UC Berkeley
Department of Electrical Engineering and Computer Science
Department of Statistics

EECS 281A / STAT 241A Statistical Learning Theory

**Solution to Problem Set 3**
Fall 2012

**Issued:** Tues. Sep. 18, 2012 **Due:** Tues. Oct 2, 2012

---

**Reading:** Chapters 2, 3 and 4

---

**Problem 3.1**
Suppose that three discrete random variables $(X, Y, Z)$ have a joint PMF such that $p(x, y, z) > 0$ for all $(x, y, z)$. Show that if $X \perp Y \mid Z$ and $X \perp Z \mid Y$, then we have $X \perp (Y, Z)$. Is this still true if we allow $p(x, y, z) = 0$ for some $(x, y, z)$?
**Solution:** For any $x, y, z$, note that from $X \perp Y \mid Z$ we have $p(x \mid y, z) = p(x \mid z)$, and from $X \perp Z \mid Y$ we have $p(x \mid y, z) = p(x \mid y)$. Thus $p(x \mid z) = p(x \mid y)$, and therefore,

$$p(y)\, p(x, z) = p(y)\, p(z)\, p(x \mid z) = p(y)\, p(z)\, p(x \mid y) = p(z)\, p(x, y).$$

Summing over $y$ gives us $p(x, z) = p(x)\, p(z)$, so $X \perp Z$, and summing over $z$ gives us $p(x, y) = p(x)\, p(y)$, so $X \perp Y$. Therefore, using the relation $p(y \mid x) = p(y)$ (from $X \perp Y$) and $p(z \mid x, y) = p(z \mid y)$ (from $X \perp Z \mid Y$), we obtain

$$p(x, y, z) = p(x)\, p(y \mid x)\, p(z \mid x, y) = p(x)\, p(y)\, p(z \mid y) = p(x)\, p(y, z).$$

Hence we conclude that $X \perp (Y, Z)$.

Counterexample for when $p(x, y, z) = 0$ for some $x, y, z$ (courtesy of Ka Kit Lam): Let $\theta \sim \text{Bernoulli}(\frac{1}{2})$. On the event $\theta = 0$, let $X, Y, Z$ be i.i.d. uniform random variables taking values in $\{0, 1\}$, and on the event $\theta = 1$, let $X, Y, Z$ be i.i.d. uniform random variables taking values in $\{2, 3\}$. Then $X \perp Y \mid Z$ because conditioning on $Z$ tells us the value of $\theta$, which makes $X$ and $Y$ independent. Similarly, we also have $X \perp Z \mid Y$. However, note that

$$p(X = 0, Y = 2, Z = 2) = 0 \neq \frac{1}{4} \cdot \frac{1}{8} = p(X = 0)\, p(Y = 2, Z = 2),$$

which shows that $X$ is *not* independent to $(Y, Z)$.

**Problem 3.2**
For each of the following statements, either give a proof of its correctness, or a counterexample to show incorrectness.

(a) If $X_1 \perp X_2$, then $X_1 \perp X_2 \mid X_3$.

**Solution:** False. Let $X_1, X_2 \sim \text{Bernoulli}(\frac{1}{2})$ i.i.d. and $X_3 = X_1 + X_2 \pmod 2$. Then $p(X_1 = x_1 \mid X_3 = x_3) = p(X_2 = x_2 \mid X_3 = x_3) = 1/2$ for all $x_1, x_2, x_3 \in \{0, 1\}$, but $p(X_1 = x_1, X_2 = x_2 \mid X_3 = 0) = 0$ if $x_1 \neq x_2$. This shows that $X_1$ is not independent of $X_2$ given $X_3$.

(b) If $X_1 \perp X_2 \mid X_4$ and $X_1 \perp X_3 \mid X_4$, then $X_1 \perp (X_2, X_3) \mid X_4$.

