Machine Learning 2 - Homework 7

Andrii Skliar, 11636785

deadline: May 21, 2018

During the process of solving the homework problems, I have collaborated with the following colleagues:

Sindy Lowe   Davide Belli   Gabriele Cesa

NB: credits for the Latex-format go to Iris Verweij, 2nd year MSc AI Student.

Problem 1: Consider a linear dynamical system model that has linear-Gaussian conditional distributions

$$\begin{aligned} p(\mathbf{z}_n|\mathbf{z}_{n-1}) &= N(\mathbf{z}_n|\mathbf{A}\mathbf{z}_{n-1}, \Gamma) \\ p(\mathbf{x}_n|\mathbf{z}_n) &= N(\mathbf{x}_n|\mathbf{C}\mathbf{z}_n, \boldsymbol{\Sigma}) \\ p(\mathbf{z}_1) &= N(\mathbf{z}_1|\boldsymbol{\mu}_0, \mathbf{V}_0) \end{aligned}$$

The log-likelihood of the data is given by

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \ln p(\mathbf{z}_1|\boldsymbol{\mu}_0, \mathbf{V}_0) + \sum_{n=1}^{N} \ln p(\mathbf{z}_n|\mathbf{z}_{n-1}, \mathbf{A}, \boldsymbol{\Gamma}) + \sum_{n=1}^{N} \ln p(\mathbf{x}_n|\mathbf{z}_n, \mathbf{C}, \boldsymbol{\Sigma})$$

In E step, we find $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \mathbb{E}_{\mathbf{z}|\boldsymbol{\theta}^{old}} [\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})]$. Your task is to perform M step.

1. Find $\mathbf{A}^{new}$ and $\boldsymbol{\Gamma}^{new}$ that optimize:

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = -\frac{N-1}{2} \ln |\boldsymbol{\Gamma}| - \mathop{\mathbb{E}}_{\mathbf{z}|\boldsymbol{\theta}^{old}} \left[\frac{1}{2} \sum_{n=2}^{N} (\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})^T \boldsymbol{\Gamma}^{-1} (\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1}) \right] + const$$

Solution: First, we derive equation to optimize $\mathbf{A}^{new}$:

$$\begin{aligned}
\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) &= -\frac{N-1}{2}\ln|\boldsymbol{\Gamma}| - \mathop{\mathbb{E}}_{\mathbf{z}|\boldsymbol{\theta}^{old}}\left[\frac{1}{2}\sum_{n=2}^{N}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})^T\boldsymbol{\Gamma}^{-1}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})\right] + const \\
&= -\frac{N-1}{2}\ln|\boldsymbol{\Gamma}| - \frac{1}{2}\sum_{n=2}^{N}\mathop{\mathbb{E}}_{\mathbf{z}|\boldsymbol{\theta}^{old}}\left[(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})^T\boldsymbol{\Gamma}^{-1}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})\right] + const \\
\frac{\partial\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})}{\partial\mathbf{A}} &= \frac{\partial\left[-\frac{N-1}{2}\ln|\boldsymbol{\Gamma}| - \frac{1}{2}\sum_{n=2}^{N}\mathbb{E}_{\mathbf{z}|\boldsymbol{\theta}^{old}}\left[(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})^T\boldsymbol{\Gamma}^{-1}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})\right] + const\right]}{\partial\mathbf{A}} \\
&= \{\text{using the dominated convergence theorem}\} \\
&= -\frac{1}{2}\sum_{n=2}^{N}\mathop{\mathbb{E}}_{\mathbf{z}|\boldsymbol{\theta}^{old}}\frac{\partial\left[(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})^T\boldsymbol{\Gamma}^{-1}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})\right]}{\partial\mathbf{A}} \\
&= \{\text{using equation (88) from Matrix Cookbook}\} \\
&= -\frac{1}{2}\sum_{n=2}^{N}\mathop{\mathbb{E}}_{\mathbf{z}|\boldsymbol{\theta}^{old}}\left[-2\boldsymbol{\Gamma}^{-1}(\mathbf{z}_n - \mathbf{A}\mathbf{z}_{n-1})\mathbf{z}_{n-1}^T\right] \\
&= \boldsymbol{\Gamma}^{-1}\sum_{n=2}^{N}\mathop{\mathbb{E}}_{\mathbf{z}|\boldsymbol{\theta}^{old}}\left[\mathbf{z}_n\mathbf{z}_{n-1}^T\right] - \mathbf{A}\,\mathbb{E}\left[\mathbf{z}_{n-1}\mathbf{z}_{n-1}^T\right] = 0
\end{aligned}$$

