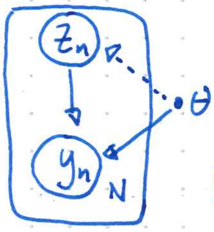
LATENT VARIABLE MODELS REVISITED



$z_n \sim p_\theta(z_n)$ FOR EXAMPLE: $z_n \sim N(0, \Sigma_z)$, $z_n \in \mathbb{R}^{D'}$

$y_n \sim p_\theta(y_n | z_n)$ $\qquad y_n \sim N(f_\theta(z_n), \Sigma_y)$, $y_n \in \mathbb{R}^{D}$

$\quad\equiv$ equiv.

ALTERNATIVELY: $y_n = f_\theta(z_n) + \varepsilon$, $\varepsilon \sim N(0, \Sigma_y)$

GOALS: 1. LEARN θ_{MLE}. I.E. LEARN TO MODEL COMPLEX DISTR. OF DATA $\{y_n\}$ BY TRANSFORMING SIMPLE DISTRIBUTION $p_\theta(z)$

2. LEARN THE POSTERIORS: $p(z_n | y_n)$

Since we don't observe z_n, we can only maximize observed data likelihood

$$\theta_{MLE} = \arg\max_\theta \log \prod_n p_\theta(y_n) = \arg\max_\theta \sum_n \log \int p_\theta(y_n | z_n) p_\theta(z_n) dz_n$$

$$= \arg\max_\theta \sum_n \log \mathbb{E}_{p_\theta(z_n)} [p_\theta(y_n | z_n)]$$

The problem: $\nabla_\theta \sum_n \log \mathbb{E}_{p_\theta(z_n)} [p_\theta(y_n | z_n)] = \sum_n \nabla_\theta \log \mathbb{E}_{p_\theta(z)} [p_\theta(y_n | z_n)]$

$$= \sum_n \frac{\nabla_\theta \mathbb{E}_{p_\theta(z_n)} [p_\theta(y_n | z_n)] \leftarrow \text{hard grad.}}{\mathbb{E}_{p_\theta(z_n)} [p_\theta(y_n | z_n)] \leftarrow \text{high var.}}$$

I. IMPORTANCE SAMPLING

$\sum_n \log \int p_\theta(y_n | z_n) p_\theta(z_n) dz_n$

$$\sum_n \log \mathbb{E}_{p_\theta(z_n)} [p_\theta(y_n | z_n)] = \sum_n \log \int \underbrace{\frac{p_\theta(z_n)}{q(z_n)}}_{\text{importance weight}} p_\theta(y_n | z_n) \cdot \underbrace{q(z_n)}_{\text{importance distribution}} dz_n$$

$$= \sum_n \log \mathbb{E}_{q(z_n)} \left[\frac{p_\theta(y_n | z_n) p_\theta(z_n)}{q(z_n)} \right]$$

$$\geq \underbrace{\sum_n \mathbb{E}_{q(z_n)} \left[\log \frac{p_\theta(y_n, z_n)}{q(z_n)} \right]}_{\text{ELBO}(\theta, q)}$$

$$\max_\theta \log \prod_n p_\theta(y_n) \geq \max_{\theta, q} \text{ELBO}(\theta, q)$$

II. OPTIMIZATION

IN EM:

Step M: $\max_\theta \text{ELBO}(\theta, q^*) \Rightarrow \nabla_\theta \mathbb{E}_{q^*(z_n)} \left[\log \frac{p_\theta(y_n, z_n)}{q(z_n)} \right] = \mathbb{E}_{q^*(z_n)} \Big[\underbrace{\nabla_\theta \log \frac{p_\theta(y_n, z_n)}{q(z_n)}}_{\text{AUTOGRAD}} \Big]$

Step E: $\max_q \text{ELBO}(\theta^*, q) \Rightarrow q^* = \arg\max_q \text{ELBO}(\theta^*, q) \Rightarrow p_{\theta^*}(z_n | y_n) = q^*(z_n)$

The problem: if $y_n = \text{nn.forward}(z_n) + \varepsilon$, $p(z_n | y_n)$ is intractable to compute analytically (unlike in the case of Gaussian Mixtures)

A. VARIATIONAL INFERENCE

Set $q(z_n) = N(z_n; \mu_n, \Sigma_n)$, Σ_n diagonal.

