z = r) = &\left((2\pi)^{d_z}|\Sigma(x)|\right)^{\frac{1}{2}} \mathcal{N}(z_0, Q_1(x) + \frac{\Sigma(x)}{\sigma^*}\nabla G_{\mathcal{D}}(Q_1(x))^T(x - G_{\mathcal{D}}(Q_1(x))), \Sigma(x)) \\
&\left(\frac{1}{2\pi\sigma^{*2}}\right)^{\frac{d_x}{2}} \exp(B)\exp(C + D + E),
\end{aligned}$$

with

$$\begin{aligned}
C &= \frac{(x - G_{\mathcal{D}}(Q_1(x)) - \sigma^* \nabla G_{\mathcal{D}}(Q_1(x))(z^0 - Q_1(x))^T}{\sigma^{*2}} R_n(x, z^0) - \frac{R_n(x, z^0)^T R_n(x, z^0)}{2\sigma^{*2}};\\
D &= -\frac{\|G_{\mathcal{D}}(r) - G_1(r)\|_2^2}{2\sigma^{*2}};\\
E &= -\frac{(G_{\mathcal{D}}(r) - G_1(r))^T (x - G_{\mathcal{D}}(r))}{\sigma^{*2}}.
\end{aligned}$$

We could then obtain,

$$\begin{aligned}
&\int_{A^1_{z^0}} p(x|r) \left\|z^0 - Q_1(x)\right\|_2 dz^0\\
&\lesssim \left((2\pi)^{d_z} |\Sigma(x)|\right)^{\frac{1}{2}} \left(\frac{1}{2\pi\sigma^{*2}}\right)^{\frac{d_x}{2}} \exp(B) \exp\left(c_1 \left(\sigma^* \left(\log\frac{1}{\sigma^*}\right)^{\frac{2}{\alpha}+3\alpha_1} + \epsilon_0 \left(\log\frac{1}{\sigma^*}\right)^{\frac{2}{\alpha}+2\alpha_1}\right)\right)\\
&\times \exp\left(c_2 \frac{\epsilon_1}{\sigma^*} \left(\log\frac{1}{\sigma^*}\right)^{\frac{1}{\alpha}+\alpha_1}\right) \left(\log\frac{1}{\sigma^*}\right)^{\alpha_1}.
\end{aligned}$$

Then we have,

$$-\sigma^* \left(\log\frac{1}{\sigma^*}\right)^{\frac{1}{2}+\alpha_1} \lesssim \frac{\sigma^{*d_z} \int_{A^1_{z^0}} p(x|r) \sigma^* \nabla \pi_{\mathcal{D}}(a_{z^0})^T (z^0 - Q_1(x)) dz^0}{((2\pi)^{d_z} |\Sigma(x)|)^{\frac{1}{2}} \left(\frac{1}{2\pi\sigma^{*2}}\right)^{\frac{d_x}{2}} \exp(B) \sigma^{*d_z} \pi_{\mathcal{D}}(Q_1(x))} \lesssim \sigma^* \left(\log\frac{1}{\sigma^*}\right)^{\frac{1}{2}+\alpha_1}.$$

Next we bound

$$\sigma^{*d_z} \int_{A^1_{z^0}} p(x|r) \pi_{\mathcal{D}}(Q_1(x)) dz^0.$$

Since for sufficient large $A^1_{z^0}$, we have

$$\int_{A^1_{z^0}} \mathcal{N}(Q_1(x) + \frac{\Sigma(x)}{\sigma^*} \nabla G_{\mathcal{D}}(Q_1(x))^T (x - G_{\mathcal{D}}(Q_1(x))), \Sigma(x)) \geq 1 - \sigma^{*2}.$$