$$\sum_{n=2}^{N}\mathop{\mathbb{E}}_{\mathbf{z}|\boldsymbol{\theta}^{old}}\left[\mathbf{z}_n\mathbf{z}_{n-1}^T\right] - \mathbf{A}\,\mathbb{E}\left[\mathbf{z}_{n-1}\mathbf{z}_{n-1}^T\right] = 0$$

$$\sum_{n=2}^{N}\mathbf{A}\mathop{\mathbb{E}}_{\mathbf{z}|\boldsymbol{\theta}^{old}}\left[\mathbf{z}_{n-1}\mathbf{z}_{n-1}^T\right] = \sum_{n=2}^{N}\mathop{\mathbb{E}}_{\mathbf{z}|\boldsymbol{\theta}^{old}}\left[\mathbf{z}_n\mathbf{z}_{n-1}^T\right]$$

$$\mathbf{A} = \left(\sum_{n=2}^{N}\mathop{\mathbb{E}}_{\mathbf{z}|\boldsymbol{\theta}^{old}}\left[\mathbf{z}_n\mathbf{z}_{n-1}^T\right]\right)\left(\sum_{n=2}^{N}\mathop{\mathbb{E}}_{\mathbf{z}|\boldsymbol{\theta}^{old}}\left[\mathbf{z}_{n-1}\mathbf{z}_{n-1}^T\right]\right)^{-1}$$

Using similar way of reasoning, we can derive equation to optimize $\boldsymbol{\Gamma}^{new}$:

$$\begin{aligned}
\frac{\partial\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})}{\partial\boldsymbol{\Gamma}} &= \{\text{using equation (57) from Matrix Cookbook:}\} \\
&= -\frac{N-1}{2}\boldsymbol{\Gamma}^{-1} - \frac{1}{2}\sum_{n=2}^{N}\mathop{\mathbb{E}}_{\mathbf{z}|\boldsymbol{\theta}^{old}}\frac{\partial\left[(\mathbf{z}_n - \mathbf{A}^{new}\mathbf{z}_{n-1})^T\boldsymbol{\Gamma}^{-1}(\mathbf{z}_n - \mathbf{A}^{new}\mathbf{z}_{n-1})\right]}{\partial\boldsymbol{\Gamma}} \\
&= \{\text{using equation (61) from Matrix Cookbook and the fact that } \boldsymbol{\Gamma} \text{ is symmetric:}\} \\
&= -\frac{N-1}{2}\boldsymbol{\Gamma}^{-1} + \frac{1}{2}\sum_{n=2}^{N}\mathop{\mathbb{E}}_{\mathbf{z}|\boldsymbol{\theta}^{old}}\left[\boldsymbol{\Gamma}^{-1}(\mathbf{z}_n - \mathbf{A}^{new}\mathbf{z}_{n-1})(\mathbf{z}_n - \mathbf{A}^{new}\mathbf{z}_{n-1})^T\boldsymbol{\Gamma}^{-1}\right] = 0 \\
\boldsymbol{\Gamma}^{-1} &= \frac{1}{N-1}\boldsymbol{\Gamma}^{-1}\left[\sum_{n=2}^{N}\mathop{\mathbb{E}}_{\mathbf{z}|\boldsymbol{\theta}^{old}}\left[(\mathbf{z}_n - \mathbf{A}^{new}\mathbf{z}_{n-1})(\mathbf{z}_n - \mathbf{A}^{new}\mathbf{z}_{n-1})^T\right]\right]\boldsymbol{\Gamma}^{-1} \\
\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{-1} &= \frac{1}{N-1}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{-1}\left[\sum_{n=2}^{N}\mathop{\mathbb{E}}_{\mathbf{z}|\boldsymbol{\theta}^{old}}\left[(\mathbf{z}_n - \mathbf{A}^{new}\mathbf{z}_{n-1})(\mathbf{z}_n - \mathbf{A}^{new}\mathbf{z}_{n-1})^T\right]\right]\boldsymbol{\Gamma}^{-1} \\
\mathbf{I}\boldsymbol{\Gamma} &= \frac{1}{N-1}\mathbf{I}\left[\sum_{n=2}^{N}\mathop{\mathbb{E}}_{\mathbf{z}|\boldsymbol{\theta}^{old}}\left[(\mathbf{z}_n - \mathbf{A}^{new}\mathbf{z}_{n-1})(\mathbf{z}_n - \mathbf{A}^{new}\mathbf{z}_{n-1})^T\right]\right]\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma} \\
\boldsymbol{\Gamma} &= \frac{1}{N-1}\left[\sum_{n=2}^{N}\mathop{\mathbb{E}}_{\mathbf{z}|\boldsymbol{\theta}^{old}}\left[(\mathbf{z}_n - \mathbf{A}^{new}\mathbf{z}_{n-1})(\mathbf{z}_n - \mathbf{A}^{new}\mathbf{z}_{n-1})^T\right]\right]
\end{aligned}$$

