CS 181: Machine Learning April 21, 2023

Midterm II Review Solutions

George Cai (georgecai@college.harvard.edu) *Matthew Qu (matthewqu@college.harvard.edu)*

1. (PCA on transformed data[1])

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a data matrix, with the j^{th} row corresponding to the j^{th} observation $\mathbf{x}_j^\top \in \mathbb{R}^d$. Assume $\mathbf{X}$ is centered, and suppose that the PCA of this data has principal components (eigenvectors) $\mathbf{v}_1, \ldots, \mathbf{v}_d$ with associated variances (eigenvalues) $\lambda_1 \geq \ldots, \geq \lambda_d \geq 0$. Now, let $\mathbf{Q}$ be a $d \times d$ orthonormal matrix and let $\mathbf{y}_j = \mathbf{Q}\mathbf{x}_j$ for all $j = 1, \ldots, n$.

(a) In words, briefly explain how the matrix $\mathbf{Q}$ transforms the data $\mathbf{x}_j$. Then, find an expression for $\mathbf{Y} \in \mathbb{R}^{n \times d}$, the new data matrix where the j^{th} row corresponds to $\mathbf{y}_j^\top$, in terms of $\mathbf{X}$ and $\mathbf{Q}$.

(b) Show that the PCA of $\mathbf{Y}$ yields principal components $\mathbf{Q}\mathbf{v}_1, \ldots, \mathbf{Q}\mathbf{v}_d$ with associated variances $\lambda_1, \ldots, \lambda_d$. Briefly explain why this result is intuitive.

(c) Now, suppose that $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$ is the SVD of $\mathbf{X}$. Noting that the columns of $\mathbf{V}$ contain the eigenvectors of the covariance matrix of $\mathbf{X}$, find an equation that relates the singular values of $\mathbf{S}$, s_j, with the PCA variances λ_j (assuming that the diagonal values of $\mathbf{S}$ are sorted in decreasing order).

(d) *With a graphical explanation*, show how performing PCA on non-centered data can yield incorrect principal components and variances.

[1] Adapted from STAT 185, Fall 2022

Solution.

(a) Orthonormal matrices represent transformations composed of rotations and reflections, so the data is being changed through these length-preserving transformations. Since $\mathbf{y}_j = \mathbf{Q}\mathbf{x}_j$, we know that $\mathbf{y}_j^\top = \mathbf{x}_j^\top \mathbf{Q}^\top$, which implies that $\mathbf{Y} = \mathbf{X}\mathbf{Q}^\top$.

(b) Let $\mathbf{\Sigma}_X = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ denote the covariance matrix of $\mathbf{X}$. Then, consider the covariance matrix of $\mathbf{Y}$:

$$\begin{aligned}\mathbf{\Sigma}_Y &= \frac{1}{n}\mathbf{Y}^\top\mathbf{Y} \\ &= \frac{1}{n}(\mathbf{X}\mathbf{Q}^\top)^\top(\mathbf{X}\mathbf{Q}^\top) \\ &= \frac{1}{n}\mathbf{Q}\mathbf{X}^\top\mathbf{X}\mathbf{Q}^\top \\ &= \mathbf{Q}\mathbf{\Sigma}_X\mathbf{Q}^\top.\end{aligned}$$

Note that in order for this derivation to be correct, we must show that $\mathbf{Y}$ is centered given that $\mathbf{X}$ is centered; this is left as an additional exercise. Since we know the principal components and variances of the data $\mathbf{X}$, we can write $\mathbf{\Sigma}_X = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$, where the columns of $\mathbf{V}$ are the eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_d$ and $\mathbf{\Lambda}$ is diagonal with entries $\lambda_1, \ldots, \lambda_d$. This means that

$$\mathbf{\Sigma}_Y = \mathbf{Q}\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top\mathbf{Q}^\top = (\mathbf{Q}\mathbf{V})\mathbf{\Lambda}(\mathbf{Q}\mathbf{V})^\top.$$

