IE 598 Inference in Graphical Models

Solution 1

Problem 1.1 (Exercise 2.5 in Koller/Friedman)

Let X, Y, Z be three disjoint subsets of random variables. We say X and Y are conditionally independent given Z if and only if

$$\mathbb{P}_{X,Y|Z}(x,y|z) \quad = \quad \mathbb{P}_{X|Z}(x|z)\,\mathbb{P}_{Y|Z}(y|z)\ .$$

Show that X and Y are conditionally independent given Z if and only if the joint distribution for the three subsets of random variables factors in the following form:

$$\mathbb{P}_{X,Y,Z}(x,y,z) \quad = \quad h(x,z)\,g(y,z)\ .$$

Solution 1.1

First, we show that conditional independence implies the desired factorization.

$$\begin{aligned}
\mathbb{P}_{X,Y,Z}(x,y,z) &= \mathbb{P}_Z(z)\,\mathbb{P}_{X,Y|Z}(x,y|z) \\
&= \mathbb{P}_Z(z)\,\mathbb{P}_{X|Z}(x|z)\,\mathbb{P}_{Y|Z}(y|z) \\
&= h(x,z)\,g(y,z)\ ,
\end{aligned}$$

where we choose $h(x,z) = \mathbb{P}_{X|Z}(x|z)$ and $g(y,z) = \mathbb{P}_{Y|Z}(y|z)\mathbb{P}_Z(z)$.

Now, we show the other direction. For $\mathbb{P}_{X,Y,Z}(x,y,z) = h(x,z)\,g(y,z)$, let $h_1(z) = \sum_x h(x,z)$ and $g_1(z) = \sum_y g(y,z)$. We need to show that for such $\mathbb{P}_{X,Y,Z}(x,y,z)$, we have $\mathbb{P}_{X,Y|Z}(z,y|z) = \mathbb{P}_{X|Z}(x|z)\mathbb{P}_{Y|Z}(y|z)$. We first compute

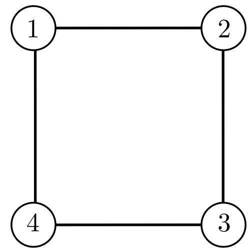
$$\begin{aligned}
\mathbb{P}_{X,Y|Z}(z,y|z) &= \frac{\mathbb{P}_{X,Y,Z}(x,y,z)}{\sum_{x,y}\mathbb{P}_{X,Y,Z}(x,y,z)} \\
&= \frac{h(x,z)g(y,z)}{\sum_{x,y}h(x,z)g(y,z)} \\
&= \frac{h(x,z)g(y,z)}{h_1(z)g_1(z)}\ .
\end{aligned}$$

Similarly, we can compute $\mathbb{P}_{X|Z}(x|z) = h(x,z)/h_1(z)$ and $\mathbb{P}_{Y|Z}(y|z) = g(y,z)/g_1(z)$. This proves the conditional independence, since $\mathbb{P}_{X,Y|Z}(z,y|z) = \mathbb{P}_{X|Z}(x|z)\mathbb{P}_{Y|Z}(y|z)$.

Problem 1.2 (Exercise 4.1 in Koller/Friedman)

In this problem, we will show by example that the distribution of a graphical model need not have a factorization of the form in the Hammersley-Clifford Theorem if the distribution is not strictly positive. In particular, we will consider a distribution on the following simple 4-cycle where each node is a binary random variable, X_i, for $i \in \{1,2,3,4\}$. Consider a probability distribution that assigns a probability 1/8 uniformly to each of the following set of values (X_1, X_2, X_3, X_4):

$$\begin{array}{cccc}
(0,0,0,0) & (1,0,0,0) & (1,1,0,0) & (1,1,1,0) \\
(0,0,0,1) & (0,0,1,1) & (0,1,1,1) & (1,1,1,1)
\end{array}$$



and assigns zero to all other configurations of (X_1, X_2, X_3, X_4).

