This handout includes space for every question that requires a written response. Please feel free to use it to handwrite your solutions (legibly, please). If you choose to typeset your solutions, the `README.md` for this assignment includes instructions to regenerate this handout with your typeset LaTeX solutions.

1

To help you get started consider this: We rely on the known property that if ψ is a strictly monotonically decreasing function, then the following two problems are equivalent:

$$\max_{\theta} f(\theta) = \min_{\theta} \psi(f(\theta))$$

2 To help you get start consider that:

$$p_\theta(y \mid x) = \frac{p_\theta(x,y)}{p_\theta(x)} = \frac{\pi_y \cdot \exp\left(-\frac{1}{2\sigma^2}(x-\mu_y)^\top (x-\mu_y)\right) \cdot Z^{-1}(\sigma)}{\sum_i \pi_i \cdot \exp\left(-\frac{1}{2\sigma^2}(x-\mu_i)^\top (x-\mu_i)\right) \cdot Z^{-1}(\sigma)}$$

where $Z(\sigma)$ is the Gaussian partition function (which is a function of σ).

3a
3b
3c

4

To provide more direction consider proving (*) or (**) below: Consider the simple case of describing a joint distribution over (X_1, X_2) using the forward versus reverse factorizations. Consider the forward factorization where

$$p_f(x_1) = \mathcal{N}(x_1 \mid 0, 1)$$
$$p_f(x_2 \mid x_1) = \mathcal{N}(x_2 \mid \mu_2(x_1), \epsilon)$$

for which

$$\mu_2(x_1) = \begin{cases} 0 & \text{if } x_1 \le 0 \\ 1 & \text{otherwise} \end{cases}$$

(*) This construction makes $p_f(x_2)$ a mixture of two distinct Gaussians, which $p_r(x_2)$ cannot match, since $p_r(x_2)$ is strictly Gaussian. Any counterexample of this form, which makes $p_f(x_2)$ non-Gaussian, suffices for full-credit.

(**) Interestingly, we can also intuit about the distribution $p_f(x_1 \mid x_2)$. If one chooses a very small positive ϵ, then the corresponding $p_f(x_1 \mid x_2)$ will approach a truncated Gaussian distribution, which cannot be approximated by the Gaussian $p_r(x_1 \mid x_2)$ [1].

Optionally, we can prove (*) and a variant of (**) which states that, any $\epsilon > 0$, the distribution:

$$p_f(x_1 \mid x_2) = \frac{p_f(x_1, x_2)}{p_f x_2}$$

is a mixture of truncated Gaussians whose mixture weights depend on ϵ.

[1] This observation will be useful when we move on to variational autoencoders $p(z, x)$ (where z is a latent variable) and discuss the importance of having good variational approximations of the true posterior $p(z \mid x)$

5a
5b

6a
6b
6g