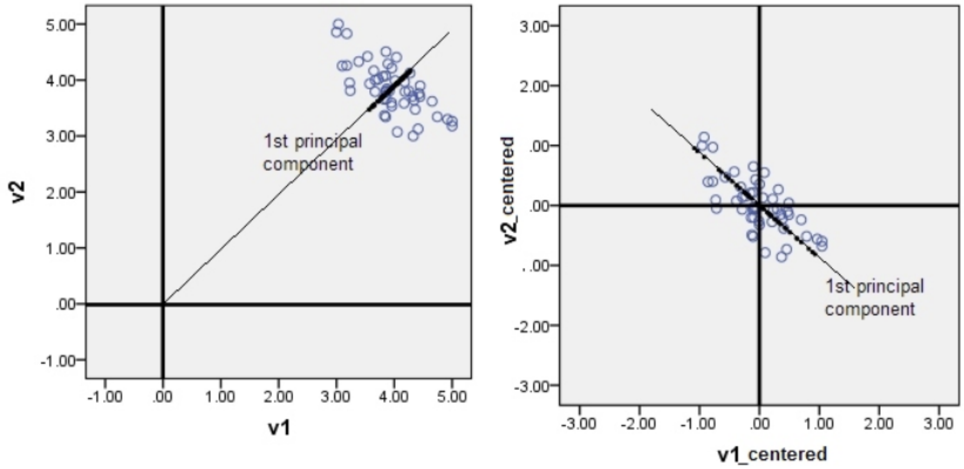
Therefore, the eigendecomposition of $\mathbf{\Sigma}_Y$ yields eigenvectors which are the columns of $\mathbf{Q}\mathbf{V}$ and eigenvalues which are the diagonal entries of $\mathbf{\Lambda}$. This corresponds exactly to $\mathbf{Y}$ having principal components $\mathbf{Q}\mathbf{v}_1, \ldots, \mathbf{Q}\mathbf{v}_d$ with associated variances $\lambda_1, \ldots, \lambda_d$. This result makes sense because since $\mathbf{Q}$ transforms the data through rotations and reflections, the principal components will also be transformed via the same transformation. Since $\mathbf{Q}$ is distance-preserving, this also means that the variances are unaffected.

(c) Given the SVD of $\mathbf{X}$, we have

$$\begin{aligned}\mathbf{\Sigma}_X &= \frac{1}{n}\mathbf{X}^\top\mathbf{X} \\ &= \frac{1}{n}(\mathbf{U}\mathbf{S}\mathbf{V}^\top)^\top(\mathbf{U}\mathbf{S}\mathbf{V}^\top) \\ &= \frac{1}{n}\mathbf{V}\mathbf{S}^\top\mathbf{U}^\top\mathbf{U}\mathbf{S}\mathbf{V}^\top \\ &= \frac{1}{n}\mathbf{V}\mathbf{S}^\top\mathbf{S}\mathbf{V}^\top \\ &= \mathbf{V}\frac{\mathbf{S}^2}{n}\mathbf{V}^\top.\end{aligned}$$

The fourth equality is due to $\mathbf{U}$ being unitary, so $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$, while the fifth equality is due to $\mathbf{S}$ being diagonal, so $\mathbf{S}^\top = \mathbf{S}$. From this derivation, we arrive at the eigendecomposition of $\mathbf{\Sigma}_X$, so we have $\mathbf{\Lambda} = \frac{\mathbf{S}^2}{n}$. This means that the singular values s_j of $\mathbf{X}$ are related to the PCA variances λ_j through the equation $\lambda_j = \frac{s_j^2}{n}$.

(d) Consider the figure below:



The general direction of the principal components can be explained by noting that PCA minimizes the reconstruction error, that is, it minimizes the sum of the distances from the original data and the data projected onto the subspace spanned by the principal component(s). By not centering the data (left), the principal component is very different from the true direction of maximal variance when the data is centered (right).

