STAT215: Assignment 3

Libby Zhang

Due: March 3, 2020

Problem 1: *Variational inference.*

Standard VI minimizes $\mathrm{KL}\left(q(z) \,\|\, p(z \mid x)\right)$, the Kullback-Leibler divergence from the variational approximation $q(z)$ to the true posterior $p(z \mid x)$. In this problem we will develop some intuition for this optimization problem. For further reference, see Chapter 10 of *Pattern Recognition and Machine Learning* by Bishop.

(a) Let $\mathcal{Q} = \{q(z) : q(z) = \prod_{d=1}^{D} \mathcal{N}(z_d \mid m_d, v_d^2)\}$ denote the set of Gaussian densities on $z \in \mathbb{R}^D$ with diagonal covariance matrices. Solve for

$$q^\star = \arg\min_{\mathcal{Q}} \mathrm{KL}\left(q(z) \,\|\, \mathcal{N}(z \mid \mu, \Sigma)\right),$$

where Σ is an arbitrary covariance matrix.

We will use the closed form expression of the KL divergences between two Gaussian distributions, $\mathcal{N}_0(\mu_0, \Sigma_0)$ and $\mathcal{N}_1(\mu_1, \Sigma_1)$,

$$\mathrm{KL}\left(\mathcal{N}_0 || \mathcal{N}_1\right) = \frac{1}{2}\left(\mathrm{tr}\left(\Sigma_1^{-1}\Sigma_0\right) + (\mu_1 - \mu_0)^\mathsf{T}\Sigma_1^{-1}(\mu_1 - \mu_0) + \log|\Sigma_1| - \log|\Sigma_0| - D\right) \tag{1}$$

Substituting in our distributions $q(z) = \prod_{d=1}^{D} \mathcal{N}(z_d \mid m_d, v_d^2) = \mathcal{N}(z \,|\, m, V = \mathrm{diag}[v_1^2, \ldots, v_D^2])$ and $p(z) = \mathcal{N}(z \mid \mu, \Sigma)$, we have

$$\mathrm{KL}(q||p) = \frac{1}{2}\left(\mathrm{tr}\left(\Sigma^{-1}V\right) + (m - \mu)^\mathsf{T}\Sigma^{-1}(m - \mu) + \log|\Sigma| - \log|V| - D\right)$$

To find the minimizing $q^\star$, we take the gradient of the KL divergence with respective to each of the parameters $\phi = (m, V)$ and set to 0 to solve for the maximizing parameter.

$$\begin{aligned}
\nabla_m \mathrm{KL}(q||p) &= \frac{1}{2}\nabla_m\left((m - \mu)^\mathsf{T}\Sigma^{-1}(m - \mu)\right) \\
&= \frac{1}{2}\nabla_m\left(m^\mathsf{T}\Sigma^{-1}m - 2\mu^\mathsf{T}\Sigma^{-1}m + \mu^\mathsf{T}\Sigma^{-1}\mu\right) \\
&= \frac{1}{2}\left(2\Sigma^{-1}m - 2\Sigma^{-}1\mu\right) \\
&= \Sigma^{-1}(m - \mu) = 0 \\
\Rightarrow m^\star &= \mu
\end{aligned} \tag{2}$$

$$
\begin{aligned}
\nabla_V \mathrm{KL}(q||p) &= \frac{1}{2}\nabla_V \left(\mathrm{tr}\left(\Sigma^{-1}V\right) - \log|V|\right) \\
&= \frac{1}{2}\nabla_V \left[\sum_d \left(\Sigma^{-1}\right)_{dd} v_d^2\right] - V^{-1} \qquad (3) \\
&= \frac{1}{2}\left(\mathrm{diag}\left[\Sigma^{-1}\right] - V^{-1}\right) \\
\Rightarrow V^\star &= \mathrm{diag}\left[\Sigma^{-1}\right]^{-1} \qquad (4)
\end{aligned}
$$

where in (3), we used the matrix derivative identity $\nabla_A \log|A| = \left(A^\top\right)^{-1}$

(b) Now solve for $q^\star \in \mathscr{Q}$ that minimizes the KL in the opposite direction,

$$
q^\star = \arg\min_{\mathscr{Q}} \mathrm{KL}\left(\mathscr{N}(z \mid \mu, \Sigma) \,||\, q(z)\right).
$$

We again use Eqn. (1) the closed form expression for the KL divergence between two Gaussians,

$$
\mathrm{KL}(p||q) = \frac{1}{2}\left(\mathrm{tr}\left(V^{-1}\Sigma\right) + (\mu - m)^\top V^{-1}(\mu - m) + \log|V| - \log|\Sigma| - D\right)
$$

