A Spectral Approach to Gradient Estimation for Implicit Distributions: Appendix

A. Proof of Proposition 1

We first introduce the following lemmas.

Lemma 1 (Liu et al. 2016, Proposition 3.5)**.** *Let $\mathcal{H}$ denote the Reproducing Kernel Hilbert Space (RKHS) induced by kernel k. If $k(\cdot,\cdot)$ has continuous second order partial derivatives, and both $k(\mathbf{x},\cdot)$ and $k(\cdot,\mathbf{x})$ are in the Stein class of q for any fixed $\mathbf{x}$, then $\forall f \in \mathcal{H}$, f is in the Stein class of q.*

Lemma 2 (Mercer's theorem)**.** *Let k be a continuous kernel on compact metric space $\mathcal{X}$. q is a finite Borel measure on $\mathcal{X}$. Then for $\{\psi_j\}_{j\geq 1}$ that satisfy eq. (1), $\forall \mathbf{x},\mathbf{y} \in \mathcal{X}$:*

$$k(\mathbf{x},\mathbf{y}) = \sum_j \mu_j \psi_j(\mathbf{x})\psi_j(\mathbf{y}).$$

Proof. See Sejdinovic & Gretton, Theorem 50. □

Lemma 3 (Sejdinovic & Gretton, Theorem 51)**.** *Let $\mathcal{X}$ be a compact metric space and k: $\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ be a continuous kernel, Define:*

$$\mathcal{H} = \left\{ f = \sum_i a_i \psi_i : \left\{ \frac{a_i}{\sqrt{\mu_i}} \right\} \in \ell^2 \right\}.$$

Then $\mathcal{H}$ is the same space as the RKHS induced by k.

Then we prove Proposition 1.

Proof. In Lemma 3 we set $a_j = 1$, $a_i = 0$ $(\forall i \neq j)$, then we have $\psi_j \in \mathcal{H}$. According to Lemma 1, we can conclude that ψ_j is in the Stein class of q . □

B. Error Bound of SSGE

Define

$$g_i(\mathbf{x}) = \sum_{j=1}^{\infty} \beta_{ij}\psi_j(\mathbf{x}), \quad g_{i,J}(\mathbf{x}) = \sum_{j=1}^{J} \beta_{ij}\psi_j(\mathbf{x}), \quad \tilde{g}_{i,J}(\mathbf{x}) = \sum_{j=1}^{J} \beta_{ij}\hat{\psi}_j(\mathbf{x}), \quad \hat{g}_{i,J}(\mathbf{x}) = \sum_{j=1}^{J} \hat{\beta}_{ij}\hat{\psi}_j(\mathbf{x}), \tag{27}$$

which correspond to the major approximations in each step.

Lemma 4 (Sinha & Belkin 2009; Izbicki et al. 2014; Rosasco et al. 2010)**.** *For all $1 \leq j \leq J$,*

$$\int \left(\hat{\psi}_j(\mathbf{x}) - \psi_j(\mathbf{x})\right)^2 dq = O_p\left(\frac{1}{\mu_j \delta_j^2 M}\right), \tag{28}$$

where $\delta_j = \mu_j - \mu_{j+1}$.

Lemma 5 (Sinha & Belkin 2009; Izbicki et al. 2014)**.** *For all $1 \leq j \leq J$,*

$$\int \hat{\psi}_j(\mathbf{x})^2 \, dq = O_p\left(\frac{1}{\mu_j \Delta_J^2 M}\right) + 1, \tag{29}$$

and for all $1 \leq i \leq J, i \neq j$,

$$\int \hat{\psi}_i(\mathbf{x})\hat{\psi}_j(\mathbf{x})\, dq = O_p\left(\left(\frac{1}{\sqrt{\mu_i}} + \frac{1}{\sqrt{\mu_j}}\right)\frac{1}{\Delta_J\sqrt{M}}\right), \tag{30}$$

where $\Delta_J = \min_{1\leq j\leq J}\delta_j$.