(a) We first need to show that this distribution is Markov on our graph. To do this, it should not be difficult to see that what we need to show are the following conditions:

- The pair of variables X_1 and X_3 are conditionally independent given (X_2, X_4).
- The pair of variables X_2 and X_4 are conditionally independent given (X_1, X_3).

First, show that if we interchange X_1 and X_4 and interchange X_2 and X_3, we obtain the same distribution, i.e.., $\mathbb{P}(x_1, x_2, x_3, x_4) = \mathbb{P}(x_4, x_3, x_2, x_1)$. This implies that if we can show the first condition, then the other is also true.

(b) Show that whatever pair of values you choose for (X_2, X_4), we then know either X_1 or X_3 with certainty. For example, $(X_2 = 0, X_4 = 0)$ implies that $X_3 = 0$. Since we know either X_1 or X_3 with certainty, then conditioning on the other one of these obviously provides no additional information, trivially proving conditional independence.

(c) What we now need to show is that the distribution cannot be factorized in the way stated in the Hammersley-Clifford Theorem. We will do this by contradiction. Noting that the maximal cliques in our graph are just the edges and absorbing the normalization $1/Z$ into any of the pairwise compatibility functions, we know that if our distribution has the factorization implied by the Hammersley-Clifford Theorem, we can write it in the following form:

$$\mathbb{P}(x_1, x_2, x_3, x_4) \quad = \quad \psi_{12}(x_1, x_2)\, \psi_{23}(x_2, x_3)\, \psi_{34}(x_3, x_4)\, \psi_{41}(x_4, x_1)\ .$$

Show that assuming that our distribution has such a factorization leads to a contradiction by examining the values of $\mathbb{P}(0, 0, 0, 0)$, $\mathbb{P}(0, 0, 1, 0)$, $\mathbb{P}(0, 0, 1, 1)$, and $\mathbb{P}(1, 1, 1, 0)$.

Solution 1.2

(a) We don't need to check $\mathbb{P}(x_1, x_2, x_3, x_4) = \mathbb{P}(x_4, x_3, x_2, x_1)$ for configurations where $x_1 = x_4$ and $x_2 = x_3$. For all others, we have

$$\begin{aligned}
\mathbb{P}(0,0,0,1) = \mathbb{P}(1,0,0,0) = \frac{1}{8} \\
\mathbb{P}(0,0,1,1) = \mathbb{P}(1,1,0,0) = \frac{1}{8} \\
\mathbb{P}(0,1,0,1) = \mathbb{P}(1,0,1,0) = 0 \\
\mathbb{P}(0,1,1,1) = \mathbb{P}(1,1,1,0) = \frac{1}{8} \\
\mathbb{P}(0,0,1,0) = \mathbb{P}(0,1,0,0) = 0 \\
\mathbb{P}(0,1,0,0) = \mathbb{P}(0,0,1,0) = 0 \\
\mathbb{P}(1,0,1,1) = \mathbb{P}(1,1,0,1) = 0 \\
\mathbb{P}(1,1,0,1) = \mathbb{P}(1,0,1,1) = 0
\end{aligned}$$

(b) If $(X_2, X_4) = (0,0)$ then $X_3 = 0$. If $(X_2, X_4) = (0,1)$ then $X_1 = 0$. If $(X_2, X_4) = (1,0)$ then $X_1 = 1$. If $(X_2, X_4) = (1,1)$ then $X_3 = 1$.

(c) The following equations are true:

$$\begin{aligned}
\psi_{12}(0,0)\,\psi_{23}(0,1)\,\psi_{34}(1,0)\,\psi_{41}(0,0) &= 0\\
\psi_{12}(0,0)\,\psi_{23}(0,0)\,\psi_{34}(0,0)\,\psi_{41}(0,0) &= 1/8\\
\psi_{12}(0,0)\,\psi_{23}(0,1)\,\psi_{34}(1,1)\,\psi_{41}(1,0) &= 1/8
\end{aligned}$$