Then, we again have

$$
\nabla_m \mathrm{KL}(p||q) = V^{-1}(\mu - m) = 0 \Rightarrow m^\star \qquad = \mu \qquad (5)
$$

We will work with $L = V^{-1}$ out of convenience here, where $\lambda_d = \frac{1}{v_d^2}$,

$$
\begin{aligned}
\nabla_L \mathrm{KL}(p||q) &= \frac{1}{2}\nabla_L \left(\mathrm{tr}(L\Sigma) + (\mu - m)^\top L(\mu - m) - \log|L|\right) \\
&= \frac{1}{2}\mathrm{diag}[\Sigma] + (\mu - m)(\mu - m)^\top - L^{-1} \\
\Rightarrow V^\star &= \mathrm{diag}[\Sigma] + (\mu - m)(\mu - m)^\top \qquad (6)
\end{aligned}
$$

(c) Plot the contour lines of your solutions to parts (a) and (b) for the case where

$$
\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}.
$$

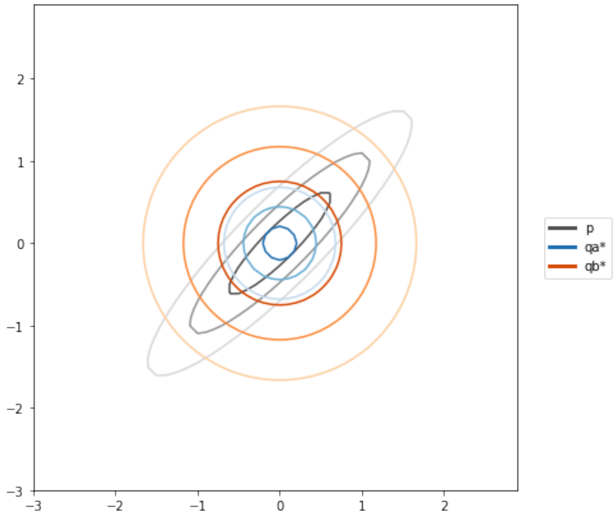

Figure 1: Contour lines of true distribution $p\ \mathcal{N}(\mu, \Sigma)$, $q_a^\star$ which minimizes $\mathrm{KL}(q||p)$ and exhibits mode-seeking behavior, and $q_b^\star$ which minimizes $\mathrm{KL}(p||q)$ and exhibits mean-seeking/mass-covering behavior.

Problem 2: *Variational autoencoders (VAE's)*

In class we derived VAE's as generative models $p(x, z; \theta)$ of observations $x \in \mathbb{R}^P$ and latent variables $z \in \mathbb{R}^D$, with parameters θ. We used variational expectation-maximization to learn the parameters θ that maximize a lower bound on the marginal likelihood,

$$\log p(x;\theta) \geq \sum_{n=1}^{N} \mathbb{E}_{q(z_n|x_n,\phi)} [\log p(x_n, z_n; \theta) - \log q(z_n \mid x_n, \phi)] \triangleq \mathscr{L}(\theta, \phi).$$

The difference between VAE's and regular variational expectation-maximization is that we constrained the variational distribution $q(z \mid x, \phi)$ to be a parametric function of the data; for example, we considered,

$$q(z_n \mid x_n, \phi) = \mathcal{N}\left(z_n \mid \mu(x_n; \phi), \operatorname{diag}([\sigma_1^2(x_n; \phi), \ldots, \sigma_D^2(x_n; \phi)])\right),$$

where $\mu : \mathbb{R}^P \to \mathbb{R}^D$ and $\sigma_d^2 : \mathbb{R}^P \to \mathbb{R}_+$ are functions parameterized by ϕ that take in a datapoint x_n and output means and variances of z_n, respectively. In practice, it is common to implement these functions with neural networks. Here we will study VAE's in some special cases. For further reference, see Kingma and Welling (2019), which is linked on the course website.

(a) Consider the linear Gaussian model factor model,

$$p(x_n, z_n; \theta) = \mathcal{N}(z_n; 0, I)\,\mathcal{N}(x_n \mid Az_n, V),$$

where $A \in \mathbb{R}^{P\times D}$, $V \in \mathbb{R}^{P\times P}$ is a diagonal, positive definite matrix, and $\theta = (A, V)$. Solve for the true posterior $p(z_n \mid x_n, \theta)$.

