A. Proofs

Proof of Theorem 1. I) Recall that $T(\theta) = \theta + \epsilon\boldsymbol{\phi}(\theta)$ and ρ_T is the density of $T(\theta)$ when $\theta \sim \rho$. . When ϵ is sufficiently large, T is a one-to-one map and $|\det(\nabla T(\theta))| \neq 0$. By the change of variables formula, we have $\rho(\theta) = \rho_T(T(\theta))|\det(\nabla T(\theta))|$. Therefore,

$$
\begin{aligned}
&J[\rho_T] - J[\rho] - (F(\rho_T) - F[\rho]) \\
&= \alpha(\mathbb{H}[\rho_T] - \mathbb{H}[\rho]) \\
&= \alpha\bigg(- \int_\theta \rho_T(\theta)\log\rho_T(\theta)d\theta + \int_\theta \rho(\theta)\log\rho(\theta)d\theta\bigg) \\
&= \alpha\bigg(- \int_\theta \rho_T(T(\theta))\log\rho_T(T(\theta))dT(\theta) + \int_\theta \rho(\theta)\log\rho(\theta)d\theta\bigg) \\
&= \alpha(- \int_\theta \frac{\rho(\theta)}{|\det(\nabla T(\theta))|}\log\frac{\rho(\theta)}{|\det(\nabla T(\theta))|}|\det(\nabla T(\theta))|d\theta + \int_\theta \rho(\theta)\log\rho(\theta)d\theta) \\
&= \alpha\int_\theta \log(|\det(\nabla T(\theta)|)\rho(\theta)d\theta \\
&= \alpha\mathbb{E}_\rho[\log\det(|\nabla_\theta T(\theta)|)].
\end{aligned}
$$

Following the definition of $\mathcal{D}F$,

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\varepsilon}(J[\rho_T] - J[\rho])\bigg|_{\varepsilon=0} &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}(F[\rho_T] - F[\rho])\bigg|_{\varepsilon=0} \ + \ \alpha\frac{\mathrm{d}}{\mathrm{d}\varepsilon}(\mathbb{H}[\rho_T] - \mathbb{H}[\rho])\bigg|_{\varepsilon=0} \\
&= \mathbb{E}_{\theta\sim\rho}[\mathcal{D}F[\rho](\theta)^\top\boldsymbol{\phi}(\theta)] + \frac{\mathrm{d}}{\mathrm{d}\varepsilon}[\alpha\log(|\det(\nabla T(\theta))|)]\bigg|_{\varepsilon=0} \\
&= \mathbb{E}_{\theta\sim\rho}[\mathcal{D}F[\rho](\theta)^\top\boldsymbol{\phi}(\theta)] + \frac{\mathrm{d}}{\mathrm{d}\varepsilon}[\alpha\log(|\det(I + \varepsilon\nabla\boldsymbol{\phi}(\theta))|)]\bigg|_{\varepsilon=0} \\
&= \mathbb{E}_{\theta\sim\rho}[\mathcal{D}F[\rho](\theta)^\top\boldsymbol{\phi}(\theta)] + \mathbb{E}_{\theta\sim\rho}[\alpha\mathrm{trace}(\nabla\boldsymbol{\phi}(\theta))] \\
&= \mathbb{E}_{\theta\sim\rho}\left[\mathcal{D}F[\rho](\theta)^\top\boldsymbol{\phi}(\theta) \ + \ \alpha\nabla^\top\boldsymbol{\phi}(\theta)\right].
\end{aligned}
$$

II) Let θ_i and $\phi_i(\theta)$ are the i-th elements of $\theta = [\theta_i]_{i=1}^d$ and $\boldsymbol{\phi}(\theta) = [\phi_i(\theta)]_{i=1}^d$, respectively. By reproducing properties of RKHS with differentiable positive definite kernels (See Steinwart & Christmann (2008); Zhou (2008)), we have:

