1 Introduction

1.1 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 $p_{\theta,\sigma^2}(x, z)$ denote the joint density under the model. It is easy to see that

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

1.2 Objective Function

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 $x^{(i)} \overset{i.i.d}{\sim} p(x)$, we wish to solve the following maximum likelihood problem:

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

2 Direct MLE Method

In this section, we directly solve the MLE problem by computing gradients of $\ell(\theta, \sigma^2)$ w.r.t θ and σ^2. 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

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

Gradient w.r.t σ^2

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

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

Algorithm 1 Direct MLE via Gradient Descent

- Initialise $\theta^{(0)}$ and $\sigma^{2(0)}$ and set $t=0$
- Repeat until convergence
  - ▷ Compute the gradient $\nabla_\theta\ell\left(\theta^{(t)},\sigma^{2(t)}\right)$ and update the parameters

$$\theta^{(t+1)}=\theta^{(t)}-\eta_1\nabla_\theta\ell\left(\theta^{(t)},\sigma^{2(t)}\right)$$

  - ▷ Compute the gradient $\nabla_{\sigma^2}\ell\left(\theta^{(t+1)},\sigma^{2(t)}\right)$ and update the parameters

$$\sigma^{2(t+1)}=\sigma^{2(t)}-\eta_2\nabla_{\sigma^2}\ell\left(\theta^{(t+1)},\sigma^{2(t)}\right)$$

  - ▷ Set $t\leftarrow t+1$

3 Variational Method

In this section, we consider variational inference and VAEs. We use the ELBO to obtain a lower bound on the likelihood $\ell(\theta,\sigma)$ and optimize the ELBO using SGD. The marginal likelihoods of individual datapoints can each be rewritten as

$$\log p_\theta\left(x^{(i)}\right)=D_{KL}\left(q_\phi\left(z\mid x^{(i)}\right)\|p_\theta\left(z\mid x^{(i)}\right)\right)+\mathcal{L}\left(\theta,\phi;x^{(i)}\right)$$

The second RHS term $\mathcal{L}\left(\theta,\phi;x^{(i)}\right)$ is called the evidence lower bound on the marginal likelihood of datapoint i, and can be written as

$$
\begin{aligned}
\log p_\theta\left(x^{(i)}\right)\geq\mathcal{L}\left(\theta,\phi;x^{(i)}\right) &= \mathbb{E}_{q_\phi(z\mid x)}\left[-\log q_\phi(z\mid x)+\log p_\theta(x,z)\right] \\
&= -D_{KL}\left(q_{\boldsymbol{\phi}}\left(\mathbf{z}\mid\mathbf{x}^{(i)}\right)\|p_{\boldsymbol{\theta}}(\mathbf{z})\right)+\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}\mid\mathbf{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}\left(\mathbf{x}^{(i)}\mid\mathbf{z}\right)\right]
\end{aligned}
$$

We want to differentiate and optimize the lower bound $\mathcal{L}\left(\boldsymbol{\theta},\phi;\mathbf{x}^{(i)}\right)$ w.r.t. both the variational parameters ϕ and generative parameters θ.