We could then obtain

$$\begin{aligned}
&\exp\left(-c_1 \left(\sigma^* \left(\log\frac{1}{\sigma^*}\right)^{\frac{2}{\alpha}+3\alpha_1} + \epsilon_0 \left(\log\frac{1}{\sigma^*}\right)^{\frac{2}{\alpha}+2\alpha_1}\right)\right) \exp\left(-c_2 \frac{\epsilon_1}{\sigma^*} \left(\log\frac{1}{\sigma^*}\right)^{\frac{1}{\alpha}+\alpha_1}\right) (1 - \sigma^{*2})\\
&\lesssim \frac{\sigma^{*d_z} \int_{A^1_{z^0}} p(x|r) \pi_{\mathcal{D}}(Q_1(x)) dz^0}{((2\pi)^{d_z} |\Sigma(x)|)^{\frac{1}{2}} \left(\frac{1}{2\pi\sigma^{*2}}\right)^{\frac{d_x}{2}} \exp(B) \sigma^{*d_z} \pi_{\mathcal{D}}(Q_1(x))}\\
&\lesssim \exp\left(c_1 \left(\sigma^* \left(\log\frac{1}{\sigma^*}\right)^{\frac{2}{\alpha}+3\alpha_1} + \epsilon_0 \left(\log\frac{1}{\sigma^*}\right)^{\frac{2}{\alpha}+2\alpha_1}\right)\right) \exp\left(c_2 \frac{\epsilon_1}{\sigma^*} \left(\log\frac{1}{\sigma^*}\right)^{\frac{1}{\alpha}+\alpha_1}\right).
\end{aligned}$$

Then by the fact that

$$\begin{aligned}
&\sigma^{*d_z}\int_{A^1_{z^0}} p(x|r)(\pi_{\mathcal{D}}(Q_1(x)) + \sigma^*\nabla\pi_{\mathcal{D}}(a_{z^0})^T(z^0 - Q_1(x)))dz^0 \\
&\leq p(x) \leq \sigma^{*d_z}\int_{A^1_{z^0}} p(x|r)(\pi_{\mathcal{D}}(Q_1(x)) + \sigma^*\nabla\pi_{\mathcal{D}}(a_{z^0})^T(z^0 - Q_1(x)))dz^0 \\
&+ \int_{\overline{B}^c_z} p(x|z=r)\pi_{\mathcal{D}}(z=r)dr + \sigma^{*d_z}\int_{A^{1,c}_{z^0}\bigcap A^2_{z^0}} p(x|r)\pi_{\mathcal{D}}(r)dz^0.
\end{aligned}$$

We could then get the conclusion of the first part of the lemma. For the second part of the lemma, since

$$\int_{A^c_{z^0}} p^0(z^0|x)dz^0 = \frac{\int_{A^c_{z^0}} p(x|z=r)\pi_{\mathcal{D}}(z=r)dz^0}{\int p(x|z=r)\pi_{\mathcal{D}}(z=r)dz^0},$$

we can get the desired conclusion using the same strategy of the proof of the first part of the lemma.

C.5.7. PROOF OF LEMMA 30

Since

$$\begin{aligned}
\log p_{\mathcal{D}}(x) &\leq \frac{d_x}{2}\log(\frac{1}{2\pi\sigma^{*2}}); \\
-\log p(x) &= \frac{d_x}{2}\log(2\pi\sigma^{*2}) - \log\int \exp\left(-\frac{(x-G_1(z))^T(x-G_1(z))}{2\sigma^{*2}}\right)\pi_{\mathcal{D}}(z)dz \qquad (19) \\
&\leq \frac{d_x}{2}\log(2\pi\sigma^{*2}) + \frac{\|x\|_2^2}{\sigma^{*2}} + \frac{1}{\sigma^{*2}}\int \|G_1(z)\|_2^2\pi_{\mathcal{D}}(z)dz,
\end{aligned}$$

where the last inequality is due to Jensen inequality. So for $B_z = [-\eta(\log\frac{1}{\sigma^*})^{\frac{1}{2}}, \eta(\log\frac{1}{\sigma^*})^{\frac{1}{2}}]^{d_z}$, $B_\epsilon = [-\gamma(\log\frac{1}{\sigma^*})^{\frac{1}{2}}, \gamma(\log\frac{1}{\sigma^*})^{\frac{1}{2}}]^{d_x}$ and $B_x = \{G_{\mathcal{D}}(z) + \sigma^*\epsilon, z \in B_z, \epsilon \in B_\epsilon\}$, if η and γ are large enough, by the assumption that $\epsilon_1 \leq \sigma^{*2}$, we have

$$\int_{B^c_x} \log\frac{p_{\mathcal{D}}(x)}{p(x)}p_{\mathcal{D}}(x) - p_{\mathcal{D}}(x) + p(x)dx \leq \sigma^{*2}.$$