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\varepsilon}(J[\rho_T] - J[\rho])\bigg|_{\varepsilon=0} &= \sum_{i=1}^d \mathbb{E}_{\theta\sim\rho}\left[(\mathcal{D}F[\rho](\theta))_i\phi_i(\theta) \ + \ \alpha\frac{\partial\phi_i(\theta)}{\partial\theta_i}\right] \\
&= \sum_{i=1}^d \mathbb{E}_{\theta\sim\rho}\left[(\mathcal{D}F[\rho](\theta))_i\left\langle\phi_i, k(\cdot,\theta)\right\rangle_{\mathcal{H}} \ + \ \alpha\frac{\langle\phi_i, \partial k(\cdot,\theta)\rangle_{\mathcal{H}}}{\partial\theta_i}\right] \\
&= \sum_{i=1}^d \left\langle\phi_i, \mathbb{E}_{\theta\sim\rho}\left[(\mathcal{D}F[\rho](\theta))_i k(\cdot,\theta) \ + \ \alpha\frac{\partial k(\cdot,\theta)}{\partial\theta_i}\right]\right\rangle_{\mathcal{H}}
\end{aligned}
$$

By the Cauchy-Schwarz inequality, we can take the optimal solution $\boldsymbol{\phi}^*$ of (6) proportional to

$$
\boldsymbol{\phi}^*(\cdot) \propto \mathbb{E}_{\theta\sim\rho}\left[\mathcal{D}F[\rho](\theta)k(\theta,\cdot) + \alpha\nabla_\theta k(\theta,\cdot)\right].
$$

□

Proof of Theorem 3. Recall that ρ_T is the density of $T(\theta) = \theta + \varepsilon\boldsymbol{\phi}(\theta)$ when $\theta \sim \rho$. Following the Fokker-Plank equation (see Appendix A.3 in Liu (2017)), we have

$$
\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\rho_T(\theta)\bigg|_{\varepsilon=0} = -\nabla_\theta^\top(\boldsymbol{\phi}(\theta)\rho(\theta)),
$$

By integration by parts, we get:

$$\begin{aligned} F[\rho_T] &= F[\rho + (\rho_T - \rho)] = F[\rho - \varepsilon \nabla_\theta^\top (\boldsymbol{\phi}\rho) + O(\varepsilon^2)] \\ &= F[\rho] + \int_\theta \mathcal{L}F[\rho](\theta) \cdot (-\varepsilon \nabla_\theta^\top (\boldsymbol{\phi}(\theta)\rho(\theta)))d\theta + O(\varepsilon^2). \end{aligned}$$

Assume $\boldsymbol{\phi}$ is a continuous differentiable function with a compact support in $\mathbb{R}^d$. Using integration by parts, we have

$$\int_\theta \mathcal{L}F[\rho](\theta) \cdot \nabla_\theta^\top (\boldsymbol{\phi}(\theta)\rho(\theta)))d\theta = -\int_\theta \nabla_\theta (\mathcal{L}F[\rho](\theta))^\top \boldsymbol{\phi}(\theta)\rho(\theta)d\theta.$$

Therefore,

$$F[\rho_T] = F[\rho] + \varepsilon \mathbb{E}_{\theta\sim\rho}[\nabla_\theta (\mathcal{L}F[\rho](\theta))^\top \boldsymbol{\phi}(\theta)] + O(\varepsilon^2).$$

Comparing this with the definition of T-derivative, we have $\mathbb{E}_{\theta\sim\rho}[(\mathcal{D}F[\rho](\theta) - \nabla_\theta (\mathcal{L}F[\rho](\theta)))^\top \boldsymbol{\phi}(\theta)] = 0$. Because ρ is positive on $\mathbb{R}^d$ and this holds for any $\boldsymbol{\phi}$, we have $\mathcal{D}F[\rho](\theta) = \nabla_\theta (\mathcal{L}F[\rho](\theta))$. □

B. Deep Embedding Clustering

We present more details on the deep embedding clustering task. Let $C = [c_{\ell i}]_{\ell i} \in \mathbb{R}^{n\times m}$ stands for the clustering prediction, which can be written as

$$c_{\ell i} = \frac{\exp(\theta_i^\top z_\ell)}{\sum_{j=1}^m \exp(\theta_j^\top z_\ell)}.$$

