Supplementary Material: Reparameterization Gradient for Non-differentiable Models

A Proof of Theorem 1

Using reparameterization, we can write ELBO_θ as follows:

$$\begin{aligned}
\mathsf{ELBO}_\theta &= \mathbb{E}_{q(\boldsymbol{\epsilon})}\left[\log \frac{\sum_{k=1}^{K} \mathbb{1}[f_\theta(\boldsymbol{\epsilon}) \in R_k] \cdot r_k(f_\theta(\boldsymbol{\epsilon}))}{q_\theta(f_\theta(\boldsymbol{\epsilon}))}\right] \\
&= \mathbb{E}_{q(\boldsymbol{\epsilon})}\left[\sum_{k=1}^{K} \mathbb{1}[f_\theta(\boldsymbol{\epsilon}) \in R_k] \cdot \log \frac{r_k(f_\theta(\boldsymbol{\epsilon}))}{q_\theta(f_\theta(\boldsymbol{\epsilon}))}\right] \qquad (6) \\
&= \sum_{k=1}^{K} \mathbb{E}_{q(\boldsymbol{\epsilon})}\left[\mathbb{1}[f_\theta(\boldsymbol{\epsilon}) \in R_k] \cdot h_k(\boldsymbol{\epsilon}, \theta)\right].
\end{aligned}$$

In (6), we can move the summation and the indicator function out of $\log$ since the regions $\{R_k\}_{1 \le k \le K}$ are disjoint. We then compute the gradient of ELBO_θ as follows:

$$\begin{aligned}
&\nabla_\theta \mathsf{ELBO}_\theta \\
&= \sum_{k=1}^{K} \nabla_\theta \mathbb{E}_{q(\boldsymbol{\epsilon})}\left[\mathbb{1}[f_\theta(\boldsymbol{\epsilon}) \in R_k] \cdot h_k(\boldsymbol{\epsilon}, \theta)\right] \\
&= \sum_{k=1}^{K} \nabla_\theta \int_{f_\theta^{-1}(R_k)} q(\boldsymbol{\epsilon}) h_k(\boldsymbol{\epsilon}, \theta) d\boldsymbol{\epsilon} \\
&= \sum_{k=1}^{K} \int_{f_\theta^{-1}(R_k)} \Big(q(\boldsymbol{\epsilon}) \nabla_\theta h_k(\boldsymbol{\epsilon}, \theta) + \nabla_{\boldsymbol{\epsilon}} \bullet \big(q(\boldsymbol{\epsilon}) h_k(\boldsymbol{\epsilon}, \theta) \boldsymbol{V}(\boldsymbol{\epsilon}, \theta)\big)\Big) d\boldsymbol{\epsilon} \qquad (7) \\
&= \mathbb{E}_{q(\boldsymbol{\epsilon})}\left[\sum_{k=1}^{K} \mathbb{1}[f_\theta(\boldsymbol{\epsilon}) \in R_k] \cdot \nabla_\theta h_k(\boldsymbol{\epsilon}, \theta)\right] + \sum_{k=1}^{K} \int_{f_\theta^{-1}(R_k)} \nabla_{\boldsymbol{\epsilon}} \bullet \big(q(\boldsymbol{\epsilon}) h_k(\boldsymbol{\epsilon}, \theta) \boldsymbol{V}(\boldsymbol{\epsilon}, \theta)\big) d\boldsymbol{\epsilon} \\
&= \underbrace{\mathbb{E}_{q(\boldsymbol{\epsilon})}\left[\sum_{k=1}^{K} \mathbb{1}[f_\theta(\boldsymbol{\epsilon}) \in R_k] \cdot \nabla_\theta h_k(\boldsymbol{\epsilon}, \theta)\right]}_{\mathsf{RepGrad}_\theta} + \underbrace{\sum_{k=1}^{K} \int_{f_\theta^{-1}(\partial R_k)} \big(q(\boldsymbol{\epsilon}) h_k(\boldsymbol{\epsilon}, \theta) \boldsymbol{V}(\boldsymbol{\epsilon}, \theta)\big) \bullet d\boldsymbol{\Sigma}}_{\mathsf{BouContr}_\theta} \qquad (8)
\end{aligned}$$

where $\nabla_{\boldsymbol{\epsilon}} \bullet \boldsymbol{U}$ denotes the column vector whose i-th component is $\nabla_{\boldsymbol{\epsilon}} \cdot \boldsymbol{U}_i$, the divergence of $\boldsymbol{U}_i$ with respect to $\boldsymbol{\epsilon}$. (8) is the formula that we wanted to prove.