**Solution:** False. Let $X_2, X_3 \sim \text{Bernoulli}(\frac{1}{2})$ i.i.d., let $X_1 = X_2 + X_3 \pmod 2$, and let $X_4$ be independent of $X_1, X_2, X_3$. Then $X_1 \perp X_2$ and $X_1 \perp X_3$, but $X_1$ is completely determined by observing the pair $(X_2, X_3)$.

(c) If $X_1 \perp (X_2, X_3) \mid X_4$, then $X_1 \perp X_2 \mid X_4$.

**Solution:** True. We have

$$p(x_2 \mid x_1, x_4) = \sum_{x_3} p(x_2, x_3 \mid x_1, x_4) = \sum_{x_3} p(x_2, x_3 \mid x_4) = p(x_2 \mid x_4),$$

where the second equality uses the assumption $X_1 \perp (X_2, X_3) \mid X_4$.

**Problem 3.3**

*Graphs and independence relations:* For $i = 1, 2, 3$, let $X_i$ be an indicator variable for the event that a coin toss comes up heads (which occurs with probability $q$). Supposing that that the $X_i$ are independent, define $Z_4 = X_1 \oplus X_2$ and $Z_5 = X_2 \oplus X_3$ where $\oplus$ denotes addition in modulo two arithmetic.

(a) Compute the conditional distributions of $(X_2, X_3)$ given $Z_5 = 0$ and $Z_5 = 1$ respectively.

**Solution:** Note that if $z_5 = x_2 \oplus x_3$, then

$$\begin{aligned}\mathbb{P}(X_2 = x_2, X_3 = x_3 \mid Z_5 = z_5) &= \frac{\mathbb{P}(X_2 = x_2, X_3 = x_3, Z_5 = z_5)}{\mathbb{P}(Z_5 = z_5)} \\ &= \frac{\mathbb{P}(X_2 = x_2, X_3 = x_3)}{\mathbb{P}(z_5)},\end{aligned}$$

and $\mathbb{P}(X_2 = x_2, X_3 = x_3 \mid Z_5 = z_5) = 0$ otherwise. Furthermore, we also have

$$\mathbb{P}(Z_5 = 0) = \mathbb{P}(X_2 = 0, X_3 = 0) + \mathbb{P}(X_2 = 1, X_3 = 1) = (1 - q)^2 + q^2$$

and

$$\mathbb{P}(Z_5 = 1) = \mathbb{P}(X_2 = 0, X_3 = 1) + \mathbb{P}(X_2 = 1, X_3 = 0) = 2q(1 - q).$$

Therefore,

$$\begin{aligned}
\mathbb{P}(X_2 = 0, X_3 = 0 \mid Z_5 = 0) &= \frac{(1-q)^2}{(1-q)^2 + q^2} \\
\mathbb{P}(X_2 = 1, X_3 = 1 \mid Z_5 = 0) &= \frac{q^2}{(1-q)^2 + q^2} \\
\mathbb{P}(X_2 = 1, X_3 = 0 \mid Z_5 = 0) &= 0 \\
\mathbb{P}(X_2 = 0, X_3 = 1 \mid Z_5 = 0) &= 0,
\end{aligned}$$

and similarly,

$$\begin{aligned}
\mathbb{P}(X_2 = 0, X_3 = 0 \mid Z_5 = 1) &= 0 \\
\mathbb{P}(X_2 = 1, X_3 = 1 \mid Z_5 = 1) &= 0 \\
\mathbb{P}(X_2 = 1, X_3 = 0 \mid Z_5 = 1) &= \frac{1}{2} \\
\mathbb{P}(X_2 = 0, X_3 = 1 \mid Z_5 = 1) &= \frac{1}{2}.
\end{aligned}$$

(b) Draw a directed graphical model (the graph and conditional probability tables) for these five random variables. What independence relations does the graph imply?