Also, there exists a constant c such that when $x \in B_x$, it holds that

$$\begin{aligned}
p_{\mathcal{D}}(x) &\gtrsim \exp\left(-c\log\frac{1}{\sigma^*}\right); \\
p(x) &\gtrsim \exp\left(-c\log\frac{1}{\sigma^*}\right).
\end{aligned}$$

We then consider a compact set of ϵ and z: $\tilde{B}_\epsilon = [-\bar{c}_1(\log\frac{1}{\sigma^*})^{\frac{1}{2}}, \bar{c}_1(\log\frac{1}{\sigma^*})^{\frac{1}{2}}]^{d_x}$ and $\tilde{B}_z = [-\bar{c}_2(\log\frac{1}{2})^{\frac{1}{\alpha}}, \bar{c}_2(\log\frac{1}{\sigma^*})^{\frac{1}{2}}]^{d_z}$ with $\tilde{B}_z \subset \overline{B}_z$. we can obtain

$$\begin{aligned}
&D_{\mathrm{KL}}(p_{\mathcal{D}}(\cdot)||p(\cdot))\\
&= \int \left(\log\frac{p_{\mathcal{D}}(x)}{p(x)}\frac{p_{\mathcal{D}}(x)}{p(x)} - \frac{p_{\mathcal{D}}(x)}{p(x)} + 1\right)p(x)dx\\
&\le \int_{\tilde{B}_z\bigcap\tilde{B}_\epsilon\bigcap B_x}\left(\frac{p_{\mathcal{D}}(G_1(z)+\sigma^*\epsilon)}{p(G_1(z)+\sigma^*\epsilon)} - 1\right)^2\pi_{\mathcal{D}}(z)p(\epsilon)dzd\epsilon\\
&+ \int_{\tilde{B}_z\bigcap\tilde{B}_\epsilon^c\bigcap B_x}\left(\log\frac{p_{\mathcal{D}}(G_1(z)+\sigma^*\epsilon)}{p(G_1(z)+\sigma^*\epsilon)}\frac{p_{\mathcal{D}}(G_1(z)+\sigma^*\epsilon)}{p(G_1(z)+\sigma^*\epsilon)} - \frac{p_{\mathcal{D}}(G_1(z)+\sigma^*\epsilon)}{p(G_1(z)+\sigma^*\epsilon)} + 1\right)\pi_{\mathcal{D}}(z)p(\epsilon)dzd\epsilon\\
&+ \int_{\tilde{B}_z^c\bigcap B_x}\left(\log\frac{p_{\mathcal{D}}(G_1(z)+\sigma^*\epsilon)}{p(G_1(z)+\sigma^*\epsilon)}\frac{p_{\mathcal{D}}(G_1(z)+\sigma^*\epsilon)}{p(G_1(z)+\sigma^*\epsilon)} - \frac{p_{\mathcal{D}}(G_1(z)+\sigma^*\epsilon)}{p(G_1(z)+\sigma^*\epsilon)} + 1\right)\pi_{\mathcal{D}}(z)p(\epsilon)dzd\epsilon\\
&+ {\sigma^*}^2.
\end{aligned} \tag{20}$$

Where we also reserve the notation B_x to be the set $\{(z,\epsilon)\,|\,x = G_1(z)+\sigma_1\epsilon \in B_x\}$.