Following the marginal and conditional Gaussian distribution expressions as given in Bishop, Chapter 2.3 (p93), the true posterior is given by

$$p(z_n \,|\, x_n) = \mathcal{N}(\Sigma A^\mathsf{T} V^{-1} x_n, \Sigma) \tag{7}$$
$$\text{where } \Sigma = (I + A^\mathsf{T} V^{-1} A)^{-1}$$

(b) Consider the variational family of Gaussian densities with diagonal covariance, as described above, and assume that $\mu(x; \phi)$ and $\log \sigma_d^2(x; \phi)$ are linear functions of x. Does this family contain the true posterior? Find the member of this variational family that maximizes $\mathscr{L}(\theta, \phi)$ for fixed θ. (Hint: use your answer to Problem 1a.)

This variational family contains the true posterior only if $A^\mathsf{T} A$ is diagonal, or equivalently if the rows of A are linearly independent/orthogonal to each other. Then, the covariance matrix of the posterior (7) Σ is also diagonal.

We minimize the KL divergence between the true and approximate posteriors to find the member of $\mathcal{Q}$ which maximizes the evidence lower bound:

$$q^\star = \arg\min_{q\in\mathcal{Q}} \operatorname{KL}(q(z_n \,|\, x_n; \phi) || p(z_n \,|\, x_n; \theta))$$

Using the results from problem 1A, we have

$$\mu^\star_\phi = \mu_\theta = (I + A^\mathsf{T} V^{-1} A)^{-1} x_n$$
$$\Sigma^\star_\phi = \mathrm{diag}\left[\Sigma_\theta^{-1}\right]^{-1} = \mathrm{diag}\left[(I + A^\mathsf{T} V^{-1} A)\right]^{-1}$$

(c) Now consider a simple nonlinear factor model,

$$p(x_n, z_n; \theta) = \mathcal{N}(z_n \mid 0, I) \prod_{p=1}^{P} \mathcal{N}(x_{np} \mid e^{a_p^\mathsf{T} z_n}, v_p),$$

parameterized by $a_p \in \mathbb{R}^D$ and $v_p \in \mathbb{R}_+$. The posterior is no longer Gaussian, since the mean of x_{np} is a nonlinear function of the latent variable.[1]

Generate a synthetic dataset by sampling $N = 1000$ datapoints from a $D = 1$, $P = 2$ dimensional model with $A = [1.2, 1]^\mathsf{T}$ and $v_p = 0.1$ for $p = 1, 2$. Use the reparameterization trick and automatic differentiation to perform stochastic gradient descent on $-\mathcal{L}(\theta, \phi)$.

Make the following plots:

- A scatter plot of your simulated data (with equal axis limits).
- A plot of $\mathcal{L}(\theta, \phi)$ as a function of SGD iteration.
- A plot of the model parameters $(A_{11}, A_{21}, v_1, v_2)$ as a function of SGD iteration.
- The approximate Gaussian posterior with mean $\mu(x; \phi)$ and variance $\sigma_1^2(x; \phi)$ for $x \in \{(0,0), (1,1), (10,7)\}$ using the learned parameters ϕ.
- The true posterior at those points. (Since z is one dimensional, you can compute the true posterior with numerical integration.)

Comment on your results.

Both the encoder and the decoder are a neural network that parameterizes a Gaussian distribution (with diagonal covariance):

$$(\boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \text{EncoderNeuralNet}_\phi(\boldsymbol{x})$$
$$q_\phi(\mathbf{z} \mid \boldsymbol{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2))$$

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; 0, \mathbb{I})$$
$$\mathrm{p} = \text{DecoderNeuralNet}_\theta(\boldsymbol{z})$$

The weights of each network is shared between their respective means and covariances. To ensure that a valid covariance is outputted, the last layer of the covariance pipe is a SoftPlus function.

[1] For this particular model, the expectations in $\mathcal{L}(\theta, \phi)$ can still be computed in closed form using the fact that $\mathbb{E}[e^z] = e^{\mu + \frac{1}{2}\sigma^2}$ for $z \sim \mathcal{N}(\mu, \sigma^2)$.

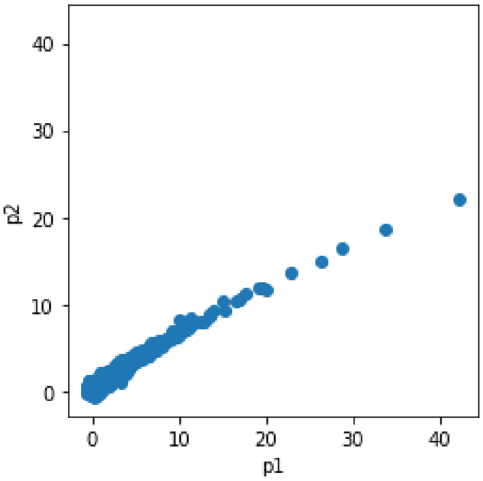

Figure 2: Synthetic data

The ELBO was then calculated using the reparameterization trick, by introducing a new random variable $\epsilon \sim \mathcal{N}(0, I)$ to take the stochasticity out of the z variable.