**Solution:**

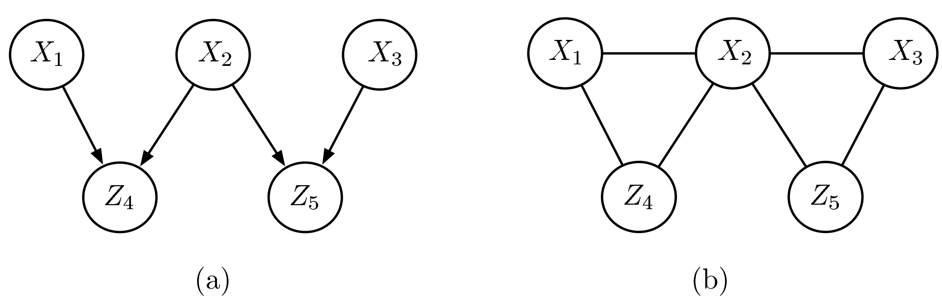


Figure 1: (a) Directed graphical model for Problem 3.3(b). (b) Undirected graphical model for Problem 3.3(c).

The directed graphical model is given in Figure 1(a). The conditional probability tables display the conditional probability distributions for each node given its parents. The tables for this problem are listed below. Note that the tables for the $X_i$'s are simply the marginal distributions, and that given their parents, the nodes $Z_j$ are specified exactly.

- For $X_1$, $X_2$, and $X_3$:

| $X_i = 0$ | $X_i = 1$ |
|---|---|
| $1 - q$ | $q$ |

- For $Z_4$:

| | $Z_4 = 0$ | $Z_4 = 1$ |
|---|---|---|
| $X_1 = 0, X_2 = 0$ | 1 | 0 |
| $X_1 = 0, X_2 = 1$ | 0 | 1 |
| $X_1 = 1, X_2 = 0$ | 0 | 1 |
| $X_1 = 1, X_2 = 1$ | 1 | 0 |

- For $Z_5$:

| | $Z_5 = 0$ | $Z_5 = 1$ |
|---|---|---|
| $X_2 = 0, X_3 = 0$ | 1 | 0 |
| $X_2 = 0, X_3 = 1$ | 0 | 1 |
| $X_2 = 1, X_3 = 0$ | 0 | 1 |
| $X_2 = 1, X_3 = 1$ | 1 | 0 |

We list the conditional independence assertions implied by the graph below. Note that if $X_A \perp X_B \mid X_C$, then certainly $X_{A'} \perp X_{B'} \mid X_C$ for any $A' \subseteq A, B' \subseteq B$ (cf. Problem 3.2(c)). Hence, we will only list the largest sets $X_A$ and $X_B$ which are conditionally independent given $X_C$; all other conditional independence assertions follow trivially from this fact.

- $X_1 \perp (X_2, X_3, Z_5)$
- $X_1 \perp X_2 \mid (X_3, Z_5)$
- $X_1 \perp Z_5 \mid (X_2, X_3, Z_4)$
- $X_1 \perp (X_2, X_3) \mid Z_5$
- $X_1 \perp (X_2, Z_5) \mid X_3$

- $X_1 \perp (X_3, Z_5) \mid (X_2, Z_4)$
- $X_2 \perp (X_1, X_3)$
- $X_2 \perp X_3 \mid (X_1, Z_4)$
- $X_3 \perp (X_1, X_2, Z_4)$
- $X_3 \perp (X_1, X_2) \mid Z_4$
- $X_3 \perp (X_2, Z_4) \mid X_1$
- $X_3 \perp (X_1, Z_4) \mid (X_2, Z_5)$
- $X_3 \perp Z_4 \mid (X_1, X_2, Z_5)$
- $Z_4 \perp (X_3, Z_5) \mid (X_1, X_2)$
- $Z_4 \perp Z_5 \mid (X_1, X_2, X_3)$
- $Z_5 \perp (X_1, Z_4) \mid (X_2, X_3)$
- $(X_1, Z_4) \perp (X_3, Z_5) \mid X_2$

(c) Draw an undirected graphical model (the graph and compatibility functions) for these five variables. What independence relations does it imply?