For the second and third part of equation (20), by (1) ϵ is gaussian noise with mean 0 and identity covariance; (2) for $Z \sim \pi_{\mathcal{D}}(z)$, $\max_{1\le j\le d_z}\|Z^j\|_{\psi_2}$ is bounded; (3) when $x \in B_x$, $p_{\mathcal{D}}(x) \gtrsim \exp\left(-c\log\frac{1}{\sigma^*}\right)$ and $p(x) \gtrsim \exp\left(-c\log\frac{1}{\sigma^*}\right)$. we can get that when $(\bar{c}_1, \bar{c}_2)$ are large enough, the second and third part of equation (20) can be upper bounded by ${\sigma^*}^2$.

For the first part of equation (20), since

$$\begin{aligned}
&\frac{p_{\mathcal{D}}(G_1(z)+\sigma^*\epsilon)}{p(G_1(z)+\sigma^*\epsilon)}\\
=&\frac{\int\exp\left(-\frac{(G_1(z)+\sigma^*\epsilon-G_{\mathcal{D}}(z'))^T(G_1(z)+\sigma^*\epsilon-G_{\mathcal{D}}(z'))}{2{\sigma^*}^2}\right)\pi_{\mathcal{D}}(z')dz'}{\int\exp\left(-\frac{(G_1(z)+\sigma^*\epsilon-G_1(z'))^T(G_1(z)+\sigma^*\epsilon-G_1(z'))}{2{\sigma^*}^2}\right)\pi_{\mathcal{D}}(z')dz'}.
\end{aligned}$$

We first consider the numerator, define

$$B_{\sigma^*}(z,\bar{c}_3) = \left[z - \bar{c}_3\sigma^*\left(\log\frac{1}{\sigma^*}\right)^{\frac{2}{\alpha^2}+\frac{1}{2}}\mathbf{1}_{d_z}, z + \bar{c}_3\sigma^*\left(\log\frac{1}{\sigma^*}\right)^{\frac{2}{\alpha^2}+\frac{1}{2}}\mathbf{1}_{d_z}\right].$$

Therefore by the fact that under $\overline{B}_z$, $\|G_{\mathcal{D}}(z) - G_{\mathcal{D}}(z')\|_2^2 \ge a\left(\log\frac{1}{\sigma^*}\right)^{\frac{4}{\alpha^2}}\|z - z'\|_2^2$, $\|G_1(z) - G_{\mathcal{D}}(z)\|_2 \le \epsilon_1$ and $(a-b)^2 \ge \frac{1}{2}a^2 - b^2$, we can get

$$\begin{aligned}
&\int_{B_{\sigma^*}(z,\bar{c}_3)^c\bigcap\overline{B}_z}\exp\left(-\frac{(G_1(z)+\sigma^*\epsilon-G_{\mathcal{D}}(z'))^T(G_1(z)+\sigma^*\epsilon-G_{\mathcal{D}}(z'))}{2{\sigma^*}^2}\right)\pi_{\mathcal{D}}(z')dz'\\
&\le \exp\left(\frac{\epsilon_1^2}{{\sigma^*}^2} - \left(\frac{dz}{4}\bar{c}_3^2a - d_x\bar{c}_1^2\right)\log\frac{1}{\sigma^*}\right).
\end{aligned}$$

Also,

$$\begin{aligned}&\int_{\overline{B}_z^c} \exp\left(-\frac{(G_1(z)+\sigma^*\epsilon - G_{\mathcal{D}}(z'))^T(G_1(z)+\sigma^*\epsilon - G_{\mathcal{D}}(z'))}{2\sigma^{*2}}\right)\pi_{\mathcal{D}}(z')dz' \\ &\leq \pi_{\mathcal{D}}(\overline{B}_z^c).\end{aligned}$$