2. Find $\mathbf{C}^{new}$ and $\mathbf{\Sigma}^{new}$ that optimize:

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = -\frac{N}{2} \ln |\mathbf{\Sigma}| - \mathop{\mathbb{E}}_{\mathbf{z}|\boldsymbol{\theta}^{old}} \left[\frac{1}{2} \sum_{n=1}^{N} (\mathbf{x}_n - \mathbf{C}\mathbf{z}_n)^T \mathbf{\Sigma}^{-1} (\mathbf{x}_n - \mathbf{C}\mathbf{z}_n) \right] + const$$

Solution: Using the previous part of the task, we can simply replace corresponding parameters in the final equation to get necessary result. This gives us the following equations:

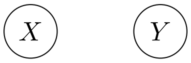
$$\mathbf{C}^{new} = \sum_{n=1}^{N} \left[\left(\mathbf{x}_n \, \mathbb{E}\left[\mathbf{z}_n{}^T \right] \right) \left(\mathbb{E}\left[\mathbf{z}_n \mathbf{z}_n{}^T \right] \right)^{-1} \right]$$

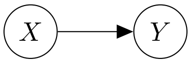
$$\mathbf{\Sigma}^{new} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}\left[(\mathbf{x}_n - \mathbf{C}^{new}\mathbf{z}_n)(\mathbf{x}_n - \mathbf{C}^{new}\mathbf{z}_n)^T \right]$$

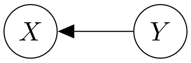
Problem 2: We consider the simplest nontrivial causal Bayesian networks: those with only two variables. Call the variables X and Y.

(a) Draw all dierent possible structures that these networks can have.

Solution:

1) 

2) 

3) 

(b) For each of the structures, write down the corresponding factorization of $p(X, Y)$.

Solution:

1) $p(X, Y) = p(X)p(Y)$

2) $p(X, Y) = p(Y|X)p(X)$

3) $p(X, Y) = p(X|Y)p(Y)$

(c) For each of the structures, write down an explicit expression for $p(Y|X)$ in terms of the factors in (b).

Solution:

1) $p(Y|X) = \frac{p(X,Y)}{p(X)} = \frac{p(X)p(Y)}{p(X)} = p(Y)$

2) $p(Y|X) = \frac{p(X)p(Y)}{p(X)} = \frac{p(Y|X)p(X)}{p(X)} = p(Y|X)$

3) $p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} = \frac{p(X|Y)p(Y)}{\int_Y p(X|Y)p(Y)dY}$

(d) For each of the structures, write down an explicit expression for $p(Y|do(X))$ in terms of the factors in (b).

Solution:

1)

$$\begin{aligned} p(Y|do(X)) &= \{\text{using action/observation exchange rule:}(Y \perp\!\!\!\perp X|\varnothing)_{\mathcal{G}_{\overline{\varnothing},\underline{Y}}}\} \\ &= p(Y|X) = p(Y) \end{aligned}$$

2)

$$\begin{aligned} p(Y|do(X)) &= \{\text{using action/observation exchange rule:}(Y \perp\!\!\!\perp X|\varnothing)_{\mathcal{G}_{\overline{\varnothing},\underline{Y}}}\} \\ &= p(Y|X) \end{aligned}$$

3)

$$\begin{aligned} p(Y|do(X)) &= \{\text{using ignoring actions rule:}(Y \perp\!\!\!\perp X|\varnothing)_{\mathcal{G}_{\overline{\varnothing},\overline{X(\varnothing)}}}\} \\ &= p(Y) \end{aligned}$$

(e) If Y means lung cancer, and X means smoking (supposing for simplicity that both are binary variables), describe in words what $p(Y|X)$ means and what $p(Y|do(X))$ means, and clearly indicate the difference in interpretation.