**Solution:** The graphical model is the moralized form of the graph in part (b), given in Figure 1(b).

The maximal cliques in this graphical model can be parametrized in two ways, either as

$$\Psi_{\{X_1,Z_4,X_2\}}(x_1, z_4, x_2) = p(x_1)p(x_2)p(z_4|x_1, x_2)$$
$$\Psi_{\{X_2,Z_5,X_3\}}(x_2, z_5, x_3) = p(x_3)p(z_5|x_2, x_3),$$

or

$$\Psi_{\{X_1,Z_4,X_2\}}(x_1, z_4, x_2) = p(x_1)p(z_4|x_1, x_2)$$
$$\Psi_{\{X_2,Z_5,X_3\}}(x_2, z_5, x_3) = p(x_3)p(x_2)p(z_5|x_2, x_3).$$

Th conditional independence assertions implied by the graph are

- $(X_1, Z_4) \perp (X_3, Z_5) \mid X_2$
- $Z_4 \perp (X_3, Z_5) \mid (X_1, X_2)$
- $(X_1, Z_4) \perp Z_5 \mid (X_2, X_3)$
- $X_1 \perp (X_3, Z_5) \mid (Z_4, X_2)$
- $(X_1, Z_4) \perp X_3 \mid (X_2, Z_5)$

- $Z_4 \perp Z_5 \mid (X_1, X_2, X_3)$
- $Z_4 \perp X_3 \mid (X_1, X_2, Z_5)$
- $X_1 \perp Z_5 \mid (Z_4, X_2, X_3)$
- $X_1 \perp X_3 \mid (Z_4, X_2, Z_5)$.

As before, we have omitted the independencies that may be derived directly from the decomposition rule of Problem 3.2(c).

(d) Under what conditions on $q$ do we have $Z_5 \perp X_3$ and $Z_4 \perp X_1$? Are either of these marginal independence assertions implied by the graphs in (b) or (c)?

**Solution:** First note that when $q \in \{0, 1\}$, $X_1, X_2$, and $X_3$ are all constant random variables, which implies that $Z_4$ and $Z_5$ are also constant, so $Z_5 \perp X_3$ and $Z_4 \perp X_1$.

Next, consider $0 < q < 1$. For independence to hold, we need

$$\mathbb{P}(Z_5 = 1) = \mathbb{P}(Z_5 = 1 \mid X_3 = 0) = \mathbb{P}(Z_5 = 1 \mid X_3 = 1).$$

Note that

$$\mathbb{P}(Z_5 = 1 \mid X_3 = 0) = \mathbb{P}(X_2 = 1) = q,$$
$$\mathbb{P}(Z_5 = 1 \mid X_3 = 1) = \mathbb{P}(X_2 = 0) = 1 - q,$$

and as before,

$$\mathbb{P}(Z_5 = 1) = 2q(1 - q).$$

Thus, we have independence for $0 < q < 1$ if and only if $q = \frac{1}{2}$. These marginal independence assertions are not implied by the graphs.

**Problem 3.4**

Consider a sequence of random variables $(X_1, \ldots, X_d)$ generated according to the following procedure:

(i) Sample $X_1 \sim N(0, 1)$.

(ii) Given some $a \in (-1, 1)$, for $t = 1, \ldots, d - 1$, set $X_{t+1} = aX_t + \sqrt{1 - a^2}\, W_t$, where the $\{W_t\}_{t=1}^{d-1}$ are independent $N(0, 1)$ variables, with $W_t$ chosen independently of $X_t$.

(a) Compute the covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ of the random vector $X \in \mathbb{R}^d$.