In E-step: $\mu_n^*, \Sigma_n^* = \arg\min_{\mu_n, \Sigma_n} D_{KL}[q_{\mu_n, \Sigma_n}(z_n) \| p_{\theta^*}(z_n | y_n)]$

$$\equiv \arg\max_{\mu_n, \Sigma_n} \underbrace{\mathbb{E}_{q_{\mu_n, \Sigma_n}(z_n)} \left[\log \frac{p_{\theta^*}(z_n, y_n)}{q_{\mu_n, \Sigma_n}(z_n)} \right]}$$

The problem: if N is huge then we need $\text{ELBO}_n(\theta^*, q_{\mu_n, \Sigma_n})$ to run VI a huge number of times. Computationally intractable.

B. AMORTIZATION

<u>Idea</u>: if $y_n = 1$ and $\mu_n = 1.01$, $\sigma_n^2 = 0.5$

then if $y_m \approx y_n$ then $p(z_n | y_n) \approx p(z_m | y_m)$ and hence $\begin{matrix}\sigma_m^2 \approx \sigma_n^2 \\ \mu_m \approx \mu_n\end{matrix}$

i.e. we can predict μ_n, σ_n^2 by looking at y_n, $\exists\, g_\phi(y_n) = \mu_n, \Sigma_n$.

We learn a function $g_{\phi^*}(y_n) = \mu_{\phi^*}(y_n), \Sigma_{\phi^*}(y_n)$ s.t.

$$\phi^* = \underset{\phi}{\operatorname{argmin}} \sum_n D_{KL}[\underbrace{N(\mu_\phi(y_n), \Sigma_\phi(y_n))}_{q_\phi(z_n)} \| P_{\theta^*}(z_n | y_n)]$$

$$\equiv \underbrace{\underset{\phi}{\operatorname{argmax}} \sum_n \mathbb{E}_{q_\phi(z_n)} \left[\frac{P_{\theta^*}(z_n, y_n)}{q_\phi(z_n)}\right]}_{\text{ELBO}(\theta^*, q_\phi)}$$

<u>Gradient descent</u>: $\nabla_\phi \text{ELBO}(\theta^*, q_\phi) = \nabla_\phi \sum_n \mathbb{E}_{q_\phi(z_n)} \left[\frac{P_{\theta^*}(z_n, y_n)}{q_\phi(z_n)}\right]$

$$= \sum_n \nabla_\phi \mathbb{E}_{\varepsilon \sim N(0,I)} \left[\frac{P_{\theta^*}(\varepsilon \Sigma_\phi^{1/2}(y_n) + \mu_\phi(y_n), y_n)}{q_\phi(\varepsilon \Sigma_\phi(y_n)^{1/2} + \mu_\phi(y_n))}\right]$$

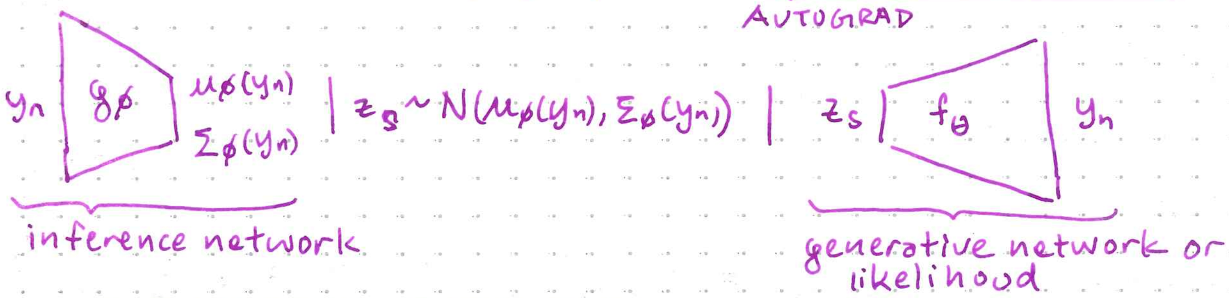
$$= \sum_n \mathbb{E}_{\varepsilon \sim N(0,I)} \Bigg[\underbrace{\nabla_\phi \frac{P_{\theta^*}(\varepsilon \Sigma_\phi(y_n)^{1/2} + \mu_\phi(y_n), y_n)}{q_\phi(\varepsilon \Sigma_\phi(y_n)^{1/2} + \mu_\phi(y_n))}}_{\text{AUTOGRAD}}\Bigg]$$