Lemma 6.

$$\int |\tilde{g}_{i,J}(\mathbf{x}) - g_{i,J}(\mathbf{x})|^2\ dq = JO_p\left(\frac{C}{\mu_J\Delta_J^2 M}\right). \tag{31}$$

Proof. By Cauchy-Schwartz inequality, Assumption 2 and Lemma 4:

$$\begin{aligned}\int |\tilde{g}_{i,J}(\mathbf{x}) - g_{i,J}(\mathbf{x})|^2\ dq &= \int \left|\sum_{j=1}^{J}\beta_{ij}\left(\psi_j(\mathbf{x}) - \hat{\psi}_j(\mathbf{x})\right)\right|^2 dq \\ &\leq \left(\sum_{j=1}^{J}\beta_{ij}^2\right)\left(\sum_{j=1}^{J}\int\left(\psi_j(\mathbf{x}) - \hat{\psi}_j(\mathbf{x})\right)^2\ dq\right) \\ &= JO_p\left(\frac{C}{\mu_J\Delta_J^2 M}\right).\end{aligned}$$

□

Lemma 7. *For all* $1 \leq j \leq J$,

$$\left(\int\left(\nabla_{x_i}\psi_j(\mathbf{x}) - \nabla_{x_i}\hat{\psi}_j(\mathbf{x})\right)\ dq\right)^2 = O_p\left(\frac{C}{\mu_j\delta_j^2 M}\right). \tag{32}$$

Proof. Denote $\delta(\mathbf{x}) = \psi_j(\mathbf{x}) - \hat{\psi}_j(\mathbf{x})$. According to Assumption 1, it is easy to see that $\hat{\psi}_j(\mathbf{x})$ satisfies the boundary condition:

$$\int \nabla_{\mathbf{x}}[\hat{\psi}_j(\mathbf{x})q(\mathbf{x})]d\mathbf{x} = \mathbf{0}. \tag{33}$$

And from the proof of Proposition 1, we know $\psi_j(\mathbf{x})$ satisfies the boundary condition. Combining the two, we have:

$$\int \nabla_{\mathbf{x}}[\delta(\mathbf{x})q(\mathbf{x})]\ d\mathbf{x} = \mathbf{0}. \tag{34}$$

By eq. (34), Lemma 4 and Assumption 2, we have

$$\begin{aligned}\left(\int \nabla_{x_i}\delta(\mathbf{x})dq\right)^2 &= \left(\int \nabla_{x_i}[\delta(\mathbf{x})q(\mathbf{x})] - \delta(\mathbf{x})\nabla_{x_i}q(\mathbf{x})\ d\mathbf{x}\right)^2 \\ &= \left(\int \delta(\mathbf{x})\nabla_{x_i}\log q(\mathbf{x})\ dq\right)^2 \\ &\leq \left(\int \delta(\mathbf{x})^2\ dq\right)\left(\int g_i(\mathbf{x})^2\ dq\right) \\ &= O_p\left(\frac{C}{\mu_j\delta_j^2 M}\right).\end{aligned} \tag{35}$$

□

Lemma 8. *For all* $1 \leq j \leq J$,

$$(\beta_{ij} - \hat{\beta}_{ij})^2 = O_p\left(\frac{1}{M}\right) + O_p\left(\frac{C}{\mu_j\delta_j^2 M}\right). \tag{36}$$

Proof.

$$\begin{aligned}\frac{1}{2}(\beta_{ij}-\hat{\beta}_{ij})^2 &\leq \left(\beta_{ij}-\frac{1}{M}\sum_{m=1}^{M}\nabla_{x_i}\psi_j(\mathbf{x}^m)\right)^2+\left(\frac{1}{M}\sum_{m=1}^{M}\left(\nabla_{x_i}\psi_j(\mathbf{x}^m)-\nabla_{x_i}\hat{\psi}_j(\mathbf{x}^m)\right)\right)^2\\ &\leq O_p\left(\frac{1}{M}\right)+2\left[\frac{1}{M}\sum_{m=1}^{M}\left(\nabla_{x_i}\psi_j(\mathbf{x}^m)-\nabla_{x_i}\hat{\psi}_j(\mathbf{x}^m)\right)-\int\left(\nabla_{x_i}\psi_j(\mathbf{x})-\nabla_{x_i}\hat{\psi}_j(\mathbf{x})\right)\,dq\right]^2\\ &\quad+2\left[\int\left(\nabla_{x_i}\psi_j(\mathbf{x})-\nabla_{x_i}\hat{\psi}_j(\mathbf{x})\right)\,dq\right]^2\\ &= O_p\left(\frac{1}{M}\right)+2O_p\left(\frac{1}{M}\right)+2\left(\int\left(\nabla_{x_i}\psi_j(\mathbf{x})-\nabla_{x_i}\hat{\psi}_j(\mathbf{x})\right)dq\right)^2.\end{aligned} \tag{37}$$

Therefore, by Lemma 7 we have

$$(\beta_{ij}-\hat{\beta}_{ij})^2=O_p\left(\frac{1}{M}\right)+O_p\left(\frac{C}{\mu_j\delta_j^2M}\right). \tag{38}$$