**Solution:** Let $W = \begin{pmatrix} W_0 & W_1 & \cdots & W_{d-1} \end{pmatrix}^\top$, where $W_0 := X_1$, and $X = \begin{pmatrix} X_1 & X_2 & \cdots & X_d \end{pmatrix}^\top$. By construction, we can write $X = AW$, where $A$ is the $d \times d$ lower triangular matrix given by

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots \\ a & \sqrt{1-a^2} & 0 & 0 & \cdots \\ a^2 & a\sqrt{1-a^2} & \sqrt{1-a^2} & 0 & \cdots \\ \vdots & \vdots & \vdots & & \ddots \end{pmatrix}.$$

Clearly $W \sim N(0, I)$. Therefore, $X$ is also normally distributed, being a linear combination of jointly Gaussian random variables, with mean $\mathbb{E}[X] = A\mathbb{E}[W] = 0$ and covariance

$$\Sigma = \mathrm{Cov}(X) = \mathbb{E}[XX^\top] = A\,\mathbb{E}[WW^\top]\,A^\top = AA^\top.$$

Given the lower triangular matrix $A$ above, the covariance matrix $\Sigma = AA^\top$ has a nice structure:

$$\Sigma = \begin{pmatrix} 1 & a & a^2 & a^3 & \cdots & a^{d-1} \\ a & 1 & a & a^2 & \cdots & a^{d-2} \\ a^2 & a & 1 & a & \cdots & a^{d-3} \\ a^3 & a^2 & a & 1 & \cdots & a^{d-3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a^{d-1} & a^{d-2} & a^{d-3} & a^{d-4} & \cdots & 1 \end{pmatrix}.$$

(b) Show that the inverse covariance matrix $\Sigma^{-1}$ is always tridiagonal, meaning that it is non-zero only on its diagonal and on the two diagonals above and below the main diagonal. (I.e., $(\Sigma^{-1})_{ij} = 0$ for all $|i-j| > 1$.)

**Solution:**

We prove a more general statement about Gaussian graphical model: in an undirected graphical model $G = (V, E)$ where the node variables $(X_1, \ldots, X_d)$ are jointly Gaussian with zero mean and covariance matrix $\Sigma$, we have $(i, j) \notin E$ if and only if $(\Sigma^{-1})_{ij} = 0$. In our particular problem, observe that the sequence $(X_1, \ldots, X_d)$ is a Markov chain, so if $|i-j| > 1$, then $(i, j) \notin E$ in the graph. Since we already know that $(X_1, \ldots, X_d)$ is jointly Gaussian, the claim above tells us that if $|i-j| > 1$ then $(\Sigma^{-1})_{ij} = 0$, i.e. $\Sigma^{-1}$ is tridiagonal.

Now to prove the claim, fix any pair of indices $(i,j) \notin E$, and let $x_{\setminus\{i,j\}}$ be the subvector of $x = (x_1, \ldots, x_d)$ with components $i$ and $j$ removed. We write $\Theta = \Sigma^{-1}$ for simplicity. Then for each $x \in \mathbb{R}^d$, we can write the multivariate Gaussian density as

$$\begin{aligned}
p(x) &= \frac{1}{Z} \exp\Big( -\frac{1}{2} x^\top \Theta x \Big) \\
&= \frac{1}{Z} \exp(-x_i x_j \Theta_{ij}) \, \exp\Big( -\frac{1}{2} \sum_{k \neq i} x_k x_j \Theta_{kj} \Big) \\
&\qquad \exp\Big( -\frac{1}{2} \sum_{l \neq j} x_i x_l \Theta_{il} \Big) \, \exp\Big( -\frac{1}{2} \sum_{k,l \neq i,j} x_k x_l \Theta_{kl} \Big) \\
&= \exp(-x_i x_j \Theta_{ij}) \, g_i(x_i, x_{\setminus\{i,j\}}) \, g_j(x_j, x_{\setminus\{i,j\}}) \qquad (1)
\end{aligned}$$

where $Z$ is the normalization constant and $g_i, g_j$ are positive functions that we constructed to make the above equation true.