Solution: $p(Y|X)$ represents probability that patient has cancer if we observe that he smokes.
$p(Y|do(X))$ represents probability that patient has cancer if we force that person to smoke.

Problem 3: Suppose that you are investigating the eectiveness of a drug against a deadly disease. You have gathered the following data on patients that have been admitted in the hospital in which you work; some of these patients have been treated with the drug, others havent. Some of them recovered, others unfortunately didnt. The reasons why some patients were treated and others were not, are unknown to you.

1a. Calculate the recovery rates (in %) for both treatment and control (i.e., untreated) group.

Solution:

$$p(R|D=1) = \frac{20}{40} = 50\%$$
$$p(R|D=0) = \frac{16}{40} = 40\%$$

1b. Would you advice a new patient to take the drug, or not?

Solution: Yes, my advice would be to take the drug as percentage of recovered patients taking the drug is higher than of ones who were not taking the drug.

Upon closer inspection of the data, you notice something peculiar when you group patients according to gender.

2a. Calculate the recovery rates (in %) for both the treatment and the control groups, for both subpopulations (males and females).

Solution:

$$Males: p(R|D=1) = \frac{18}{30} = 60\% \qquad p(R|D=0) = \frac{7}{10} = 70\%$$
$$Females: p(R|D=1) = \frac{2}{10} = 20\% \qquad p(R|D=0) = \frac{9}{30} = 30\%$$

2b. Would you advice a male patient to take the drug, or not? And a female patient?

Solution: In general, it depends on a causal model that we use. However, assuming that we base our decision on the new data, no in both cases. The reason is that for both cases percentage of recovered patients not taking the drug is higher than of ones who were taking the drug.

2c. With hindsight, what would be your advice if the gender of the patient is unknown? Is this in contradiction with your earlier advice?

Solution: Before, the recommendation was to take the drug for a patient with any gender. Now, without knowing the gender, an advice would be to not take the drug. This indeed contradicts each other.

This phenomenon is known as Simpsons paradox. A lot has been written about this paradox, but it dissolves once you recognize that you should not make the mistake of interpreting correlations as causations. Indeed, whether or not you should prescribe the drug depends on which causal model you believe to apply to this situation. The fact that different causal models will lead to dierent conclusions is not paradoxical.

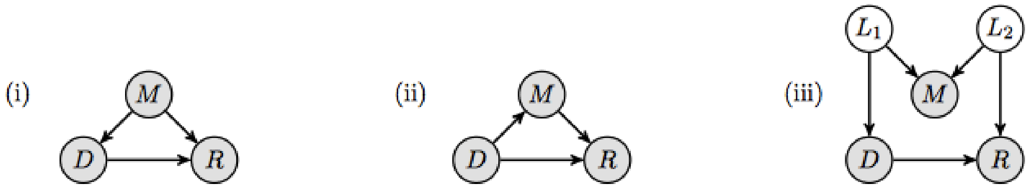


Figure 1: Different hypothetical causal models, where R stands for *Recovery*, D for taking the *Drug*, and M has different interpretations in cases (i), (ii) and (iii).

Suppose you believe that the causal model in Figure 1(i) applies, where M denotes gender of the patient (male/female).

3a. Apply Pearls back-door criterion to obtain a formula that expresses $p(R|do(D))$ in terms of observable quantities (i.e., in terms of marginal or conditional distributions where the do-operator does not appear).

Solution: As $S = \{M\}$ blocks the only backdoor path between D and R, it is admissible for an adjustment to find the causal effect of D on R. Therefore,

$$p(R|do(D = d)) = \int p(R|D, M)p(M)dM$$

3b. Is $p(R|do(D)) = p(R|D)$ in this case?

Solution: This equation would be correct in case $\varnothing$ would be admissible for an adjustment to find the causal effect of D on R. However, in this case it would not be correct. In this case: $p(R|D) = \int p(R|D, M)p(M|D)dM$.

3c. What would be your advice for a patient with unknown gender?

Solution: My advice would be to not take the drug. This is because we have seen from the previous tasks, namely item 2a., that if we use the assumption that the recovery rates are different based on the gender, taking a drug would lead to worse recovery rates than not taking it.