Then, it follows that $\psi_{34}(1,0)$ must be zero. This is a contradiction, since we know that

$$\psi_{12}(1,1)\,\psi_{23}(1,1)\,\psi_{34}(1,0)\,\psi_{41}(0,1) = 1/8\ .$$

Problem 1.3

Given a graph $G = (V, E)$, an *independent set* of G is a subset $S \subseteq V$ of the vertices, such that no two vertices in S is connected by an edge in E. Precisely, if $i, j \in S$ then $(i, j) \notin E$. We let $\mathrm{IS}(G)$ denote the set of all independent sets of G, and let $Z(G) = |\mathrm{IS}(G)|$ denote its size, i.e. the total number of independent sets in G. The number of independent sets $Z(G)$ is at least $1 + |V|$, since the empty set and all subsets with single vertex are always independent sets. We are interested in the uniform probability measure over S:

$$\mathbb{P}_{\mathrm{IS}(G)}(S) = \frac{1}{Z(G)}\mathbb{I}(S \in \mathrm{IS}(G))\ ,$$

where $\mathbb{I}(A)$ is an indicator function which is one if event A is true and zero if false.

(a) The set S can be naturally encoded by a binary vector $x \in \{0,1\}^{|V|}$ by letting $x_i = 1$ if and only if $i \in S$. Denote by $\mathbb{P}_G(x)$ the probability distribution induced on this vector x according to $\mathbb{P}_{\mathrm{IS}(G)}(S)$. Show that $\mathbb{P}_G(x)$ is a pairwise graphical model on G.
[Hint: A pairwise graphical model on a graph $G = (V, E)$ is defined by a factorization of the form $\mathbb{P}_G(x) = (1/Z)\prod_{(i,j)\in E}\psi_{i,j}(x_i, x_j)$.]

(b) Let L_n be the line graph with n vertices, i.e. the graph with vertex set $V(L_n) = \{1, 2, 3, \ldots, n\}$ and edge set $E(L_n) = \{(1,2), (2,3), \ldots, (n-1, n)\}$. Derive a formula for $Z(L_n)$.
[Hint: Write a recursion over n, and solve it using a matrix representation.]

(c) With the above definitions, derive a formula for $\mathbb{P}_{L_n}(x_i = 1)$, for each $i \in \{1, \ldots, n\}$. Plot $\mathbb{P}_{L_n}(x_i)$ versus i for $n = 11$. Describe the main features of this plot. Can you give an intuitive explanation?
[Hint: Use the recursion from previous subproblem.]

(d) The same probability distribution $\mathbb{P}_{L_n}(x)$ can be also represented with a Bayesian network. Using the recursions used in (b) and (c), write the conditional probability distributions for this Bayesian network.

Solution 1.3

(a) Define the compatibility functions to be

$$\psi_{i,j}(x_i, x_j) = \mathbb{I}((x_i, x_j) \neq (1,1))$$

Then, with this definition the joint distribution factorizes in the following form:

$$\mathbb{P}_G(x) = \frac{1}{Z} \prod_{(i,j)\in E} \psi_{i,j}(x_i, x_j)$$

since the product of $\psi_{i,j}$'s yield the indicator $\mathbb{I}(S \in \mathrm{IS}(G))$ for the subset S encoded by x. Hence, $\mathbb{P}_G(x)$ is a pairwise graphical model.

(b) We know that $X(L_n)$ is the number of independent sets in the graph L_n. Let $Z(L_n) = A_n + B_n$ where A_n denotes the number of independent sets in L_n containing the vertex n, and B_n denotes the number of independent sets in L_n excluding the vertex n. We can write the following recursion:

$$\begin{aligned} A_n &= B_{n-1} \\ B_n &= A_{n-1} + B_{n-1} \end{aligned}$$

The first recurrence follows from the fact that if $S \subseteq [n]$ containing n is an independent set of L_n, then $S \setminus \{n\}$ is an independent set of L_{n-1}. The second, similarly, is because an independent set of L_n not containing vertex n is basically an independent set of L_{n-1}. Defining $X_n = [A_n \; B_n]^T$, we can write the recurrence relation as:

$$\begin{aligned} X_n &= PX_{n-1} \\ \text{where } P &= \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \\ \text{and } X_1 &= \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{aligned}$$