2. (**Expectation maximization on multinomial data**[2])

Recall that the multinomial distribution is a generalization of the binomial distribution where there are $k \geq 3$ possible categories. If $\mathbf{x} = (x_1, \ldots, x_k) \sim \text{Mult}(n, \boldsymbol{\pi})$, its PMF is given by

$$p(\mathbf{x}) = \frac{n!}{x_1! \cdots x_k!} \pi_1^{x_1} \ldots \pi_k^{x_k},$$

where $x_j \geq 0$ for all $j = 1, \ldots, k$, $\sum_{j=1}^k x_j = n$, and $\sum_{j=1}^k \pi_j = 1$. Suppose we have a single observation $\mathbf{x} = (x_1, x_2, x_3, x_4)$ from a $\text{Mult}(n, \boldsymbol{\pi}_\theta)$ distribution, where

$$\boldsymbol{\pi}_\theta = \left(\frac{1}{2} + \frac{1}{4}\theta, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{1}{4}\theta \right).$$

However, assume that the complete data is given by $\mathbf{z} = (z_0, z_1, x_2, x_3, x_4) \sim \text{Mult}(n, \boldsymbol{\pi}_\theta^*)$, where

$$\boldsymbol{\pi}_\theta^* = \left(\frac{1}{2}, \frac{1}{4}\theta, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{1}{4}\theta \right).$$

That is, we have the latent variables z_0 and z_1, but we only observe $x_1 = z_0 + z_1$.

(a) Write out both the observed data log-likelihood, $\ell(\theta; \mathbf{x})$, and the complete data log-likelihood, $\ell(\theta; \mathbf{x}, z_1)$, up to an additive constant with respect to θ. Why does the complete data log-likelihood not depend on z_0?

(b) What is the conditional distribution $p(z_1 \mid \mathbf{x}, \theta)$? Briefly justify your answer.

(c) Let θ^t denote the current value of θ at the t^{th} iteration of the EM algorithm. Write out the ELBO function $\text{ELBO}(\theta \mid q, \theta^t)$, where $q(z_1)$ is the posterior distribution $p(z_1 \mid \mathbf{x}, \theta^t)$. You may leave your answer in terms of named distributions and the complete data log-likelihood (i.e., you do not need to plug in the PMF/PDFs of any distributions in your answer).

(d) Now, write out the expression which must be maximized in the M-step of the EM algorithm. Simplify as much as possible: your answer should be in the form $\underset{\theta}{\text{argmax}}\, g(\theta)$, where $g(\theta)$ contains no expectations or other terms not dependent on θ.

(e) Finally, compute θ^{t+1} in terms of θ^t and $\mathbf{x}$ by finding the maximum value of the expression you derived in part (d).

[2] Adapted from `http://www.columbia.edu/~mh2078/MachineLearningORFE/EM_Algorithm.pdf`

Solution.

(a) The observed data likelihood is given by

$$L(\theta; \mathbf{x}) = \frac{n!}{x_1!x_2!x_3!x_4!} \left(\frac{1}{2} + \frac{1}{4}\theta\right)^{x_1} \left(\frac{1}{4}(1-\theta)\right)^{x_2} \left(\frac{1}{4}(1-\theta)\right)^{x_3} \left(\frac{1}{4}\theta\right)^{x_4}$$

Therefore, the observed data log-likelihood is

$$\ell(\theta; \mathbf{x}) = x_1 \log\left(\frac{1}{2} + \frac{1}{4}\theta\right) + (x_2 + x_3)\log(1-\theta) + x_4 \log\theta + C,$$

where C is a constant not depending on θ. Similarly, the complete data log-likelihood is

$$\ell(\theta; \mathbf{x}, z_1) = z_0 \log\frac{1}{2} + (z_1 + x_4)\log\theta + (x_2 + x_3)\log(1-\theta) + C.$$

Note that the first term has no dependence on θ, so it can be absorbed into the constant C. Thus, the complete data log-likelihood does not depend on z_0.

(b) The posterior distribution is

$$p(z_1 \mid \mathbf{x}, \theta) \sim \text{Bin}\left(x_1, \frac{\theta/4}{1/2 + \theta/4}\right).$$

This is because given x_1, we know that the conditional distribution of z_1 reduces to a binomial distribution, where $x_1 = z_0 + z_1$ is the number of trials, and the probability has been appropriately normalized.

(c) The ELBO function is

$$\begin{aligned}\text{ELBO}(\theta \mid q, \theta^t) &= \mathbb{E}_{z_1 \sim q(z_1)}\left[\log\frac{p(\mathbf{x}, z_1|\theta)}{q(z_1)}\right] \\ &= \mathbb{E}_{z_1 \sim q(z_1)}[\ell(\theta; \mathbf{x}, z_1)] - \mathbb{E}_{z_1 \sim q(z_1)}[\log q(z_1)],\end{aligned}$$

where $q(z_1) \sim \text{Bin}\left(x_1, \frac{\theta^t/4}{1/2+\theta^t/4}\right)$. Note that $q(z_1)$, as found in the E-step of the EM algorithm, has no dependence on θ.