Now suppose that instead, you believe the causal model in Figure 1(ii) to apply. Intuitively, this would be quite unlikely, as we know that most drugs dont change gender, but we could have used a slightly dierent story where the variable M has a dierent interpretation (for example, blood pressure), and then this causal structure would also be a plausible one.

4a. Again, use Pearls back-door criterion to express $p(R|do(D))$ in terms of observable quantities.

Solution: In that case we don't have back-door path at all and therefore $\varnothing$ is admissible for an adjustment to find the causal effect of D on R. Therefore, $p(R|do(D)) = p(R|D)$.

4b. Is $p(R|do(D)) = p(R|D)$ in this case?

Solution: Yes.

4c. What would be your advice for a patient with unknown gender (or if you prefer, blood pressure) in this case?

Solution: My advice would be take a drug as in this case we would get recovery rates as calculated in item 1a. (first task of this exercise)

Finally, suppose that you believe that the causal model in Figure 1(iii) applies.

5a. Invent an interpretation of M and the two latent variables L_1, L_2 yourself that could match the causal model in Figure 1(iii).

Solution: L_1 - specific congenital disorder, which can't be diagnosed (not easily, at least) and is therefore unobservable. This contributes to the fact that patient might not be able to take a drug.
L_2 - presence of specific diseases among patient's ancestors. We don't have information about the ancestors and this variable is therefore unobservable. This contributes to the chances of recovery.
M - presence of some rare disease, i.e. rare form of influenza, in patient. Both L_1 and L_2 contribute to chances of this happening.

5b. Express $p(R|do(D))$ in terms of observable quantities.

Solution: In that case, As $S = \varnothing$ blocks the only backdoor path between D and R, it is admissible for an adjustment to find the causal effect of D on R. Therefore, $p(R|do(D)) = p(R|D)$.

5c. Is $p(R|do(D)) = p(R|D)$ in this case?

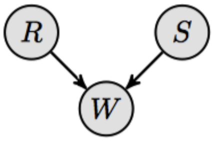
Solution: Yes.

5d. Again, what would be your advice for a patient with unknown gender in this case?

Solution: My advice would be take a drug as, again, in this case we would get recovery rates as calculated in item 1a. (first task of this exercise)

Problem 4: Consider three binary variables: R (cycled through the rain), S (taken a shower), W (wet hair). As R and S are *independent* causes of W, we consider the following SCM (where $\vee$ denotes logical or): with the following probability distribution for the exogenous variables:

$$\begin{aligned} R &= E_R \\ S &= E_S \\ W &= R \vee S \\ p(E_R, E_S) &= p(E_R)p(E_S) \end{aligned}$$



$$p(E_R = 1) = 0.7, \qquad p(E_S = 1) = 0.4$$

1. Calculate the induced distribution on the endogenous variables. How does it factorize?

Solution:

$$\begin{aligned} p(W, R, S) &= p(W|R, S)p(R)p(S) \\ p(R = 1) &= 0.7 \\ p(R = 0) &= 1 - p(R = 1) = 0.3 \\ p(S = 1) &= 0.4 \\ p(S = 0) &= 1 - p(S = 1) = 0.6 \\ p(W = 1) &= p(R = 1) + p(S = 1) - p(R = 1) \cdot p(S = 1) = 0.7 + 0.4 - 0.28 = 0.82 \\ p(W = 0) &= 1 - p(W = 1) = 0.18 \end{aligned}$$

2. What is the probability that it rained, given that your hair is wet?

Solution:

$$p(R = 1|W = 1) = \frac{p(W = 1|R = 1)p(R = 1)}{p(W = 1)} = \frac{1 \cdot 0.7}{0.82} \approx 0.85366$$

3. The probability of R increases when we observe W (in other words, R and W are correlated). Does this mean that W causes R?

Solution: No, because correlation doesn't imply causation. Also, observation of W updates our beliefs about R through conditioning.

We now would like to model what happens if we intervene on W by throwing a bucket of water over somebody.

