Supplementary Material:

Variational Inference with Non-invertible Flow

A The analysis of the regularization term under 1D linear transformation

Consider a 1D random variable z following a normal distribution $q(z) = \mathcal{N}(z|0,1)$ and apply a linear function $x = f(z) = wz$. According to the change of variable, the PDF of the random variable x is

$$\log q(x) = \log \phi(\frac{x}{w}) + \log \left| \frac{\mathrm{d}f^{-1}(x)}{\mathrm{d}x} \right|^{-1} = \log \phi(\frac{x}{w}) - \log |w|, \tag{16}$$

where $\phi(\cdot)$ is the PDF of the standard normal distribution. From the above equation, we denote the regularization term of NF to be $\gamma_{\text{NF}}(w) = \log |w|$.

With the same initial distribution $q(z)$ and the transformation function $f(z)$, we apply NFW for VI by assuming $q(x|z) = \mathcal{N}(x|f(z,\alpha)$ and $\tilde{q}(z|x) = \mathcal{N}\left(z|\tilde{f}(x,\beta\right)$. Plugging them into (11), the regularization term becomes

$$\gamma_{\text{NFW}} = \left\langle \log \frac{\tilde{q}(z|x)}{q(x|z)} \right\rangle_{q(x,z)} = \left\langle \log \frac{\mathcal{N}\left(z|\tilde{f}(x),\beta\right)}{\mathcal{N}(x|f(z),\alpha)} \right\rangle_{q(x,z)}. \tag{17}$$

Assume the reverse transformation is also a linear function $\tilde{f}(x) = \tilde{w}x$ and plug it into the above equation. The regularization term of NFW can be derived as

$$\begin{aligned}
\gamma_{\text{NFW}} &= \left\langle \log \frac{\phi(\frac{\tilde{f}(f(z)+\alpha^{\frac{1}{2}}\epsilon)-z}{\beta^{\frac{1}{2}}})\beta^{-\frac{1}{2}}}{\phi(\frac{f(z)+\alpha^{\frac{1}{2}}\epsilon-f(z)}{\alpha^{\frac{1}{2}}})\alpha^{-\frac{1}{2}}} \right\rangle_{q(z)q(\epsilon)} \\
&= \left\langle \log \phi(\frac{\tilde{f}(f(z)+\alpha^{\frac{1}{2}}\epsilon)-z}{\beta^{\frac{1}{2}}}) \right\rangle_{q(z)q(\epsilon)} - \frac{1}{2}\log\beta - \langle \log\phi(\epsilon)\rangle_{q(\epsilon)} + \frac{1}{2}\log\alpha \\
&= -\frac{1}{2\beta}\left\langle \left(\tilde{w}(wz+\alpha^{\frac{1}{2}}\epsilon)-z\right)^2 \right\rangle_{q(z)q(\epsilon)} - \frac{1}{2}\log\beta + \frac{1}{2} + \frac{1}{2}\log\alpha \\
&= -\frac{1}{2\beta}\left(1 - 2\tilde{w}w + \tilde{w}^2w^2 + \tilde{w}^2\alpha\right) - \frac{1}{2}\log\beta + \frac{1}{2} + \frac{1}{2}\log\alpha
\end{aligned}$$

By taking $\frac{\mathrm{d}\gamma_{\text{NFW}}}{\mathrm{d}\tilde{w}} = 0$, the optimal value of $\tilde{w}$ can be derived as

$$\tilde{w}^* = \frac{w}{w^2+\alpha}.$$

By plugging in the optimal $\tilde{w}$, we get the regularization term as a function of w,

$$\gamma_{\text{NFW}}(w) = -\frac{1}{2\beta}\frac{\alpha}{w^2+\alpha} - \frac{1}{2}\log\beta + \frac{1}{2} + \frac{1}{2}\log\alpha.$$

Fig. 1b in the main text shows the comparison of $\gamma_{\text{NF}}(w)$ and $\gamma_{\text{NFW}}(w)$ with three different choices of α and β. These two regularization terms are both concave with the same minimal. The smaller α and β is, the steeper the regularization term is in NFW.

B Local linear approximation

Another way to understand to the connection between the regularization term in NFW and in NF is to make a local linear approximation of the reverse transformation $\tilde{f}(\mathbf{x})$. In (15), the transformed value $f(\mathbf{z})$ is corrupted with a small Gaussian noise $\mathbf{x} = f(\mathbf{z}) + \alpha^{\frac{1}{2}}\epsilon$. As the noise corruption will be small, we can do a Taylor expansion on $\tilde{f}$ around $f(\mathbf{z})$ and approximate it with a local linear function,

$$\tilde{f}(\mathbf{x}) \approx \tilde{f}(\mathbf{x}_0) + \tilde{\mathbf{J}}_{\mathbf{x}_0}^{\top}(\mathbf{x} - \mathbf{x}_0),$$

where $\mathbf{x}_0 = f(\mathbf{z})$ and $\tilde{\mathbf{J}}_{\mathbf{x}_0}$ is the Jacobian of $\tilde{f}$ at $\mathbf{x}_0$.

Plugging the above approximation into (11), we derive an approximated regularization term,

$$\begin{aligned}\gamma_{\text{NFW}} &\approx \left\langle \log \frac{\mathcal{N}\left(\mathbf{z}|\tilde{f}(\mathbf{x}_0) + \tilde{\mathbf{J}}_{\mathbf{x}_0}^{\top}(\mathbf{x} - \mathbf{x}_0), \beta\right)}{\mathcal{N}\left(\mathbf{x}|f(\mathbf{z}), \alpha\right)} \right\rangle_{q(\mathbf{x},\mathbf{z})} \\ &= \left\langle -\frac{1}{2\beta}\left(|\tilde{\mathbf{z}} - \mathbf{z}|^2 + \alpha \text{tr}\left(\tilde{\mathbf{J}}_{\mathbf{x}_0}^{\top}\tilde{\mathbf{J}}_{\mathbf{x}_0}\right)\right)\right\rangle_{q(\mathbf{z})} - \frac{D_z}{2}\log(2\pi\beta) + \frac{D_x}{2} + \frac{D_x}{2}\log(2\pi\alpha)\end{aligned}$$

where $\tilde{\mathbf{z}} = \tilde{f}(f(\mathbf{z}) + \alpha^{\frac{1}{2}}\epsilon)$. Compare the above term with the regularization term of NF in (14), which is

$$\gamma_{NF} = -\left\langle \log \left| \frac{\mathrm{d}f^{-1}(\mathbf{x})}{\mathrm{d}\mathbf{x}} \right| \right\rangle_{q(\mathbf{z})}.$$

In the ideal situation, if the reverse transformation $\tilde{f}$ recovers the initial variable $\mathbf{z}$ with sufficient accuracy, we can ignore the term $|\tilde{\mathbf{z}} - \mathbf{z}|^2$. Then, γ_{NFW} is dominated by the trace term. In this case, if f is invertible, $\tilde{f}$ will be close to f^{-1}. The term $-\frac{\alpha}{2\beta}\text{tr}\left(\tilde{\mathbf{J}}_{\mathbf{x}_0}^{\top}\tilde{\mathbf{J}}_{\mathbf{x}_0}\right)$ encourages the Eigen values of the Jacobian of f^{-1} to be smaller, which has a similar effect as $-\log\left|\frac{\mathrm{d}f^{-1}(\mathbf{x})}{\mathrm{d}\mathbf{x}}\right|$ due to the connection between determinant and trace.