(d) In the M-step of the EM algorithm, we want to maximize the ELBO function found in part (c) with respect to θ. The second term $\mathbb{E}_{z_1 \sim q(z_1)}[\log q(z_1)]$ has no dependence on θ, so we have

$$\begin{aligned}\underset{\theta}{\text{argmax}}\ \text{ELBO}(\theta \mid q, \theta^t) &= \underset{\theta}{\text{argmax}}\, \mathbb{E}_{z_1 \sim q(z_1)}[\ell(\theta; \mathbf{x}, z_1)] \\ &= \underset{\theta}{\text{argmax}}\, \mathbb{E}_{z_1 \sim q(z_1)}[z_1 \log\theta] + x_4 \log\theta + (x_2 + x_3)\log(1-\theta) \\ &= \underset{\theta}{\text{argmax}}\, x_1 p^t \log\theta + x_4 \log\theta + (x_2 + x_3)\log(1-\theta),\end{aligned}$$

where $p^t = \frac{\theta^t/4}{1/2+\theta^t/4}$.

(e) To maximize the expression found in part (d) with respect to θ, we take the derivative and set it equal to 0:

$$0 = \frac{x_1 p^t + x_4}{\theta} - \frac{x_2 + x_3}{1-\theta}.$$

Solving for θ yields

$$\theta^{t+1} = \frac{x_4 + x_1 p^t}{x_2 + x_3 + x_4 + x_1 p^t}.$$

3. (Markov decision process of a caterpillar[3])

George the very hungry caterpillar loves to eat. Because George wants to grow up and become a butterfly, he is trying to eat as many calories as possible. At every meal, he decides between eating watermelon and strawberry ice cream. Eating watermelon gives him 4 calories, while eating ice cream gives him 10 calories. However, eating too much ice cream may cause George to become sick, and eating ice cream while sick may cause him to die! (This would be bad because then George can no longer eat.) On the other hand, eating watermelon will keep George healthy and make him healthy if he is sick. George will always be in one of these three states: healthy, sick, or dead—the transitions are given in the table below.

Health condition	Watermelon or Ice Cream?	Next condition	Probability
healthy	watermelon	healthy	1
healthy	ice cream	healthy	1/4
healthy	ice cream	sick	3/4
sick	watermelon	healthy	1/4
sick	watermelon	sick	3/4
sick	ice cream	sick	7/8
sick	ice cream	dead	1/8

(a) Model this problem as an MDP by specifying the states $\mathcal{S}$, actions $\mathcal{A}$, transition functions $T^a(s, s') = P(s' \mid s, a)$, and reward function $R(a)$. Note that in this context, the reward function does not depend on the current state s or subsequent state s'.

(b) Run **value iteration** for 2 iterations on this MDP with $\gamma = 0.8$, specifying how the functions $Q_t(s, a)$ and $V_t(s)$ change over each iteration. That is, start with $Q_0 = 0, V_0 = 0$ and compute $Q_1(s, a), V_1(s), Q_2(s, a)$, and $V_2(s)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

(c) Let π_0 be the policy that George always eats ice cream regardless of his health condition. Find the value of this policy, $V^{\pi_0}(s)$, for each state $s \in \mathcal{S}$ in terms of the discount factor γ. Then, for each state-action pair (s, a), find $Q^{\pi_0}(s, a)$ in terms of V^{π_0}.

(d) Now, run **policy iteration**, starting with the initial policy π_0 where George always eats ice cream. Setting $\gamma = 0.8$, use your results in part (c) to find the optimal policy. How many iterations does your algorithm take?

(e) Run policy iteration again, but now with $\gamma = 0.9$. How many iterations does it take for your policy to converge? If your optimal policy differs that of part (d), briefly describe your intuition for why this might occur.

(f) (**Bonus**) Does there exist a reward function $R'(a)$ with $R'(\text{watermelon}) < R'(\text{ice cream})$ and some $\gamma \in [0, 1)$ such that the optimal policy π^* satisfies $\pi^*(\text{healthy}) = \text{watermelon}$? If so, find R' and γ, or prove that such an R' cannot exist.

