A OBJECTIVE FOR DENSITY ESTIMATION

When performing density estimation for a random variable X, we only have access to samples from the unknown target distribution $X \sim p_\star$ (i.e., the unknown data distribution) but we do not have access to $p_\star$ directly (Papamakarios et al., 2017). Using Equation 1, we can use a normalizing flow to transform a complex parametric model $p_{X|\theta}$ of the target distribution into a simpler distribution p_Y (i.e., a uniform or a Normal distribution), which can be easily evaluated. In this case, we will learn the parameters θ of the model by minimizing $\mathbb{KL}(p_\star \| p_{X|\theta})$:

$$\theta^* = \min_\theta \ \mathbb{KL}(p_\star \| p_{X|\theta}) \tag{13a}$$

$$= \min_\theta \ \mathbb{E}_{p_\star(x)} \left[\log \frac{p_\star(x)}{p_{X|\theta}(x)} \right] \tag{13b}$$

$$= \min_\theta \ \underbrace{\mathbb{E}_{p_\star(x)}[\log p_\star(x)]}_{\text{=constant}} - \mathbb{E}_{p_\star(x)}[\log p_{X|\theta}(x)] \tag{13c}$$

$$= \max_\theta \ \mathbb{E}_{p_\star(x)} \left[\log p_Y(f_\theta(x)) + \log \left| \det \mathbf{J}_{f_\theta(x)} \right| \right] \ . \tag{13d}$$

where $p_{X|\theta}(x) = p_Y(y) \left| \det \mathbf{J}_{f_\theta(x)} \right|$ and $y = f_\theta(t)$. Notice that minimizing the KL is equivalent of doing maximum likelihood estimation (MLE).

B OBJECTIVE FOR DENSITY MATCHING

We can learn how to sample from a complex target distribution $p_\star$ (or, more generally, an energy function) for which we have access to its analytical form but we do not have an available sampling procedure. Using Equation 1, we can use a normalizing flow to transform samples from a simple distribution p_X, which we can easily evaluate and sample from, to a complex one (the target). In this case, we estimate θ by minimizing $\mathbb{KL}(p_{Y|\theta} \| p_\star)$:

$$\theta^* = \min_\theta \ \mathbb{KL}(p_{Y|\theta} \| p_\star) \tag{14a}$$

$$= \min_\theta \ \mathbb{E}_{p_{Y|\theta}(y)} \left[\log \frac{p_{Y|\theta}(y)}{p_\star(y)} \right] \tag{14b}$$

$$= \min_\theta \ \mathbb{E}_{p_{Y|\theta}(y)} [\log p_{Y|\theta}(y) - \log p_\star(y)] \tag{14c}$$

$$= \min_\theta \ \mathbb{E}_{p_X(x)} [\log p_X(x) - \log \left| \det \mathbf{J}_{f_\theta(x)} \right| - \log p_\star(f_\theta(x))] \ . \tag{14d}$$

where $p_{Y|\theta}(y) = p_X(x) \left| \det \mathbf{J}_{f_\theta(x)} \right|^{-1}$ and $y = f_\theta(x)$. Notice that in general, with normalizing flows, it is possible to learn a flexible distribution from which we can sample and evaluate the density of its samples. These two proprieties are particularly useful in the context of variational inference (Rezende and Mohamed, 2015).

C WEIGHT INITIALIZATION AND NORMALIZATION

Since the weight matrix W has some strictly positive and some zero entries, we need to take care of a proper initialization. Indeed, it is well known that careful parameter initialization benefits not only training but also the generalization of neural networks (Glorot and Bengio, 2010). For instance, Xavier initialization is commonly used and it takes into account the size of the input and output spaces in the affine transformations. However, since we have some zero entries, we cannot benefit from it. We choose instead to initialize all blocks with a simple distribution and to apply weight normalization (Salimans and Kingma, 2016) to better regulate the effect of such initialization. Weight normalization decomposes each row $w \in \mathbb{R}^{b \cdot d}$ of W in terms of the new parameters using $w = \exp(s) \cdot v / \|v\|$ where v has the same dimensionality of w and s is a scalar. We initialize v with a simple Normal distribution of zero mean and unit variance and $s = \log(u)$ with $u \sim \mathcal{U}(0, 1)$. Such reparametrization disentangles the direction and magnitude of w and it is known to improve and speed up optimization.

D LOGARITHMIC MATRIX MULTIPLICATION

For two arbitrary matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, their *matrix product* $C = AB \in \mathbb{R}^{m \times p}$ is defined such that

$$C_{ij} = \sum_{k=1}^{n} A_{ik} \cdot B_{kj} \ . \tag{15a}$$

Notice the latter holds only for the real semiring. In general, we can define *matrix multiplication* in any semiring as

$$C_{ij} = \bigoplus_{k=1}^{n} A_{ik} \otimes B_{kj} \ . \tag{15b}$$

In the logarithmic-semiring, the addition is defined as $a \oplus b = \log\big(e^a + e^b\big)$ and the product is defined as $a \otimes b = a + b$. Thus, the logarithmic-matrix multiplication $C = A \star B$ operation is

$$C_{ij} = \log \sum_{k=1}^{n} \exp\left(A_{ik} + B_{kj}\right) \ , \tag{15c}$$

which can be implemented with a stable *log-sum-exp* that is

$$\text{log-sum-exp}(x_1, \ldots, x_n) = \log\left(\sum_{i=1}^{n} \exp x_i\right) \tag{15d}$$

$$= x^* + \log\left(\sum_{i=1}^{n} \exp(x_i - x^*)\right) \quad \text{where} \quad x^* = \max\{x_1, \cdots, x_n\} \ . \tag{15e}$$