4. Write down the intervened structural causal model for this intervention, $do(W = 1)$

Solution:

$$R = E_R; S = E_S; W = 1; P(E_R, E_S) = p(E_R)p(E_S) \implies E_R \perp\!\!\!\perp E_S$$

5. Calculate $p(R|do(W = w))$ and compare with $p(R)$ and $p(R|W = w)$.

Solution:

$$\begin{aligned}
p(R|do(W = w)) &= \{\text{using ignoring actions rule:}(R \perp\!\!\!\perp W|\varnothing)_{\mathcal{G}_{\overline{\varnothing},\overline{W(\varnothing)}}}\} \\
&= p(R) \\
p(R = 1) &= 0.7 \\
p(R = 0) &= 1 - p(R = 1) = 0.3 \\
p(R|W = w) &= \frac{p(W = w|R)p(R)}{p(W = w)} \\
p(R = 1|W = 1) &\approx 0.85366 \\
p(R = 1|W = 0) &= 0 \\
p(R = 0|W = 1) &= \frac{p(W = 1|R = 0)p(R = 0)}{p(W = 1)} = \frac{0.4 \cdot 0.3}{0.82} \approx 0.14643 \\
p(R = 0|W = 0) &= \frac{p(W = 0|R = 0)p(R = 0)}{p(W = 0)} = \frac{0.6 \cdot 0.3}{0.18} = 1
\end{aligned}$$

As you can see, $p(R|do(W = w))$ is generally equal to $p(R)$ and is generally different from $p(R|W = w)$.

Instead of intervening on W, we now consider an intervention on S.

6. Write down the intervened structural causal model for the intervention $do(S = s)$.

Solution:

$$R = E_R; S = s; W = R \vee S; P(E_R, E_S) = p(E_R)p(E_S) \implies E_R \perp\!\!\!\perp E_S$$

7. Calculate $p(W|do(S=s))$ and compare with $p(W|S=s)$ and $p(W)$.

Solution:

$$\begin{aligned}
p(W|do(S=s)) &= \{\text{using action/observation exchange rule:}(W \perp\!\!\!\perp S|\varnothing)_{\mathcal{G}_{\overline{\varnothing},\underline{S}}}\} \\
&= p(W|S=s)) \\
p(W=1|S=1) &= 1 \\
p(W=1|S=0) &= 0.7 \\
p(W=0|S=1) &= 0 \\
p(W=0|S=0) &= 0.3 \\
p(W) &= p(W|do(S=s))p(S=s) \\
p(W=1) &= 0.82 \\
p(W=0) &= 0.18
\end{aligned}$$

As you can see, $p(W|do(S=s))$ is generally equal to $p(W|S=s)$ and is generally different from $p(W)$.

Problem 5: Given a Markovian Structural Causal Model, or alternatively, a causal Bayesian network. Let X be an observed variable, $\mathbf{X}_{pa(X)}$ denote all observed parents of X, and let Y be another observed variable (i.e., neither X nor one of the parents of X).

1. By applying the truncated factorization theorem, show that

$$p(Y|do(X), \mathbf{X}_{pa(X)}) = p(Y|X, \mathbf{X}_{pa(X)})$$

Solution:

$$\begin{aligned}
p(Y|do(X), \mathbf{X}_{pa(X)}) &= \frac{p(Y, \mathbf{X}_{pa(X)}|do(X))}{p(\mathbf{X}_{pa(X)}|do(X))} \\
&= \frac{\frac{p(Y,X,\mathbf{X}_{pa(X)})}{\cancel{p(X|\mathbf{X}_{pa(X)})}}}{\frac{p(X,\mathbf{X}_{pa(X)})}{\cancel{p(X|\mathbf{X}_{pa(X)})}}} \\
&= \frac{p(Y, X, \mathbf{X}_{pa(X)})}{p(X, \mathbf{X}_{pa(X)})} \\
&= p(Y|X, \mathbf{X}_{pa(X)}
\end{aligned}$$

2. Use that result and show by marginalizing over $\mathbf{X}_{pa(X)}$ that

$$p(Y|do(X)) = \int p(Y|X, \mathbf{X}_{pa(X)})p(\mathbf{X}_{pa(X)})d\mathbf{X}_{pa(X)}.$$

Solution: Using that $p(\mathbf{X}_{pa(X)}|do(X)) = p(\mathbf{X}_{pa(X)})$:

$$\begin{aligned}
p(Y|do(X)) &= \int p(Y|do(X), \mathbf{X}_{pa(X)}) p(\mathbf{X}_{pa(X)}|do(X)) d\mathbf{X}_{pa(X)} \\
&= \int p(Y|X, \mathbf{X}_{pa(X)}) p(\mathbf{X}_{pa(X)}) d\mathbf{X}_{pa(X)}
\end{aligned}$$