[3] Adapted from CS 182, Fall 2020

Solution.

(a) We have states $\mathcal{S} = \{H, S, D\}$ representing the health conditions healthy, sick, and dead, respectively. The actions are $\mathcal{A} = \{W, I\}$, which represent eating watermelon and ice cream, respectively. The transition functions are given as matrices below:

$$T^W(s,s') = P(s' \mid s, W) = \begin{array}{c} \\ H \\ S \\ D \end{array}\begin{array}{c} \begin{array}{ccc} H & S & D \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 1/4 & 3/4 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array}, \quad T^I(s,s') = P(s' \mid s, I) = \begin{array}{c} \\ H \\ S \\ D \end{array}\begin{array}{c} \begin{array}{ccc} H & S & D \end{array} \\ \begin{pmatrix} 1/4 & 3/4 & 0 \\ 0 & 7/8 & 1/8 \\ 0 & 0 & 1 \end{pmatrix} \end{array},$$

where s is given by the row label and s' is given by the column label. Finally, we define $R(W) = 4$ and $R(I) = 10$.

(b) Recall the value iteration algorithm:

1. Initialize $Q_0(s,a) = 0$, $V_0(s) = 0$ for all $s \in \mathcal{S}$, $a \in \mathcal{A}$.
2. For $t = 1, 2, \ldots$, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, compute

$$Q_{t+1}(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s,a) V_t(s').$$

3. For all $s \in \mathcal{S}$, set

$$V_{t+1}(s) = \max_a Q_{t+1}(s,a).$$

4. Repeat steps 2 and 3 until convergence.

In all subsequent parts of this question, we will denote the Q function as a 3×2 matrix corresponding to all pairs $(s,a) \in \mathcal{S} \times \mathcal{A}$. We have

$$Q_1(s,a) = \begin{array}{c} \\ H \\ S \\ D \end{array}\begin{array}{c} \begin{array}{cc} W & I \end{array} \\ \begin{pmatrix} 4 & 10 \\ 4 & 10 \\ 0 & 0 \end{pmatrix} \end{array},$$

and $V_1(H) = V_1(S) = 10$ (we won't specify $V(D)$ because it will always be 0). From this, we get

$$Q_2(s,a) = \begin{array}{c} \\ H \\ S \\ D \end{array}\begin{array}{c} \begin{array}{cc} W & I \end{array} \\ \begin{pmatrix} 12 & 18 \\ 12 & 17 \\ 0 & 0 \end{pmatrix} \end{array},$$

with $V_2(H) = 18$ and $V_2(S) = 17$.

(c) For a policy π, we know that V^π satisfies the following equations for all $s \in \mathcal{S}$:

$$V^\pi = r(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, \pi(s)) V^\pi(s').$$

Similarly, the Q function can be evaluated for all $(s,a) \in \mathcal{S} \times \mathcal{A}$:

$$Q^\pi(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s,a) Q^\pi(s', \pi(s'))$$

Given the policy π_0, we have the following system of equations:

$$\begin{aligned} V^{\pi_0}(H) &= 10 + \frac{3}{4}\gamma V^{\pi_0}(S) + \frac{1}{4}\gamma V^{\pi_0}(H) \\ V^{\pi_0}(S) &= 10 + \frac{7}{8}\gamma V^{\pi_0}(S) + \frac{1}{8}\gamma V^{\pi_0}(D) \\ V^{\pi_0}(D) &= 0. \end{aligned}$$

These yield the solutions

$$V^{\pi_0}(H) = \frac{320 - 40\gamma}{(4-\gamma)(8-7\gamma)}, \quad V^{\pi_0}(S) = \frac{80}{8-7\gamma}.$$

Since $Q(s, \pi(s) = V(s)$, we know that $Q^{\pi_0}(H, I) = V^{\pi_0}(H)$ and $Q^{\pi_0}(S, I) = V^{\pi_0}(S)$. Finally, we also have

$$\begin{aligned} Q^{\pi_0}(S, W) &= 4 + \frac{1}{4}\gamma V^{\pi_0}(H) + \frac{3}{4}\gamma V^{\pi_0}(S) \\ Q^{\pi_0}(H, W) &= 4 + \gamma V^{\pi_0}(H). \end{aligned}$$