□

Lemma 9.

$$\int|\tilde{g}_{i,J}(\mathbf{x})-\hat{g}_{i,J}(\mathbf{x})|^2\,dq=J^2\left(O_p\left(\frac{1}{M}\right)+O_p\left(\frac{C}{\mu_j\Delta_j^2M}\right)\right) \tag{39}$$

Proof. By applying Minkowski inequality, Cauchy-Schwartz inequality, Lemma 8 and Lemma 5, we have

$$\begin{aligned}&\int|\tilde{g}_{i,J}(\mathbf{x})-\hat{g}_{i,J}(\mathbf{x})|^2\,dq=\int\left|\sum_{j=1}^{J}\beta_{ij}\hat{\psi}_j(\mathbf{x})-\sum_{j=1}^{J}\hat{\beta}_{ij}\hat{\psi}_j(\mathbf{x})\right|^2dq=\int\left|\sum_{j=1}^{J}(\beta_{ij}-\hat{\beta}_{ij})\hat{\psi}_j(\mathbf{x})\right|^2dq\\ &\leq\left\{\sum_{j=1}^{J}\left[\int\left|(\beta_{ij}-\hat{\beta}_{ij})\hat{\psi}_j(\mathbf{x})\right|^2dq\right]^{\frac{1}{2}}\right\}^2\leq\left\{\sum_{j=1}^{J}\left[\int\left|(\beta_{ij}-\hat{\beta}_{ij})\right|^2dq\int\hat{\psi}_j^2(\mathbf{x})\,dq\right]^{\frac{1}{2}}\right\}^2\\ &=\left\{\sum_{j=1}^{J}\left[O_p\left(\frac{1}{M}\right)+O_p\left(\frac{C}{\mu_j\delta_j^2M}\right)\right]^{\frac{1}{2}}\left[O_p\left(\frac{1}{\mu_j\Delta_J^2M}\right)+1\right]^{\frac{1}{2}}\right\}^2\\ &=J^2\left(O_p\left(\frac{1}{M}\right)+O_p\left(\frac{C}{\mu_J\Delta_J^2M}\right)\right)\end{aligned} \tag{40}$$

□

Theorem 3 (Estimation Error)**.**

$$\int|\hat{g}_{i,J}(\mathbf{x})-g_{i,J}(\mathbf{x})|^2\,\,dq=J^2\left(O_p\left(\frac{1}{M}\right)+O_p\left(\frac{C}{\mu_J\Delta_J^2M}\right)\right)+JO_p\left(\frac{C}{\mu_J\Delta_J^2M}\right) \tag{41}$$

Proof. By lemma 6 and lemma 9.

$$\begin{aligned}\int|\hat{g}_{i,J}(\mathbf{x})-g_{i,J}(\mathbf{x})|^2\,\,dq&\leq\int|\tilde{g}_{i,J}(\mathbf{x})-g_{i,J}(\mathbf{x})|^2\,\,dq+\int|\tilde{g}_{i,J}(\mathbf{x})-\hat{g}_{i,J}(\mathbf{x})|^2\,dq\\ &=J^2\left(O_p\left(\frac{1}{M}\right)+O_p\left(\frac{C}{\mu_J\Delta_J^2M}\right)\right)+JO_p\left(\frac{C}{\mu_J\Delta_J^2M}\right)\end{aligned} \tag{42}$$

□

Theorem 4 (Approximation Error)**.**

$$\int |g_{i,J}(\mathbf{x}) - g_i(\mathbf{x})|^2 \, dq = \|g_i\|_{\mathcal{H}}^2 O(\mu_J) \tag{43}$$

Proof.

$$\int |g_{i,J}(\mathbf{x}) - g_i(\mathbf{x})|^2 \, dq = \sum_{j>J} \beta_{ij}^2 = \sum_{j>J} \frac{\beta_{ij}^2}{\mu_j} \mu_j \leq \mu_J \sum_{j>J} \frac{\beta_{ij}^2}{\mu_j} = \mu_J \|g_i\|_{\mathcal{H}}^2 \tag{44}$$

□

Theorem 5 (Error Bound of SSGE)**.**

$$\int \left(\hat{g}_{i,J}(\mathbf{x}) - g_i(\mathbf{x})\right)^2 \, dq = J^2 \left(O_p\left(\frac{1}{M}\right) + O_p\left(\frac{C}{\mu_J \Delta_J^2 M}\right)\right) + J O_p\left(\frac{C}{\mu_J \Delta_J^2 M}\right) + \|g_i\|_{\mathcal{H}}^2 O(\mu_J). \tag{45}$$