This yields: $X_n = P^{n-1}X_1$. As $Z(L_n) = [1 \;\; 1]^T X_n$, diagonalizaing P yields the following closed form solution:

$$Z(L_n) = \Big(1 + \frac{2}{\sqrt{5}}\Big)\Big(\frac{1+\sqrt{5}}{2}\Big)^{n-1} + \Big(1 - \frac{2}{\sqrt{5}}\Big)\Big(\frac{1-\sqrt{5}}{2}\Big)^{n-1}$$

(c) For $i \in \{1, n\}$, i.e. i being an end vertex, the number of independent sets containing i is simply $Z(L_{n-2})$. If i is an intermediate vertex, then an independent set containing i is formed by choosing an independent set from $[i-2]$ and an independent set from $[n] \setminus [i+1]$. Thus we obtain the marginal as:

$$\mathbb{P}_{L_n}(x_i = 1) = \begin{cases} \frac{Z(L_{n-2})}{Z(L_n)} & \text{if } i \in \{1, n\} \\ \frac{Z(L_{i-2})Z(L_{n-i-1})}{Z(L_n)} & \text{otherwise} \end{cases}$$

The following MATLAB code prodces the required plots:

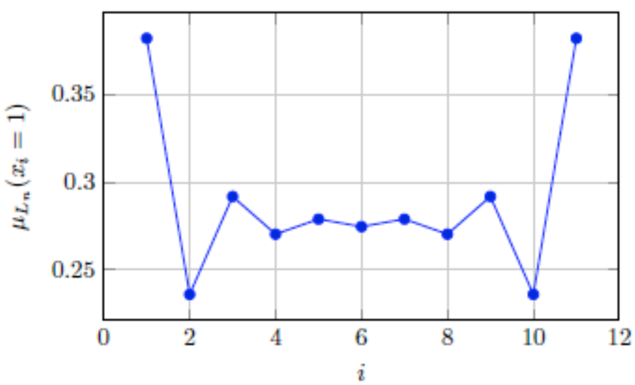
```
n = 1 1 ;
nrange = 0 : n ;
c1 = 1+2/sqrt ( 5 ) ;
c2 = 1-2/sqrt ( 5 ) ;
r1 = (1+sqrt ( 5 ) ) / 2 ;
r2 = (1-sqrt ( 5 ) ) / 2 ;
%z (1) . . . z (12) contains Z0 to Z11
z = c1∗ r1 . ∧ ( nrange-1) + c2∗ r2 . ∧ ( n range -1);

% compute marginal s mu
mu = zeros ( 1 , n ) ;
```

```
mu( 1 ) = z (end-2)/z (end ) ;
mu(n) = mu( 1 ) ;
for i = 2 : ( n-1)
mu( i ) = z ( i -1) z (n-i )/ z (n+1);
end

plot ( 1 : n , mu)
```

The plot is as follows:



The exponent in the numerator is constant for $i = 2, \ldots, n-1$, hence we see a relatively at marginal curve in this region. The marginal increases at either end, since the end vertices impose fewer restrictions on the inclusion of other vertices in the independent set.

(d) By the law of conditional probability, we have:

$$\mathbb{P}_{L_n}(x) = \mathbb{P}_{L_n}(x_1)\mathbb{P}_{L_n}(x_2|x_!) \cdots \mathbb{P}_{L_n}(x_n|x_1 \cdots x_{n-1})$$

Since the inclusion of vertex i is dependent only on its neighbors, we have $\mathbb{P}_{L_n}(x_i|x_1, \cdots, x_{i-1}) = \mathbb{P}_{L_n}(x_i|x_{i-1})$. This is equivalent to creating a Bayesian network by directing all the edges in L_n towards the larger index, i.e. letting the parent $\pi(k)$ of a vertex k be $k-1$, $k = 2, \cdots, n$. Using similar arguments as before, we have that:

$$\mathbb{P}_{L_n}(x_i = 1|x_{i-1}) = \begin{cases} 0 & \text{if } x_{i-1} = 1 \\ \frac{Z(L_{n-i-1})}{Z(L_{n-i+1})} & \text{otherwise} \end{cases}$$