(d) Recall the policy iteration algorithm:

1. Start with some policy $\pi_0 : \mathcal{S} \to \mathcal{A}$.
2. For $t = 0, 1, \ldots$, compute V^{π_t} and Q^{π_t} by policy evaluation (via the exact or iterative method).
3. For all $s \in \mathcal{S}$, set
$$\pi_{t+1}(s) = \operatorname*{argmax}_a Q^{\pi_t}(s, a).$$
4. Repeat until $\pi_{t+1} = \pi_t$.

Using this and the equations from part (c), we have the following Q function for $\gamma = 0.8$:

$$Q^{\pi_0}(s, a) = \begin{array}{c} \\ H \\ S \\ D \end{array} \begin{array}{c} \begin{array}{cc} W & I \end{array} \\ \begin{pmatrix} 34 & 37.5 \\ 31.5 & 33.3 \\ 0 & 0 \end{pmatrix} \end{array}.$$

Therefore, the policy π_1 is to eat ice cream at every state (once again, disregarding state D). Thus, the policy converges after a single iteration.

(e) Now, with $\gamma = 0.9$, we get the following result:

$$Q^{\pi_0}(s, a) = \begin{array}{c} \\ H \\ S \\ D \end{array} \begin{array}{c} \begin{array}{cc} W & I \end{array} \\ \begin{pmatrix} 52.5 & 53.9 \\ 47.9 & 47.1 \\ 0 & 0 \end{pmatrix} \end{array}.$$

Thus, we have $\pi_1(H) = I$, but $\pi_1(S) = W$. Then, evaluating this policy yields

$$Q^{\pi_1}(s, a) = \begin{array}{c} \\ H \\ S \\ D \end{array} \begin{array}{c} \begin{array}{cc} W & I \end{array} \\ \begin{pmatrix} 57.6 & 59.5 \\ 53.5 & 52.1 \\ 0 & 0 \end{pmatrix} \end{array}.$$

This yields $\pi_2(H) = I$ and $\pi_2(S) = W$, so at this point, our policy has converged after two iterations. The reason why the policy differs for different values of γ is because values of γ closer to 1 prioritize long-term gains (i.e. not dying), whereas smaller values of γ yield policies that look for immediate gain (i.e. eating as much ice cream immediately).

(f) Suppose we have reward values $R(W)$ and $R(I)$ for the reward of eating watermelon and ice cream, respectively, with $R(W) < R(I)$. Also, suppose that the optimal policy when healthy is to eat watermelon. Then, we must have $V(H) = R(W) + \gamma V(H)$, or $V(H) = (1-\gamma)^{-1}R(W)$. Furthermore, because it is the optimal policy, the value of eating vegetables when healthy must be at least the value of eating ice cream; that is, we have

$$R(W) + \gamma V(H) \geq R(I) + \frac{3}{4}\gamma V(S) + \frac{1}{4}\gamma V(H).$$

This simplifies to $V(H) \geq \frac{4}{3\gamma}(R(I) - R(W)) + V(S)$. In particular, because $R(I) - R(W) > 0$, we have $V(H) > V(S)$. Under the policy of eating vegetables while sick, we also have

$$V(S) = R(W) + \frac{3}{4}\gamma V(S) + \frac{1}{4}\gamma V(H).$$

Substituting $V(H) = \frac{1}{1-\gamma}R(W)$, we get

$$\begin{aligned} V(S) &= \frac{4}{4-3\gamma}\left(R(W) + \frac{\gamma}{4-4\gamma}R(W)\right) \\ &= \frac{4}{4-3\gamma}R(W)\left(1 + \frac{\gamma}{4-4\gamma}\right) \\ &= \frac{4}{4-3\gamma}R(W)\left(\frac{4-3\gamma}{4-4\gamma}\right) \\ &= \frac{1}{1-\gamma}R(W) \\ &= V(H). \end{aligned}$$

In fact, $V(H)$ is a lower bound on $V(S)$ because the optimal policy could be to eat ice cream while sick. In other words, the true value of $V(S)$ under the optimal policy cannot be less than $V(H)$ because existence of the policy of eating watermelon while sick forces $V(S) \geq V(H)$. Thus, we have $V(H) > V(S) \geq V(H)$, which is impossible. Thus, there are no reward values satisfying $R(W) < R(I)$ such that the optimal policy is to eat watermelon while healthy.