$$\begin{aligned}
\epsilon &\sim \mathcal{N}(0, \mathbb{I}) \\
\mu_\phi, \sigma_\phi^2 &\leftarrow \text{EncoderNN} \\
z &= \mu + \sqrt{\sigma^2} * \epsilon) \\
\log q &= \sum_p \log \mathcal{N}(z \,|\, \mu, \sigma^2)
\end{aligned}$$

$$\begin{aligned}
\mu_\theta, \sigma_\theta^2 &\leftarrow \text{DecoderNN} \\
\log p &= \sum_p \log \mathcal{N}(z \,|\, 0, \mathbb{I}) + \log \mathcal{N}(\text{batch} \,|\, \mu_\theta, \sigma_\theta^2) \\
\mathcal{L}(\theta, \phi) &= \frac{1}{N} \sum_n \log p - \log q
\end{aligned}$$

I was unable to figure out how to save intermediate values of the means and variances through iteration, since they are Traced objects and therefore abstracted. Associated code can be found in the colab notebook (link in the comments section of the upload, I was having trouble exporting it as a PDF due to Latex compiltion issues and my write-up is not letting me write the link down verbatim).

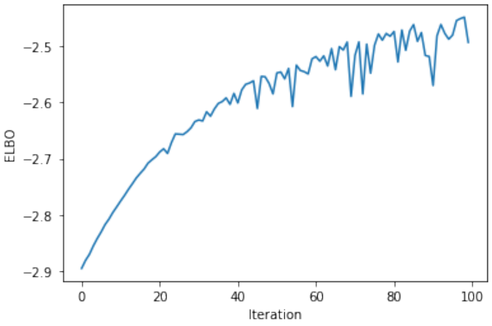

Figure 3: ELBO as function of SGD iteration

Problem 3: *Semi-Markov models*

Consider a Markov model as described in class and in, for example, Chapter 13 of *Pattern Recogntion and Machine Learning* by Bishop,

$$p(z_{1:T} \mid \pi, A) = p(z_1 \mid \pi) \prod_{t=2}^{T} p(z_t \mid z_{t-1}, A),$$

where $z_t \in \{1, \dots, K\}$ denotes the "state," and

$$p(z_1 = i) = \pi_i$$
$$p(z_t = j \mid z_{t-1} = i, A) = A_{ij}.$$

We will study the distribution of state durations—the length of time spent in a state before transitioning. Let $d \geq 1$ denote the number of time steps before a transition out of state z_1. That is, $z_1 = i, \dots, z_d = i$ for some i, but $z_{d+1} \neq i$.

(a) Show that $p(d \mid z_1 = i, A) = \text{Geom}(d \mid p_i)$, the probability mass function of the geometric distribution. Solve for the parameter p_i as a function of the transition matrix A.

$$\begin{aligned} p(d \,|\, z_1 = i, A) &= p(\{z_t = i\}_{t=1}^{d}, z_{d+1} \neq i \,|\, z_1 = i,\, A) \\ &= p(z_{d+1} \neq i \,|\, z_d = i, A)\, p(\{z_t = i\}_{t=1}^{d} \,|\, z_1 = i,\, A) \\ &= p(z_{d+1} \neq i \,|\, z_d = i, A)\, p(z_t = i \,|\, z_{t-1} = i,\, A) \\ &= (1 - A_{ii}) A_{ii}^{d-1} = \text{Geom}(d - 1 \,|\, 1 - A_{ii}) \end{aligned}$$

so the parameter $p_i = 1 - A_{ii}$, i.e. the probability of transitioning to another state.

(b) We can equivalently represent $z_{1:T}$ as a set of states and durations $\{(\tilde{z}_n, d_n)\}_{n=1}^{N}$, where $\tilde{z}_n \in \{1, \dots, K\} \setminus \{\tilde{z}_{n-1}\}$ denotes the index of the n-th visited state and $d_n \in \mathbb{N}$ denotes the duration spent in that state before transition. There is a one-to-one mapping between states/durations and the original state sequence:

$$(z_1, \dots, z_T) = (\underbrace{\tilde{z}_1, \dots, \tilde{z}_1}_{d_1 \text{ times}}, \underbrace{\tilde{z}_2, \dots, \tilde{z}_2}_{d_2 \text{ times}}, \dots \underbrace{\tilde{z}_N, \dots, \tilde{z}_N}_{d_N \text{ times}}).$$

Show that the probability mass function of the states and durations is of the form

$$p(\{(\tilde{z}_n, d_n)\}_{n=1}^{N}) = p(\tilde{z}_1 \mid \pi) \left[\prod_{n=1}^{N-1} p(d_n \mid \tilde{z}_n, A)\, p(\tilde{z}_{n+1} \mid \tilde{z}_n, A) \right] p(d_N \mid \tilde{z}_N, A),$$

and derive each conditional probability mass function.