Proof. By theorem 3 and theorem 4, we have

$$\begin{aligned} \int \left(\hat{g}_{i,J}(\mathbf{x}) - g_i(\mathbf{x})\right)^2 \, dq &\leq \int |\hat{g}_{i,J}(\mathbf{x}) - g_{i,J}(\mathbf{x})|^2 \, dq + \int |g_{i,J}(\mathbf{x}) - g_i(\mathbf{x})|^2 \, dq \\ &= J^2 \left(O_p\left(\frac{1}{M}\right) + O_p\left(\frac{C}{\mu_J \Delta_J^2 M}\right)\right) + J O_p\left(\frac{C}{\mu_J \Delta_J^2 M}\right) + \|g_i\|_{\mathcal{H}}^2 O(\mu_J) \end{aligned} \tag{46}$$

□

C. Derivation of Gradient Estimates for Entropy

First, we decompose the gradients into two terms:

$$\nabla_\phi \mathbb{H}(q) = -\nabla_\phi \mathbb{E}_{q_\phi} \log q_\phi(\mathbf{x}) = -\nabla_\phi \mathbb{E}_q \log q_\phi(\mathbf{x}) - \nabla_\phi \mathbb{E}_{q_\phi} \log q(\mathbf{x}).$$

Then it is easy to see that the first term is zero:

$$\nabla_\phi \mathbb{E}_q \log q_\phi(\mathbf{x}) = \int q(\mathbf{x}) \nabla_\phi \log q_\phi(\mathbf{x}) \, d\mathbf{x} = \int \nabla_\phi q_\phi(\mathbf{x}) \, d\mathbf{x} = \nabla_\phi \int q_\phi(\mathbf{x}) \, d\mathbf{x} = 0.$$

So we have

$$\nabla_\phi \mathbb{H}(q) = -\nabla_\phi \mathbb{E}_{q_\phi} \log q(\mathbf{x}). \tag{47}$$

A low-variance Monte-Carlo estimate of eq. (47) can be obtained when $q(\mathbf{x})$ is continuous and reparameterizable (Kingma & Welling, 2013). For example, if there exists a random variable $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, so that $\mathbf{x} = f(\boldsymbol{\epsilon}; \phi)$ follows the same distribution as $q(\mathbf{x})$, we have

$$\begin{aligned} \nabla_\phi \mathbb{H}(q) &= -\nabla_\phi \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \log q(f(\boldsymbol{\epsilon}; \phi)) \\ &= -\mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \nabla_\phi \log q(f(\boldsymbol{\epsilon}; \phi)) \\ &= -\mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \nabla_{\mathbf{x}} \log q(f(\boldsymbol{\epsilon}; \phi)) \nabla_\phi f(\boldsymbol{\epsilon}; \phi), \end{aligned}$$

where the intractable term $\nabla_{\mathbf{x}} \log q(f(\boldsymbol{\epsilon}; \phi))$ can be easily estimated by SSGE.

D. MNIST Results

We did an MNIST experiment on a VAE with an 8-dim latent space. In Figures 5a to 5c we plot random generations by a plain VAE, an Implicit VAE with the entropy term removed, and an Implicit VAE trained by SSGE, all with the same decoder structures. In Figure 5d we show the log likelihoods of the trained models (VAE and the Implicit VAE with SSGE) on 2048 test images using Annealed Importance Sampling (AIS). The results are averaged over 10 runs. For reference, we also include the results of 8-dim VAEs from the AVB paper (Mescheder et al., 2017). Though their decoder structure is not the same as ours (the structure is even not the same for their three models), we can see our method is slightly better than plain VAE without other tricks, while AVB has to rely on Adaptive Contrast (AC) and different decoder structures.

(a) VAE

(b) Implicit VAE, w/o entropy

(c) Implicit VAE, Spectral

METHOD	TEST LL.
VAE	-89.5 $\pm$ 0.6
SPECTRAL	-89.4 $\pm$ 0.7
VAE	-90.9 $\pm$ 0.6
AVB	-91.2 $\pm$ 0.6
AVB + AC	-89.6 $\pm$ 0.6

(d)

Figure 5. Results on the MNIST dataset: (a)-(c) Samples generated by VAE, Implicit VAE trained without the entropy term, and Implicit VAE trained by SSGE; (d) Test log likelihoods evaluated by AIS.