4. (**Everything is a graphical model**)

For the following models[4], write the following: (i) graphical model, (ii) "generative story" of the model, (iii) observed data log likelihood, (iv) ELBO function, (v) E-step (what are you maximizing, with respect to what?), (vi) M-step (what are you maximizing, with respect to what?).

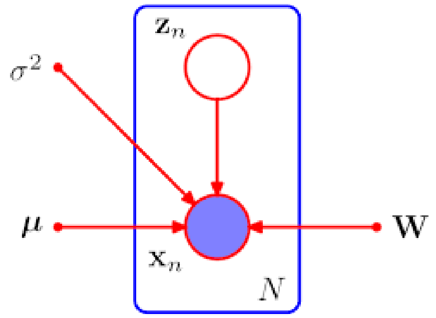
(a) pPCA: latent variables $\mathbf{z}_n \sim \mathcal{N}(0, I)$, $\mathbf{z} \in \mathbb{R}^k$, observed variables $\mathbf{x}_n | \mathbf{z}_n \sim \mathcal{N}(\mathbf{W}\mathbf{z}_n, \sigma^2 \mathbf{I})$, $\mathbf{x} \in \mathbb{R}^d$. You may find the multivariate Gaussian $\mathbf{r} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{r} \in \mathbb{R}^d$ PDF helpful:

$$p(\mathbf{r}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2} (\det \boldsymbol{\Sigma})^{-1/2} \exp - \left(\frac{(\mathbf{r} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{r} - \boldsymbol{\mu})}{2} \right)$$

(b) Hidden Markov Model: latent variables $(\mathbf{s}_1, \ldots, \mathbf{s}_n), \mathbf{s}_1 \sim \text{Cat}(\boldsymbol{\theta})$, $\mathbf{s}_t \in [0, 1]^K$ (i.e. the s_t are one-hot encoded vectors), $\boldsymbol{\theta} \in \mathbb{R}^K$, observed variables $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{x}_t | \mathbf{s}_t \sim p(\mathbf{x}_t | \mathbf{s}_t)$ (this is arbitrary, one example is linear Gaussian noise $\mathcal{N}(\mathbf{D}s_t + \mathbf{E}, \sigma^2 \mathbf{I})$, where $\mathbf{D} \in \mathbb{R}^{K \times d}, \mathbf{E} \in \mathbb{R}^d$, and $\mathbf{I}$ is the $d \times d$ identity matrix), $\mathbf{x}_t \in \mathbb{R}^d$, latent state transition probabilities $\mathbf{T}_{ij} \in \mathbb{R}^{K \times K}$, $\mathbf{T}_{ij} = P(s_{t+1} = j | s_t = i)$.

[4] For more practice, go through the midterm skills checklist and do this process for each model listed.

Solution.



(a) (i) Note that the parameter $\boldsymbol{\mu}$ is unnecessary if the data $\mathbf{X}$ is mean-centered.

(ii) For each data point, we sample $\mathbf{z}_n$ from the latent Gaussian distribution, and then we linearly map from the latent space to the observation space. Finally, we sample Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ and add it to observation.

(iii) We can marginalize over $\mathbf{z}_n$ by integrating: $p(\mathbf{x}_n; \mathbf{W}, \sigma) = \int_{-\infty}^{\infty} p(\mathbf{x}_n, \mathbf{z}_n) d\mathbf{z}_n = \int_{-\infty}^{\infty} p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n) d\mathbf{z}_n$ and substituting in the PDFs. Alternatively, we can use the property that the marginal over any particular variable in a multivariate Gaussian is going to be Gaussian, meaning we only need to find the mean and variance of $\mathbf{x}_n$. First, note that $\mu_{\mathbf{x}} = \mathbb{E}[\mathbf{x}] = \mathbb{E}[\mathbf{W}\mathbf{z} + \boldsymbol{\epsilon}] = \mathbb{E}[\mathbf{W}\mathbf{z}] + \mathbb{E}[\boldsymbol{\epsilon}] = 0$ since both our noise and our $\mathbf{z}$ are zero centered. Next,