Problem 1.4

We again consider the independent set explained in the previous problem. Now let $G = T_{k,\ell}$ denote the rooted tree with branching factor k and ℓ generations, that is the root has k descendants and each other node has one ancestor and k descendants except for the leaves. The total number of vertices is $(k^{\ell+1} - 1)/(k-1)$, and $T_{k,\ell=0}$ is the graph consisting only of the root. We let ϕ denote the root of $T_{k,\ell}$.

(a) Let $Z_\ell = Z(T_{k,\ell})$ denote the total number of independent sets of $G = T_{k,\ell}$. Let $Z_\ell(0)$ be the number of independent sets in $T_{k,\ell}$ such that the root is $x_\phi = 0$, and $Z_\ell(1)$ be the number of independent sets such that $x_\phi = 1$. It is immediate that $Z_0(0) = Z_0(1) = 1$. Derive a recursion expressing $(Z_{\ell+1}(0), Z_{\ell+1}(1))$ as a function of $(Z_\ell(0), Z_\ell(1))$.

(b) Using the above recursion, derive a recursion for the probability that the root belongs to a uniformly random independent set. Explicitly, derive a recursion for

$$p_\ell \quad = \quad \mathbb{P}_{T_{k,\ell}}(\{x_\phi = 1\}) \; .$$

(c) Program this recursion and plot p_ℓ as a function of $\ell \in \{0, 1, \dots, 50\}$ for four values of k, e.g. $k \in \{1, 2, 3, 10\}$. Comment on the qualitative behavior of these plots.

(d) Prove that, for $k \leq 3$, the recursion converges to a unique value using Banach's fixed point theorem.

Solution 1.4

(a) As before, we assume that an empty set is an independent set, by definition. An independent set not containing the root ϕ is formed by choosing an independent set from each subtree rooted at one of the children of ϕ. Also, an independent set containing the root ϕ cannot have any of the children of ϕ and thus is formed of in addition to independent sets not containing the root in the subtrees of the children of . This yields the following recursion equations:

$$\begin{aligned} Z_{\ell+1}(0) &= (Z_\ell(0) + Z_\ell(1))^k \\ Z_{\ell+1}(1) &= Z_\ell(0)^k \\ Z_0(0) &= Z_1 \;=\; 1 \end{aligned}$$

(b) We have the following immediately:

$$\begin{aligned} p_{\ell+1} &= \frac{Z_{\ell+1}(1)}{Z_{\ell+1}(0) + Z_{\ell+1}(1)} \\ &= \frac{Z_\ell(0)^k}{(Z_\ell(0) + Z_\ell(1))^k + Z_\ell(0)^k} \\ &= \frac{1}{1 + \left(\frac{1}{1-p_\ell}\right)^k} \end{aligned}$$