The durations of each state is only conditionally dependent on the state itsel, and the next state is only conditionally dependent on its immediately preceding state. This Markov/duration model provides us with strong structures to simplify the joint distribution as follows:

$$\begin{aligned} p(\{(\tilde{z}_n, d_n)\}_{n=1}^{N}) &= p(\tilde{z}_1 \,|\, \pi)\, p(d_1 \,|\, \tilde{z}_1,\, A)\, p(\tilde{z}_2 \,|\, \tilde{z}_1,\, A)\, p(d_2 \,|\, \tilde{z}_2,\, A) \dots p(d_N \,|\, \tilde{z}_N,\, A) \\ &= p(\tilde{z}_1 \mid \pi) \left[\prod_{n=1}^{N-1} p(d_n \mid \tilde{z}_n, A)\, p(\tilde{z}_{n+1} \mid \tilde{z}_n, A) \right] p(d_N \mid \tilde{z}_N, A) \end{aligned}$$

Then, the conditional PMF of each term is

$$\begin{aligned} p(\tilde{z}_1 \mid \pi) &= p(z_1 = \tilde{z}_1 \,|\, \pi) = \pi_{\tilde{z}_1} \\ p(\tilde{z}_{n+1} \mid \tilde{z}_n, A) &= A_{\tilde{z}_{n+1}\tilde{z}_n} && \text{for } n = 1, \dots, N-1 \\ p(d_n \mid \tilde{z}_n, A) &= \mathrm{Geom}(d_n - 1 \,|\, p_{\tilde{z}_n}) && \text{for } p_{\tilde{z}_n} = A_{\tilde{z}_n \tilde{z}_n}, n = 1, \dots, N \end{aligned}$$

(c) *Semi-Markov* models replace $p(d_n \mid \tilde{z}_n)$ with a more flexible duration distribution. For example, consider the model,

$$p(d_n \mid \tilde{z}_n) = \mathrm{NB}(d_n \mid r, \theta_{\tilde{z}_n}),$$

where $r \in \mathbb{N}$ and $\theta_k \in [0,1]$ for $k = 1, \dots, K$. Recall from Assignment 1 that the negative binomial distribution with integer r is equivalent to a sum of r geometric random variables. Use this equivalence to write the semi-Markov model with negative binomial durations as a Markov model on an extended set of states $s_n \in \{1, \dots, Kr\}$. Specifically, write the transition matrix for $p(s_n \mid s_{n-1})$ and the mapping from s_n to z_n.

Let us index the states by m, so that there are a total of $M = Kr$ states in the extended set. This model is called semi-Markovian because for every new state $\tilde{z}_n$ that it enters, it must then sequentially visit every r associated state (non-Markov process) before it can transition to a different state $\tilde{z}_{n+1}$ (Markovian).

As an example, suppose $K = 4$ and $R = 2$, then our sets look like

$$\begin{aligned} &\tilde{z} = \{1, 2, 3, 4\} \\ &s = \{1_1, 1_2, 2_1, 2_2, 3_1, 3_2, 4_1, 4_2\} = 1, 2, 3, 4, 5, 6, 7, 8 \end{aligned}$$

When we enter $\tilde{z}_n = 1$ for example, $s_m = 1_1$ and $s_{m+1} = 1_2$ before it can transition to any other $\tilde{z}_{n+1} \neq \tilde{z}_n$. Then, let the map from s_m to $\tilde{z}_n$ be defined by $M(s_m) = \lceil s_m / r \rceil$. We also define the boolean functions identifying if a state is a entry/beginning state ($B(s_m) = [s_m \% R == 1]$) or a terminal state $T(s_m) = [s_m \% R == 0]$. Therefore,

$$p(s_m \,|\, s_{m-1}) = \begin{cases} \theta_{\tilde{z}_n} & \text{if } M(s_{m-1}) == M(s_m) \\ A_{\tilde{z}_{n+1}\tilde{z}_n} & \text{if } M(s_{m-1})\,! = M(s_m),\ T(s_{m-1}) \text{ and } B(s_m) \\ 0 & \text{otherwise} \end{cases}$$