1 Introduction

Problem Setup

Suppose $X \in \mathbb{R}^n$ and $Z \in \mathbb{R}^m$ with $X = f(Z) + \varepsilon$, where $f : \mathbb{R}^m \to \mathbb{R}^n$ and $\varepsilon \sim N\left(0, \sigma^2 I\right)$. Assuming Z is normally distributed with independent marginals, this is equivalent to the following latent variable model (a special case of the well-known *nonlinear ICA* model):

$$\begin{aligned} Z &\sim N(0, I) \\ X \mid Z &\sim N\left(f(Z), \sigma^2 I\right). \end{aligned}$$

Let $\varphi(u; \mu, \Sigma)$ denote the density of a $N(\mu, \Sigma)$ random variable and $p(x, z)$ denote the joint density under the model. It is easy to see that

$$\begin{aligned} p(x, z) &= p(x \mid z)\, p(z) = \varphi\left(x; f(z), \sigma^2 I\right) \varphi(z; 0, I) \\ p(x) &= \int \varphi\left(x; f(z), \sigma^2 I\right) \varphi(z; 0, I) dz \end{aligned}$$

Marginal Density and Inference

Now, suppose we let g_θ denote a family of deep neural network distributions parametrized by θ. To approximate the marginal density $p(x)$, we replace f with g_θ and try to find the choice of θ that maximizes the observed data likelihood. Given k observations $x^{(i)} \overset{\text{iid}}{\sim} p(x)$, we wish to solve the following maximum likelihood problem:

$$\max_{\theta, \sigma} \underbrace{\sum_{i=1}^{k} \int \varphi\left(x^{(i)}; g_\theta(z), \sigma^2 I\right) \varphi(z; 0, I) dz}_{:= \ell(\theta, \sigma)}$$

2 Direct MLE Method

Directly solve the MLE problem by computing gradients of $\ell(\theta, \sigma)$ w.r.t θ and σ. This is, in general, intractable for arbitrary nonlinear ICA models but worst-case thinking does not apply to our special cases.

Gradient w.r.t θ

$$\begin{aligned} \nabla_\theta \ell(\theta, \sigma) &= \sum_{i=1}^{k} \int \nabla_\theta \varphi\left(x^{(i)}; g_\theta(z), \sigma^2 I\right) \varphi(z; 0, I) dz \\ &= \sum_{i=1}^{k} \int \left(2\pi\sigma^2\right)^{-n/2} \nabla_\theta \exp\left(-\frac{\|x^{(i)} - g_\theta(z)\|_2^2}{2\sigma^2}\right) \varphi(z; 0, I) dz \\ &= \sum_{i=1}^{k} \int \nabla_\theta \left(-\frac{\|x^{(i)} - g_\theta(z)\|_2^2}{2\sigma^2}\right) \varphi\left(x^{(i)}; g_\theta(z), \sigma^2 I\right) \varphi(z; 0, I) dz \\ &= \sum_{i=1}^{k} \int \left[\frac{1}{\sigma^2} \cdot \left(x^{(i)} - g_\theta(z)\right)^T \nabla_\theta g_\theta(z)\right] \varphi\left(x^{(i)}; g_\theta(z), \sigma^2 I\right) \varphi(z; 0, I) dz \end{aligned}$$

Gradient w.r.t σ^2

$$\begin{aligned}
\nabla_{\sigma^2}\ell(\theta,\sigma) &= \sum_{i=1}^{k}\int \nabla_{\sigma^2}\varphi\left(x^{(i)}; g_\theta(z), \sigma^2 I\right)\varphi(z;0,I)dz \\
&= \sum_{i=1}^{k}\int \nabla_{\sigma^2}\left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{\|x^{(i)}-g_\theta(z)\|_2^2}{2\sigma^2}\right)\varphi(z;0,I)dz \\
&= \sum_{i=1}^{k}\int \left[\frac{1}{2\sigma^2}\cdot\left(-n+\frac{\|x^{(i)}-g_\theta(z)\|_2^2}{\sigma^2}\right)\right]\varphi\left(x^{(i)}; g_\theta(z), \sigma^2 I\right)\varphi(z;0,I)dz
\end{aligned}$$

From the two results above, we can iteratively update θ and σ^2 via gradient descent. The integrals can be approximated via numerical integration.