Since $p$ is positive and obeys the graph $G$, the Hammersley-Clifford Theorem tells us that there exist positive potential functions $\Psi_c$ defined on the maximal cliques, $\mathcal{C}$, of $G$ such that $p(x) = \prod_{c \in \mathcal{C}} \Psi_c(x_c)$. Since $(i,j) \notin E$, $i$ and $j$ do not appear together in any clique, so, by collecting those potential terms that depend on $x_j$ and those that do not, we can write

$$p(x) = f_i(x_i, x_{\setminus\{i,j\}}) f_j(x_j, x_{\setminus\{i,j\}}) \qquad (2)$$

for some positive functions $f_i, f_j$.

Combining equations 1 and 2, we have

$$\exp(-x_i x_j \Theta_{ij}) = \frac{f_i(x_i, x_{\setminus\{i,j\}})}{g_i(x_i, x_{\setminus\{i,j\}})} \, \frac{f_j(x_j, x_{\setminus\{i,j\}})}{g_j(x_j, x_{\setminus\{i,j\}})} = h_i(x_i) \, h_j(x_j),$$

for positive functions $h_i, h_j$ (the second equality follows as the left-hand side does not depend on $x_{\setminus\{i,j\}}$). When $x_i = 0$, we see $h_j(x_j) = 1/h_i(0)$ for all $x_j$, implying that $h_j(x_j)$ is a constant function of $x_j$. Similarly, $h_i(x_i)$ is constant, and so $\exp(-x_i x_j \Theta_{ij})$ is a constant function of $x_i, x_j$. Hence we have $\exp(-x_i x_j \Theta_{ij}) = \exp(-0 \cdot 0 \cdot \Theta_{ij}) = 1$ and thus we conclude that $\Theta_{ij} = 0$.

(*Hint:* You may want to simulate this numerically just to confirm the intuition. In proving the result, the Hammersley-Clifford theorem could be helpful.)

**Problem 3.5**
Consider the directed graph shown in Figure 2(a). For each of the following conditional independence statements, verify whether or not they hold. In each case, be explicit using the Bayes ball algorithm, indicating how the ball gets through, or how it is blocked for each possible path.

(a) $X_2 \perp X_8 \mid \{X_3, X_4, X_5\}$.

**Solution:** False. The ball can go directly from 2 to 8.

(b) $X_8 \perp X_9 \mid \{X_3, X_4, X_5\}$.

**Solution:** False. The ball can go from 8 to 9 following the path $8-2-6-4-7-9$.

(c) $X_7 \perp X_{10} \mid \{X_3, X_4, X_5\}$.

**Solution:** False. The ball can go from 7 to 10 following the path $7-9-5-9-10$.

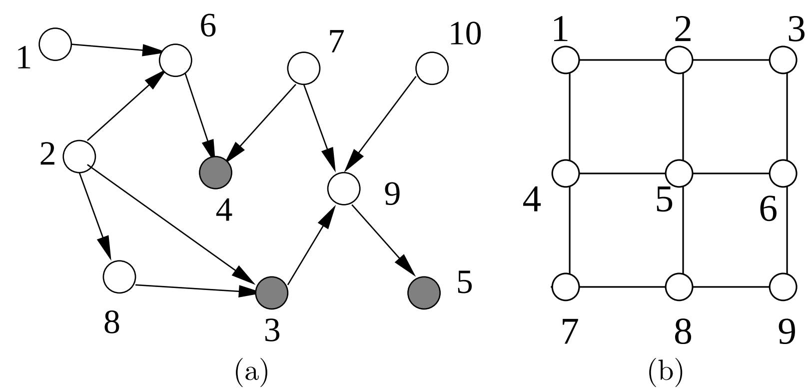


Figure 2: (a) A directed graph. (b) An undirected graphical model: a $3 \times 3$ grid, frequently used in spatial statistics and image processing.

**Problem 3.6**
*Undirected graphs and elimination:* Consider the undirected graph in Figure 2(b): it is a $3 \times 3$ grid or lattice graph.