DEPICT (Dizaji et al., 2017) is an iterative procedure which updates the parameters by minimizing a proximal point like KL divergence loss function:

$$\begin{aligned} R(\Theta) &= \mathrm{KL}(Q||C) + \mathrm{KL}(f||u) \\ &= \frac{1}{n}\sum_{\ell=1}^n \sum_{i=1}^m q_{\ell i} \log \frac{q_{\ell i}}{c_{\ell i}} + \frac{1}{n}\sum_{i=1}^m f_i \log \frac{f_i}{\mu_i}, \end{aligned}$$

where $f_i = \frac{1}{n}\sum_{\ell=1}^n q_{\ell i}$, $\mu_i = \frac{1}{m}$, and $q_{\ell i}$ is a distribution constructed based on the $C^{old} = [c_{\ell i}^{old}]$ from the last iteration:

$$q_{\ell i} = \frac{c_{\ell i}^{old}/(\sum_{\ell'} c_{\ell' i}^{old})^{1/2}}{\sum_{i'=1}^m \left(c_{\ell i'}^{old}/(\sum_{\ell'} c_{\ell' i'}^{old})^{1/2} \right)},$$

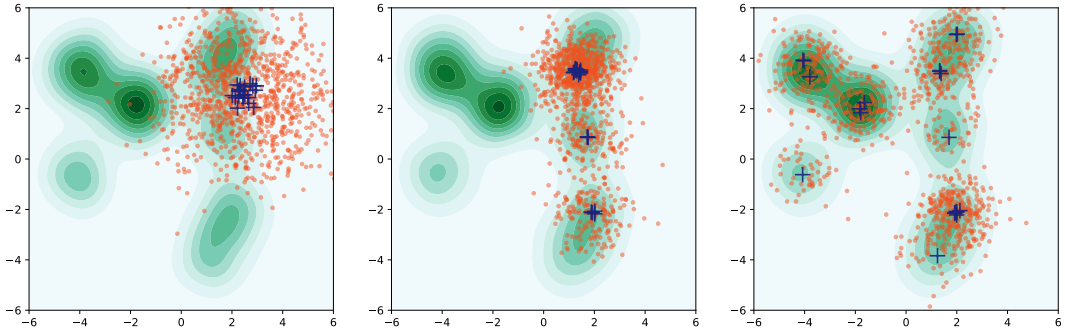
Here, the first term refines the cluster by learning from their high confidence assignments with the help of the auxiliary distribution Q and the second KL term balances the frequency of clusters. We refer readers to Dizaji et al. (2017) for more details.

C. Variational Inference with Nonlinear SVGD

To demonstrate the broad applicability of our method, here we present an experiment on learning diversified mixture models in variational inference settings (Blei et al., 2017). In this case, we are given the density of the target distribution $p(x)$ (instead of the data drawn from p), and we are interested in finding a mixture model $q_\rho(x) = \int q(x|\theta)\rho(\theta)d\theta$ to approximate p, by minimizing an entropy regularized KL divergence objective:

$$\min_{\rho} \mathrm{KL}(q_\rho \;||\; p) - \alpha\mathbb{H}[\rho],$$

where $\mathbb{H}[\rho]$ enforces the diversity of the components of q_ρ, and $\mathrm{KL}(q_\rho \;||\; p)$ enforces q_ρ to form a close approximation to p. For our experiment, we set the target distribution p to be a GMM $p(x) = \sum_{i=1}^{10} \frac{1}{10}\mathcal{N}(x; v_i, \mathrm{I})$, with $x \in \mathbb{R}^{10}$, and the elements of each mean vector v_i drawn from $\mathrm{uniform}[-4, 4]$. We set q_ρ to be Gaussian mixture model, $q(x|\theta) = \mathcal{N}(x;\ \mu,\ \sigma^2)$, and $\theta = [\mu, \sigma]$, and run the algorithm with 20 components, with initialization from $\mu_i \sim \mathrm{uniform}[2, 3]$ and $\sigma_i = 3$. As shown in Figure 4, applying our method (see Algorithm 1) to VI yields good mass-covering property and finds all modes in p accurately, while standard VI baseline lacks effective exploration and suffers from mode-collapsing.



(a) Initlizations (b) VI (c) VI + Ours

Figure 4. Results on VI with Gaussian mixture models. Green contours represent the KDE of the first two dimension of the true target distribution p; Blue "+" stands for the means $\{\mu_i\}$ of each Gaussian component of the proposal distribution q_Θ; Orange dots represent samples drawn from q.