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{xx}} &= \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\top] \\
&= \mathbb{E}[(\mathbf{W}\mathbf{z} + \boldsymbol{\epsilon} - 0)(\mathbf{W}\mathbf{z} + \boldsymbol{\epsilon} - 0)^\top] \\
&= \mathbb{E}[\mathbf{W}\mathbf{z}\mathbf{z}^T\mathbf{W}^\top + \boldsymbol{\epsilon}\mathbf{z}^\top\mathbf{W}^\top + \mathbf{W}\mathbf{z}\boldsymbol{\epsilon}^\top + \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] \\
&= \mathbf{W}\mathbb{E}[\mathbf{z}\mathbf{z}^\top]\mathbf{W}^\top + \mathbb{E}[\epsilon\mathbf{z}^\top]\mathbf{W}^\top + \mathbf{W}\mathbb{E}[\mathbf{z}\boldsymbol{\epsilon}^\top] + \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] \\
&= \mathbf{W}(\mathbf{I})\mathbf{W}^T + \sigma^2\mathbf{I}.
\end{aligned}$$

With these two parameters, we have $\mathbf{x} \sim \mathcal{N}(0, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})$. The observed data log likelihood is

$$\begin{aligned}
\ell(\mathbf{x}_n; \mathbf{W}, \sigma) &= \sum_{n=1}^{N} \log p(\mathbf{x}_n|\mathbf{W}, \sigma) \\
&= \sum_{n=1}^{N} \log\left(\frac{1}{(2\pi)^{d/2}(\det \boldsymbol{\Sigma}_{\mathbf{xx}})^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_n^\top \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{x}_n)\right\}\right) \\
&= \frac{Nd}{2}\log 2\pi + \log\det(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}) + \sum_{n=1}^{N} -\frac{1}{2}(\mathbf{x}_n^\top \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{x}_n).
\end{aligned}$$

(iv) ELBO: Recall that ELBO is defined as $\sum_{n=1}^{N} \mathbb{E}_{\mathbf{z}_n \sim q(\mathbf{z})}\left[\log \frac{p(\mathbf{x}_n, \mathbf{z}_n; \mathbf{W}, \sigma)}{q(\mathbf{z})}\right]$.

(v) E-step: In lecture, we showed that $\underset{q}{\operatorname{argmax}}$ ELBO $= \underset{q}{\operatorname{argmin}} D_{KL}(q \,||\, p(\mathbf{z}|\mathbf{x}; \mathbf{W}, \sigma))$. Therefore, assuming some initial values of $\mathbf{W}^t, \sigma^t$, we set $q^*(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{x}_n; \mathbf{W}^t, \sigma^t)$.
For some calculation practice (you would not have to do this on an exam): To find the posterior distribution, which we also know is Gaussian, we can use the following formula (found in the STAT 110 textbook, but would be provided on exam if needed): a joint Gaussian distribution $p(\mathbf{x}_1, \mathbf{x}_2) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has conditional distribution

$$p(\mathbf{x}_1|\mathbf{x}_2) \sim \mathcal{N}(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

So $p(\mathbf{z}|\mathbf{x}; \mathbf{W}, \boldsymbol{\sigma}) \sim \mathcal{N}(\boldsymbol{\Sigma}_{zx}\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{x}), \boldsymbol{\Sigma}_{zz}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xz})$, where $\boldsymbol{\Sigma}_{zz} = \mathbf{I}$, and $\boldsymbol{\Sigma}_{zx} = \mathbf{W}, \boldsymbol{\Sigma}_{xz} = \mathbf{W}^\top$.

(vi) M-step: $\underset{\mathbf{W},\sigma}{\operatorname{argmax}} \operatorname{ELBO}(\mathbf{x}_n, \mathbf{z}_n, q^*, \mathbf{W}, \sigma) = \underset{\mathbf{W},\sigma}{\operatorname{argmax}} \mathbb{E}_{z_n \sim q(\mathbf{z})}[\log p(\mathbf{x}_n, \mathbf{z}_n; \mathbf{W}, \sigma)]$.

(b) See HMM section notes.