C. JOINT TRAINING:

$$\theta^*, \phi^* = \underset{\theta,\phi}{\operatorname{argmax}}\ \text{ELBO}(\theta, q_\phi)$$

$$\nabla_{\phi,\theta} \text{ELBO}(\theta, q_\phi) = \nabla_{\phi,\theta} \sum_n \mathbb{E}_{\varepsilon \sim N(0,I)} \left[\frac{P_\theta(\varepsilon \cdot \Sigma_\phi(y_n)^{1/2} + \mu_\phi(y_n), y_n)}{q_\phi(\varepsilon \Sigma_\phi(y_n)^{1/2} + \mu_\phi(y_n))}\right]$$

$$= \sum_n \mathbb{E}_{\varepsilon \sim N(0,I)} \Bigg[\underbrace{\nabla_{\phi,\theta} \frac{P_\theta(\varepsilon \cdot \Sigma_\phi(y_n)^{1/2} + \mu_\phi(y_n), y_n)}{q_\phi(\varepsilon \Sigma_\phi(y_n)^{1/2} + \mu_\phi(y_n))}}_{\text{AUTOGRAD}}\Bigg]$$



DIFFERENT PRESENTATIONS OF THE ELBO

I. JOINT PLUS ENTROPY

$$\text{ELBO}(\theta, q_\phi) = \sum_n \mathbb{E}_{q_\phi} \left[\log \frac{P_\theta(y_n, z_n)}{q_\phi(z_n)}\right] = \sum_n \mathbb{E}_{q_\phi} \log P_\theta(y_n, z_n) - \mathbb{E}_{q_\phi}[\log q_\phi]$$

$$= \sum_n \left[\mathbb{E}_{q_\phi} \log P_\theta(y_n, z_n) + \mathbb{H}[q_\phi]\right]$$

II. EXPECTED LIKELIHOOD MINUS KL

$$\text{ELBO}(\theta, q_\phi) = \sum_n \mathbb{E}_{q_\phi} \left[\log \frac{P_\theta(y_n, z_n)}{q_\phi(z_n)}\right] = \sum_n \mathbb{E}_{q_\phi} \left[\log \frac{P_\theta(y_n | z_n) P_\theta(z_n)}{q_\phi(z_n)}\right]$$

$$= \sum_n \left[\mathbb{E}_{q_\phi}[\log P_\theta(y_n | z_n)] - \mathbb{E}_{q_\phi}\left[\log \frac{q_\phi(z_n)}{P_\theta(z_n)}\right]\right]$$

$$= \sum_n \left[\mathbb{E}_{q_\phi}[\log P_\theta(y_n | z_n)] - D_{KL}[q_\phi(z_n) \| P_\theta(z_n)]\right]$$

III. OBSERVED LOG LIKELIHOOD MINUS KL

$$\text{ELBO}(\theta, q_\phi) = \sum_n \mathbb{E}_{q_\phi}\left[\log \frac{p_\theta(y_n, z_n)}{q_\phi(z_n)}\right] = \sum_n \mathbb{E}_{q_\phi}\left[\log \frac{p_\theta(y_n, z_n)\, p_\theta(y_n)}{q_\phi(z_n)\, p_\theta(y_n)}\right]$$

$$= \sum_n \mathbb{E}_{q_\phi}\left[\log \frac{p_\phi(z_n | y_n)}{q_\phi(z_n)} \cdot p_\theta(y_n)\right]$$

$$= \sum_n \left[\mathbb{E}_{q_\phi} \log p_\theta(y_n) - \mathbb{E}_{q_\phi} \log \frac{q_\phi(z_n)}{p_\phi(z_n | y_n)}\right]$$

$$= \sum_n \left[\log p_\theta(y_n) - D_{KL}\left[q_\phi(z_n) \,\|\, p_\phi(z_n | y_n)\right]\right]$$