

1 Introduction

1.1 Problem Setup

Suppose $X \in \mathbb{R}^n$ and $Z \in \mathbb{R}^m$ with $X = f(Z) + \varepsilon$, where $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$ and $\varepsilon \sim N(0, \sigma^2 I)$. 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):

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

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

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

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

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

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

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

Gradient w.r.t σ^2

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

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 $\nabla_{\theta} \ell(\theta^{(t)}, \sigma^{2(t)})$ and update the parameters

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

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

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

- ▷ 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 $\ell(\theta, \sigma^2)$ and optimize the ELBO using SGD. The marginal likelihoods of individual datapoints can each be rewritten as

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

The term $\mathcal{L}(\theta, \phi; x^{(i)})$ is called the evidence lower bound on the marginal likelihood of datapoint i and can be written as ¹

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

We want to differentiate and optimize the lower bound $\mathcal{L}(\theta, \phi; x^{(i)})$ w.r.t. both the variational parameters ϕ and generative parameters θ . The KL-divergence $D_{KL}(q_\phi(z^{(i)} | x^{(i)}) || p_\theta(z^{(i)}))$ can be integrated analytically, such that only the reconstruction error $\mathbb{E}_{q_\phi(z^{(i)} | x^{(i)})} [\log p_\theta(x^{(i)} | z^{(i)})]$ requires estimation by sampling. The *stochastic gradient variational bayes* (SGVB) estimator

$$\tilde{\mathcal{L}}(\theta, \phi; x^{(i)}) = -D_{KL}(q_\phi(z^{(i)} | x^{(i)}) || p_\theta(z^{(i)})) + \frac{1}{L} \sum_{l=1}^L \log p_\theta(x^{(i)} | z^{(i,l)})$$

KL-divergence

When both the prior $p_\theta(z) = \mathcal{N}(0, I)$ and the posterior approximation $q_\phi(z^{(i)} | x^{(i)})$ are Gaussian, the KL term that can be integrated analytically. Let J be the dimensionality of z . Let μ and σ denote the variational mean and std evaluated at datapoint i , and let μ_j and σ_j denote the j -th element of these vectors.

$$\begin{aligned} -D_{KL}(q_\phi(z) || p_\theta(z)) &= \int q_\phi(z) (\log p_\theta(z) - \log q_\phi(z)) dz \\ &= \frac{1}{2} \sum_{j=1}^J \left(1 + \log((\sigma_j)^2) - (\mu_j)^2 - (\sigma_j)^2 \right) \end{aligned}$$

Probabilistic encoders and decoders

In variational auto-encoders, neural networks are used as probabilistic encoders and decoders. For both the encoder and decoder, we use a MLP with Gaussian outputs. Let the decoder be a multivariate Gaussian with a diagonal covariance structure

$$\begin{aligned} \log p(x | z) &= \log \mathcal{N}(x; m, s^2 I) \\ \text{where } h &= h(z) \\ m &= W_1 h + b_1 \\ \log s^2 &= W_2 h + b_2 \end{aligned}$$

where $\{W_1, W_2, b_1, b_2\}$ are the weights and biases of the MLP (as part of θ) with $m \in R^n$ and $s^2 \in R$. The reconstruction error can be expanded as

$$\log p(x | z) = -\frac{n}{2} \cdot \log(2\pi s^2) - \frac{\|x - m\|_2^2}{2s^2}$$

¹An equivalent concept to ELBO is the variational free energy. The variational free energy in a latent variable model $p_\theta(x, z)$ is defined as

$$\mathcal{L}(\theta, q) = \mathbb{E}_{z \sim q} [-\log q(z) + \log p_\theta(x, z)],$$

where q is any probability density/mass function over the latent variables z . The first term is the Shannon entropy $H(q) = -\mathbb{E}_{z \sim q} \log q(z)$ of the variational distribution $q(z)$ and does not depend on θ . The second term is sometimes referred to as the energy.