The two non-trivial steps in the above derivation are (7) and (8). First, (7) is a direct consequence of the following theorem, existing yet less well-known, on exchanging integration and differentiation under moving domain:

Theorem 6. *Let $D_\theta \subset \mathbb{R}^n$ be a smoothly parameterized region. That is, there exist open sets $\Omega \subset \mathbb{R}^n$ and $\Theta \subset \mathbb{R}$, and twice continuously differentiable $\widehat{\boldsymbol{\epsilon}} : \Omega \times \Theta \to \mathbb{R}^n$ such that $D_\theta = \widehat{\boldsymbol{\epsilon}}(\Omega, \theta)$ for each $\theta \in \Theta$. Suppose that $\widehat{\boldsymbol{\epsilon}}(\cdot, \theta)$ is a C^1-diffeomorphism for each $\theta \in \Theta$. Let $f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ be a differentiable function such that $f(\cdot, \theta) \in \mathcal{L}^1(D_\theta)$ for each $\theta \in \Theta$. If there exists $g : \Omega \to \mathbb{R}$ such that $g \in \mathcal{L}^1(\Omega)$ and $\left|\nabla_\theta\big(f(\widehat{\boldsymbol{\epsilon}}, \theta)\big|\frac{\partial \widehat{\boldsymbol{\epsilon}}}{\partial \boldsymbol{\omega}}\big|\big)\right| \le g(\boldsymbol{\omega})$ for any $\theta \in \Theta$ and $\boldsymbol{\omega} \in \Omega$, then*

$$\nabla_\theta \int_{D_\theta} f(\boldsymbol{\epsilon}, \theta) d\boldsymbol{\epsilon} = \int_{D_\theta} \Big(\nabla_\theta f + \nabla_{\boldsymbol{\epsilon}} \cdot (f \mathbf{v})\Big)(\boldsymbol{\epsilon}, \theta) d\boldsymbol{\epsilon}.$$

Here $\mathbf{v}(\boldsymbol{\epsilon}, \theta)$ denotes $\nabla_\theta \widehat{\boldsymbol{\epsilon}}(\boldsymbol{\omega}, \theta)\big|_{\boldsymbol{\omega} = \widehat{\boldsymbol{\epsilon}}_\theta^{-1}(\boldsymbol{\epsilon})}$, the velocity of the particle $\boldsymbol{\epsilon}$ at time θ.

The statement of Theorem 6 (without detailed conditions as we present above) and the sketch of its proof can be found in [3]. One subtlety in applying Theorem 6 to our case is that R_k (which corresponds to Ω in the theorem) may not be open, so the theorem may not be immediately applicable. However, since the boundary ∂R_k has Lebesgue measure zero in $\mathbb{R}^n$, ignoring the reparameterized boundary $f_\theta^{-1}(\partial R_k)$ in the integral of (7) does not change the value of the integral. Hence, we apply Theorem 6 to $D_\theta = \mathrm{int}(f_\theta^{-1}(R_k))$ (which is possible because $\Omega = \mathrm{int}(R_k)$ is now open), and this gives us the desired result. Here $\mathrm{int}(T)$ denotes the interior of T.

Second, to prove (8), it suffices to show that

$$\int_V \nabla_{\boldsymbol{\epsilon}} \bullet \boldsymbol{U}(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon} = \int_{\partial V} \boldsymbol{U}(\boldsymbol{\epsilon}) \bullet d\boldsymbol{\Sigma}$$

where $\boldsymbol{U}(\boldsymbol{\epsilon}) = q(\boldsymbol{\epsilon}) h_k(\boldsymbol{\epsilon}, \theta) \boldsymbol{V}(\boldsymbol{\epsilon}, \theta)$ and $V = f_\theta^{-1}(R_k)$. To prove this equality, we apply the divergence theorem:

Theorem 7 (Divergence theorem)**.** *Let V be a compact subset of $\mathbb{R}^n$ that has a piecewise smooth boundary ∂V. If $\boldsymbol{F}$ is a differentiable vector field defined on a neighborhood of V, then*

$$\int_V (\nabla \cdot \boldsymbol{F})\, dV = \int_{\partial V} \boldsymbol{F} \cdot d\boldsymbol{\Sigma}$$

where $d\boldsymbol{\Sigma}$ is the outward pointing normal vector of the boundary ∂V.

