MLE/VAE for nonlinear ICA

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 $\varepsilon \sim N\left(0, \sigma^2 I_n\right)$. 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):

$$\begin{aligned} Z &\sim N(0, I) \\ X \mid Z &\sim N\left(f(Z), \sigma^2 I\right). \end{aligned} \tag{1}$$

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

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

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:

$$\max_{\theta,\sigma^2} \underbrace{\sum_{i=1}^{k} \log \int \varphi\left(x^{(i)}; g_\theta(z), \sigma^2 I\right) \varphi(z; 0, I) dz}_{:=\ell(\theta,\sigma^2)} \tag{3}$$

2 Previous Literature

It was widely believed in previous literature that directly optimizing the marginal likelihood in latent variable models is hard, without making common simplifying assumptions about the marginal or posterior probabilities. In particular, we are interested in the intractability settings as described by [3], where

- The integral of the marginal likelihood $p_\theta(x) = \int p_{\boldsymbol{\theta}}(z) p_{\boldsymbol{\theta}}(x \mid z)\, dz$ is intractable (so we cannot evaluate or differentiate the marginal likelihood)
- The true posterior density $p_{\boldsymbol{\theta}}(z \mid x) = p_{\boldsymbol{\theta}}(x \mid z) p_{\boldsymbol{\theta}}(z) / p_{\boldsymbol{\theta}}(x)$ is intractable (so the EM algorithm cannot be used)

- The required integrals for any reasonable mean-field VB algorithm are also intractable.

The intractability condition is very common, and can be easily established for moderately complicated likelihood functions $p_{\boldsymbol{\theta}}(x \mid z)$ (e.g. a neural network with a nonlinear hidden layer). Under the intractabilities, the direct MLE method is hard:

- [4] established that the log probability of generating a particular example d, from a model with parameters θ is

  $$\log p(d \mid \theta) = \log \left[\sum_{\alpha} p(\alpha \mid \theta) p(d \mid \alpha, \theta) \right]$$

  where the α are explanations. The posterior probability of an explanation given d and θ is related to its energy by the equilibrium or Boltzmann distribution, which at a temperature of 1 gives

  $$P_{\alpha}(\theta, d) = \frac{p(\alpha \mid \theta) p(d \mid \alpha, \theta)}{\sum_{\alpha'} p\left(\alpha' \mid \theta\right) p\left(d \mid \alpha', \theta\right)} = \frac{e^{-E_{\alpha}}}{\sum_{\alpha'} e^{-E_{\alpha'}}}$$

  However, it was claimed in the paper that the posterior distribution is computationally intractable. It has exponentially many terms and cannot be factored into a product of simpler distributions.

- [3] established the naïve Monte Carlo gradient estimator

  $$\begin{aligned} \nabla_{\phi} \mathbb{E}_{q_{\phi}(z)}[f(z)] &= \mathbb{E}_{q_{\phi}(z)} \left[f(z) \nabla_{q_{\phi}(z)} \log q_{\phi}(z) \right] \\ &\simeq \frac{1}{L} \sum_{l=1}^{L} f(z) \nabla_{q_{\phi}\left(z^{(l)}\right)} \log q_{\phi} \left(z^{(l)} \right) \end{aligned}$$

  where $z^{(l)} \sim q_{\phi}\left(z \mid x^{(i)}\right)$ exhibits very high variance and is impractical for the purposes of learning [5].

In addition to the hardness proposed in previous literature, we have demonstrated in our own experiments that directly optimizing likelihood 1 is computationally challenging due to the following two facts:

- The density $\varphi\left(x^{(i)}; g_{\theta}(z), \sigma^2 I\right)$ becomes vanishingly small when the dimension of the observed space becomes large, incuring numerical underflow when performing evaluation of the likelihood and computing optimization

- Approximating the integral with numerical integration is challenging – which requires large number of Monte-Carlo samples such that the integral can be evaluted to the precision desired. In contrast, [3] has found that the number of Monte-Carlo samples from the latent space can be set to one, as long as the minibatch size M is large enough – this avoids the computational burden of large number of Monte-Carlo samples.

3 MLE with Gradient Descent

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

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

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 $t \leftarrow t + 1$
```

4 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

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

The 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 [1]

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

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

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

KL-divergence

When both the prior $p_\theta(z) = \mathcal{N}(0, I)$ and the posterior approximation $q_\phi\left(z^{(i)} \mid x^{(i)}\right)$ 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.

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

Reconstruction Error

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

[1] An 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

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

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.

multivariate Gaussian with a diagonal covariance structure

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

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

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

References

[1] Khemakhem, Ilyes, et al. “Variational Autoencoders and Nonlinear ICA: A Unifying Framework.” ArXiv.org, 21 Dec. 2020, arxiv.org/abs/1907.04809.

[2] Kingma, Diederik P., and Max Welling. “An Introduction to Variational Autoencoders.” ArXiv.org, 11 Dec. 2019, https://arxiv.org/abs/1906.02691.

[3] Kingma, Diederik P, and Max Welling. “Auto-Encoding Variational Bayes.” ArXiv.org, 1 May 2014, https://arxiv.org/abs/1312.6114.

[4] RS;, Dayan P;Hinton GE;Neal RM;Zemel. “The Helmholtz Machine.” Neural Computation, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/7584891/.

[5] Paisley, John, et al. “Variational Bayesian Inference with Stochastic Search.” ArXiv.org, 27 June 2012, arxiv.org/abs/1206.6430.