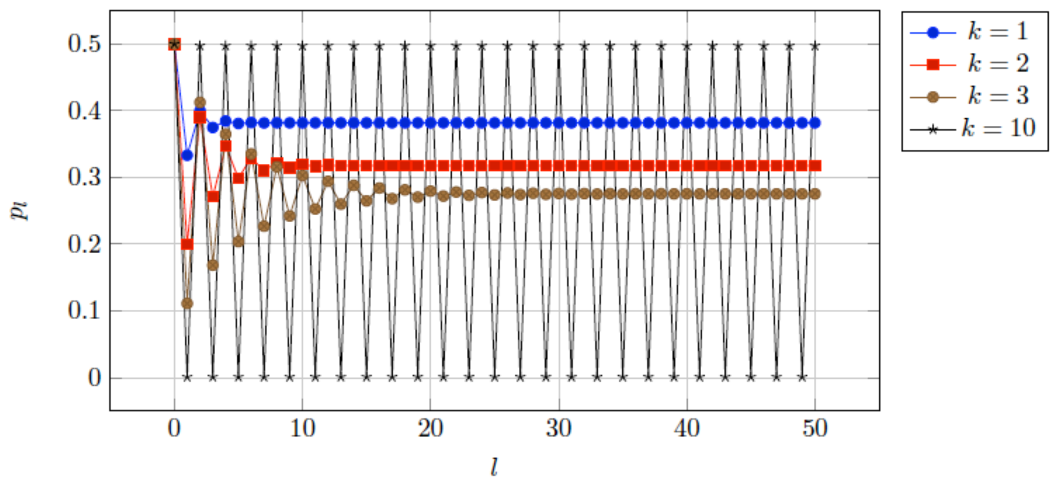
(c) The following code plots pl for the relevant values of k and ℓ:

```
kvals = [1 2 3 10];
iters = length ( kvals ) ;
lvals = 1:50;
p = zeros ( iters , 1+length ( lvals ) ) ;
p ( : , 1 ) = 0 . 5 ∗ ones ( iters , 1 ) ; % initialization
spec = { 'b' 'g' 'r' 'k' } ;

figure (1)
hold on
```

```
for i = 1 : iters
        k = kvals ( i );
        for l = 1 : length ( lvals )
                p ( i , l +1) = (1p ( i , l ) )∧k/(1+(1p ( i , l ) )∧k ) ;
        end
        plot ( [ 0 lvals ] , p ( i , : ) , spec { i } ) ;
end
hold off
legend ( '1' , '2' , '3' , '10' ) ;
```

The plot is as follows:



The recursion converges to a fixed point for $k \leq 3$ but fails to (or appears to fail to) converge for $k > 3$.

(d) **Proof of convergence for** $k \leq 3$

Let $f_k : [0, 1] \to [0, 1]$ be the mapping (as in the recursion) parametrized by k:

$$f_k(x) \quad = \quad \frac{(1-x)^k}{1+(1-x)^k}$$

We use Banachs fixed point theorem to prove convergence.
Theorem 1 (Banach). *Let X be a complete metric space and $f : X \to X$ be a contraction mapping. Then f has a unique fixed point $x*$. Also the sequence $\{x_i\}_{i=1}^{\infty}$ generated by $x_i = f(x_{i-1})$ converges to $x*$.*

Definition 1. *Let X be a metric space and $d(\cdot, \cdot)$ be the associated metric. $f : X \to X$ is a contraction mapping with parameter β on X if $\exists 0 \leq \beta < 1$ such that:*

$$\forall x_1, x_2 \in X : \quad d(f(x_1), f(x_2)) \leq \beta d(x_1, x_2)$$

In our case, the space is the interval $[0, 1]$ with the associated metric being the absolute value of the difference: $d(x_1, x_2) = |x_1 - x_2|$.

We know that a fixed point exists as the mappings f_k are continuous, decreasing and map onto $[0, 0.5]$ for all k. However, this does not guarantee that the recursion converges. To use the fixed point theorem, we prove that for $k = 1, 2, 3$, f_k are contraction mappings. For this we use the following lemma:

Lemma 1. *Let $f : [1, b] \to [1, b]$ be a differentiable function such that $|f'(x)|$ is bounded uniformly by $\beta < 1$ in its domain. Then f is a contraction mapping with parameter β, the distance metric being the absolute value of the difference.*

Proof. Consider $a \leq x_1 < x_2 \leq b$. By the intermediate value theorem, $\exists c \in [x_1, x_2]$ such that $f(x_2) - f(x_1) = f'(c)(x_2 - x_1)$. Thus $|f(x_2) - f(x_1)| = |f'(c)|\,|x_2 - x_1| \leq \beta|x_2 - x_1|$. $\square$