In our case, the region $V = f_\theta^{-1}(R_k)$ may not be compact, so we cannot directly apply Theorem 7 to $\boldsymbol{U}$. To circumvent the non-compactness issue, we assume that $q(\boldsymbol{\epsilon})$ is in $\mathcal{S}(\mathbb{R}^n)$, the Schwartz space on $\mathbb{R}^n$. That is, assume that every partial derivative of $q(\boldsymbol{\epsilon})$ of any order decays faster than any polynomial. This assumption is reasonable in that the probability density of many important probability distributions (e.g., the normal distribution) is in $\mathcal{S}(\mathbb{R}^n)$. Since $q \in \mathcal{S}(\mathbb{R}^n)$, there exists a sequence of test functions $\{\phi_j\}_{j\in\mathbb{N}}$ such that each ϕ_j has compact support and $\{\phi_j\}_{j\in\mathbb{N}}$ converges to q in $\mathcal{S}(\mathbb{R}^n)$, which is a well-known result in functional analysis. Since each ϕ_j has compact support, so does $\boldsymbol{U}^j(\boldsymbol{\epsilon}) \triangleq \phi_j(\boldsymbol{\epsilon}) h_k(\boldsymbol{\epsilon}, \theta) \boldsymbol{V}(\boldsymbol{\epsilon}, \theta)$. By applying Theorem 7 to $\boldsymbol{U}^j$, we have

$$\int_V \nabla_{\boldsymbol{\epsilon}} \bullet \boldsymbol{U}^j(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon} = \int_{\partial V} \boldsymbol{U}^j(\boldsymbol{\epsilon}) \bullet d\boldsymbol{\Sigma}.$$

Because $\{\phi_j\}_{j\in\mathbb{N}}$ converges to q in $\mathcal{S}(\mathbb{R}^n)$, taking the limit $j \to \infty$ on the both sides of the equation gives us the desired result.

B Proof of Theorem 3

Theorem 3 is a direct consequence of the following theorem called "area formula":

Theorem 8 (Area formula)**.** *Suppose that $g : \mathbb{R}^{n-1} \to \mathbb{R}^n$ is injective and Lipschitz. If $A \subset \mathbb{R}^{n-1}$ is measurable and $\boldsymbol{H} : \mathbb{R}^n \to \mathbb{R}^n$ is measurable, then*

$$\int_{g(A)} \boldsymbol{H}(\boldsymbol{\epsilon}) \cdot d\boldsymbol{\Sigma} = \int_A \Big(\boldsymbol{H}(g(\boldsymbol{\zeta})) \cdot \boldsymbol{n}(\zeta)\Big) |Jg(\boldsymbol{\zeta})|\, d\boldsymbol{\zeta}$$

where $Jg(\boldsymbol{\zeta}) = \det\Big[\frac{\partial g(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}_1}\Big|\frac{\partial g(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}_2}\Big|\cdots\Big|\frac{\partial g(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}_{n-1}}\Big|\boldsymbol{n}(\boldsymbol{\zeta})\Big]$, and $\boldsymbol{n}(\boldsymbol{\zeta})$ is the unit normal vector of the hypersurface $g(A)$ at $g(\boldsymbol{\zeta})$ such that it has the same direction as $d\boldsymbol{\Sigma}$.

A more general version of Theorem 8 can be found in [2]. In our case, the hypersurface $g(A)$ for the surface integral on the LHS is given by $\{\boldsymbol{\epsilon} \mid \boldsymbol{a} \cdot \boldsymbol{\epsilon} = c\}$, so we use $A = \mathbb{R}^{n-1}$ and $g(\boldsymbol{\zeta}) = \big(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{j-1}, \frac{1}{\boldsymbol{a}_j}(c - \boldsymbol{a}_{-j} \cdot \boldsymbol{\zeta}), \boldsymbol{\zeta}_j, \ldots, \boldsymbol{\zeta}_{n-1}\big)^\intercal$ and apply Theorem 8 with $\boldsymbol{H}(\boldsymbol{\epsilon}) = q(\boldsymbol{\epsilon}) \boldsymbol{F}(\boldsymbol{\epsilon})$. In this settings, $\boldsymbol{n}(\boldsymbol{\zeta})$ and $|Jg(\boldsymbol{\zeta})|$ are calculated as

$$\boldsymbol{n}(\boldsymbol{\zeta}) = \mathrm{sgn}(-\boldsymbol{a}_j) \frac{|\boldsymbol{a}_j|}{\|\boldsymbol{a}\|_2} \Big(\frac{\boldsymbol{a}_1}{\boldsymbol{a}_j}, \ldots, \frac{\boldsymbol{a}_{j-1}}{\boldsymbol{a}_j}, 1, \frac{\boldsymbol{a}_{j+1}}{\boldsymbol{a}_j}, \ldots, \frac{\boldsymbol{a}_n}{\boldsymbol{a}_j}\Big)^\intercal \quad \text{and} \quad |Jg(\boldsymbol{\zeta})| = \frac{\|\boldsymbol{a}\|_2}{|\boldsymbol{a}_j|},$$

and this gives us the desired result.