Then, by the fact that when $x \in B_x$, $p(x) \gtrsim \exp\left(-c\log\frac{1}{\sigma^*}\right)$, we can choose a large enough $\bar{c}_3$ and $\overline{\eta}$, such that

$$\begin{aligned}&\left(\int_{B_{\sigma^*}(z,\bar{c}_3)^c \bigcap \overline{B}_z} \exp\left(-\frac{(G_1(z)+\sigma^*\epsilon - G_{\mathcal{D}}(z'))^T(G_1(z)+\sigma^*\epsilon - G_{\mathcal{D}}(z'))}{2\sigma^{*2}}\right)\pi_{\mathcal{D}}(z')dz'\right. \\ &\left.+\int_{\overline{B}_z^c} \exp\left(-\frac{(G_1(z)+\sigma^*\epsilon - G_{\mathcal{D}}(z'))^T(G_1(z)+\sigma^*\epsilon - G_{\mathcal{D}}(z'))}{2\sigma^{*2}}\right)\pi_{\mathcal{D}}(z')dz\right)\exp\left(c\log\frac{1}{\sigma^*}\right) \leq \sigma^*.\end{aligned}$$

So we have

$$\begin{aligned}&\frac{p_{\mathcal{D}}(G_1(z)+\sigma^*\epsilon)}{p(G_1(z)+\sigma^*\epsilon)} \\ \leq&\frac{\int_{B_{\sigma^*}(z,c_3)} \exp\left(-\frac{(G_1(z)+\sigma^*\epsilon - G_{\mathcal{D}}(z'))^T(G_1(z)+\sigma^*\epsilon - G_{\mathcal{D}}(z'))}{2\sigma^{*2}}\right)\pi_{\mathcal{D}}(z')dz'}{\int \exp\left(-\frac{(G_1(z)+\sigma^*\epsilon - G_1(z'))^T(G_1(z)+\sigma^*\epsilon - G_1(z'))}{2\sigma^{*2}}\right)\pi_{\mathcal{D}}(z')dz'} + \sigma^* \\ &\leq \sup_{z'\in B_{\sigma^*}(z,c_3)} \frac{\exp\left(-\frac{(G_1(z)+\sigma^*\epsilon - G_{\mathcal{D}}(z'))^T(G_1(z)+\sigma^*\epsilon - G_{\mathcal{D}}(z'))}{2\sigma^{*2}}\right)}{\exp\left(-\frac{(G_1(z)+\sigma^*\epsilon - G_1(z'))^T(G_1(z)+\sigma^*\epsilon - G_1(z'))}{2\sigma^{*2}}\right)} + \sigma^* \\ &\leq \exp\left(\frac{\epsilon_1^2}{2\sigma^{*2}} + \frac{\epsilon_1}{\sigma^*}\left(\bar{c}_1\sqrt{d_x}(\log\frac{1}{\sigma^*})^{\frac{1}{2}} + \bar{c}_3\sqrt{d_z}(\log\frac{1}{\sigma^*})^{\frac{2}{\alpha^2}+\frac{1}{2}+\frac{1}{\alpha}} + \epsilon_1\right)\right) + \sigma^*.\end{aligned}$$

Therefore we can get

$$\log\frac{p_{\mathcal{D}}(G_1(z)+\sigma^*\epsilon)}{p(G_1(z)+\sigma^*\epsilon)} \lesssim \sigma^* + \frac{\epsilon_1}{\sigma^*}(\log\frac{1}{\sigma^*})^{\frac{2}{\alpha^2}+\frac{1}{2}+\frac{1}{\alpha}}.$$

Similarly,

$$\log\frac{p(G_1(z)+\sigma^*\epsilon)}{p_{\mathcal{D}}(G_1(z)+\sigma^*\epsilon)} \lesssim \sigma^* + \frac{\epsilon_1}{\sigma^*}(\log\frac{1}{\sigma^*})^{\frac{2}{\alpha^2}+\frac{1}{2}+\frac{1}{\alpha}}.$$

So we can bound the first part of equation (20) by $O\left(\frac{\epsilon_1^2}{\sigma^2}(\log\frac{1}{\sigma^*})^{\frac{4}{\alpha^2}+\frac{2}{\alpha}+1} + \sigma^{*2}\right)$. We can then get the desired conclusion by combining all those facts.