A little calculus shows that the maximum value of $|f_k'(x)|$ occurs at $x = 1 - \left(\frac{k-1}{k+1}\right)^{1/k}$, whereby we get:

$$|f_k'(x)| \quad \leq \quad \frac{(k+1)^2}{4k}\left(\frac{k-1}{k+1}\right)^{\frac{k-1}{k}}$$

For $k = 2, 3$ this yields that f_k is indeed a contraction map by Lemma 1. Thus, by the fixed point theorem, the recursion converges to its unique fixed point. For $k = 1$, we cannot use this directly as the maximum is at $x = 0$ and $f_1(0) = -1$. However this can be remedied by restricting the domain of f_1 to $[\epsilon, 1]$ for some small $\epsilon > 0$ whereupon it becomes a contraction map on the restricted domain since $|f_1'(x) \leq \frac{1}{(1+\epsilon)^2}$.

Non-convergence for $k > 4$

One argument for the non-convergence of the recursion for larger k is the following condition: there must exist a neighborhood around the fixed point $x*$ in which $|f_k(x)| < 1$ holds. This is because the linearization of f_k around its fixed point must be a stable linear system, i.e. have eigenvalues within the unit circle. For values of $k > 4$, this condition fails to hold.

Problem 1.5 (Intersection lemma) In proving that pairwise Markov property implies global Markov property for undirected graphical models, we used the intersection lemma which states that if μ is strictly positive and

$$A\text{–}(C \cup D)\text{–}B, \qquad A\text{–}(B \cup D)\text{–}C,$$

then

$$A\text{–}D\text{–}(B \cup C)\,.$$

Here A–B–C if and only if $\mu(x_A, x_C|x_B) = \mu(x_A|x_B)\,\mu(x_C|x_B)$. From previous homework, we know that A–$(C \cup D)$–B if and only if $\mu(x_A, x_B, x_C, x_D) = a(x_A, x_C, x_D)\,b(x_B, x_C, x_D)$ for some function $a(\cdot)$ and $b(\cdot)$. Similarly, we have $\mu(x_A, x_B, x_C, x_D) = f(x_A, x_B, x_D)\,g(x_B, x_C, x_D)$.

(*a*) Show $f(x_A, x_B, x_D) = a'(x_A, x_D)\,b'(x_B, x_D)$ for some $a'(\cdot)$ and $b'(\cdot)$ and find one such pair of functions a' and b' in terms of $a(), b(), g()$.

(*b*) Substitute $f(\cdot)$ and prove A–D–$(B \cup C)$.

(*c*) Find a counter example when μ is not strictly positive.

Solution 1.5

(a) since $\mu = a{\cdot}b = a'{\cdot}b'{\cdot}g$, we can set $a'(x_A, x_D) = a(x_A, x_C^*, x_D)$ and $b'(x_B, x_D) = b(x_B, x_C^*, x_D)/g(x_B, x_C^*, x_D)$, for any values of x_C^* of choice. Note this is only well-defined for positive $g(\cdot)$, i.e. $g(x_B, x_C^*, x_D) > 0$ for all x_A and x_D.

(b) Substituting $f(\cdot)$ we get

$$\mu(x_A, x_B, x_C, x_D) \quad = \quad a'(x_A, x_D)\,\underbrace{b'(x_B, x_D)\,g(x_B, x_C, x_D)}_{\text{a function of } x_B, x_C, x_D}$$

which proves A–D–$(B \cup C)$.

(c) a counter example:

ignore x_D (make it deterministic for instance) and let

$$x_A = \begin{cases} 1 & \text{w.p. } 1/2 \\ -1 & \text{w.p. } 1/2 \end{cases}$$

$$x_B = \begin{cases} x_A & \text{w.p. } 1/2 \\ 2x_A & \text{w.p. } 1/2 \end{cases}$$

$$x_C = \begin{cases} x_A & \text{w.p. } 1/2 \\ 2x_A & \text{w.p. } 1/2 \end{cases}$$

given x_B we know x_A, so conditional $x_A \perp x_B | x_C$ and also $x_A \perp x_C | x_B$, but it is clear that $x_A \not\perp (x_B, x_C)$.