(a) Sketch the sequence of graphs obtained by running the algorithm `Graph-eliminate`:

(i) Following the ordering $\{5, 4, 8, 6, 2, 9, 3, 7, 1\}$?

(ii) Following the ordering $\{1, 7, 3, 9, 2, 4, 6, 8, 5\}$?

What is the largest clique formed by each graph sequence? Which ordering is preferable?

**Solution:**

(i) The sequence of graph elimination is given in Figure 3. The largest *elimination* cliques, as defined in Section 3.2.1 of the reader, are $T_4 = \{1, 2, 4, 6, 7, 8\}$, $T_8 = \{1, 2, 6, 7, 8, 9\}$, and $T_6 = \{1, 2, 3, 6, 7, 9\}$, so the largest intermediate clique is of size 6.

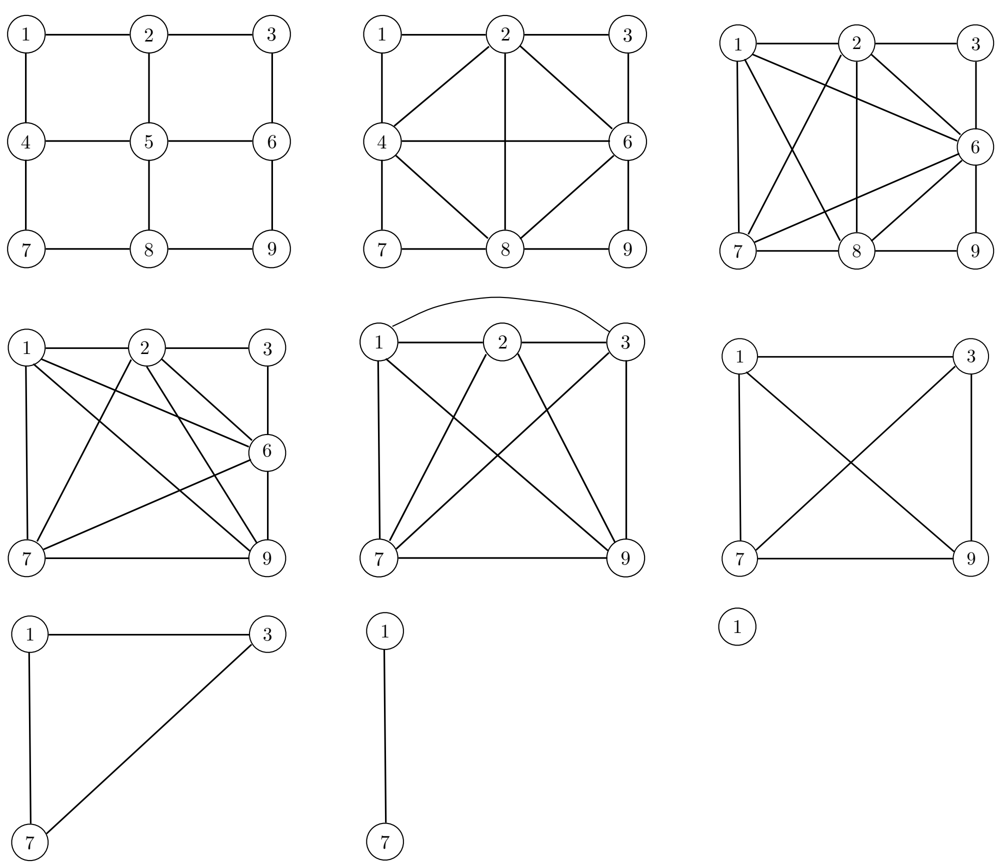

Figure 3: Sequence of graph elimination following the ordering $\{5, 4, 8, 6, 2, 9, 3, 7, 1\}$.

(ii) The sequence of graph elimination is given in Figure 4. The largest elimination cliques are $T_2 = \{2, 4, 5, 6\}$ and $T_4 = \{4, 5, 6, 8\}$, so in this case the largest intermediate clique is of size 4. Thus this ordering is more preferable.

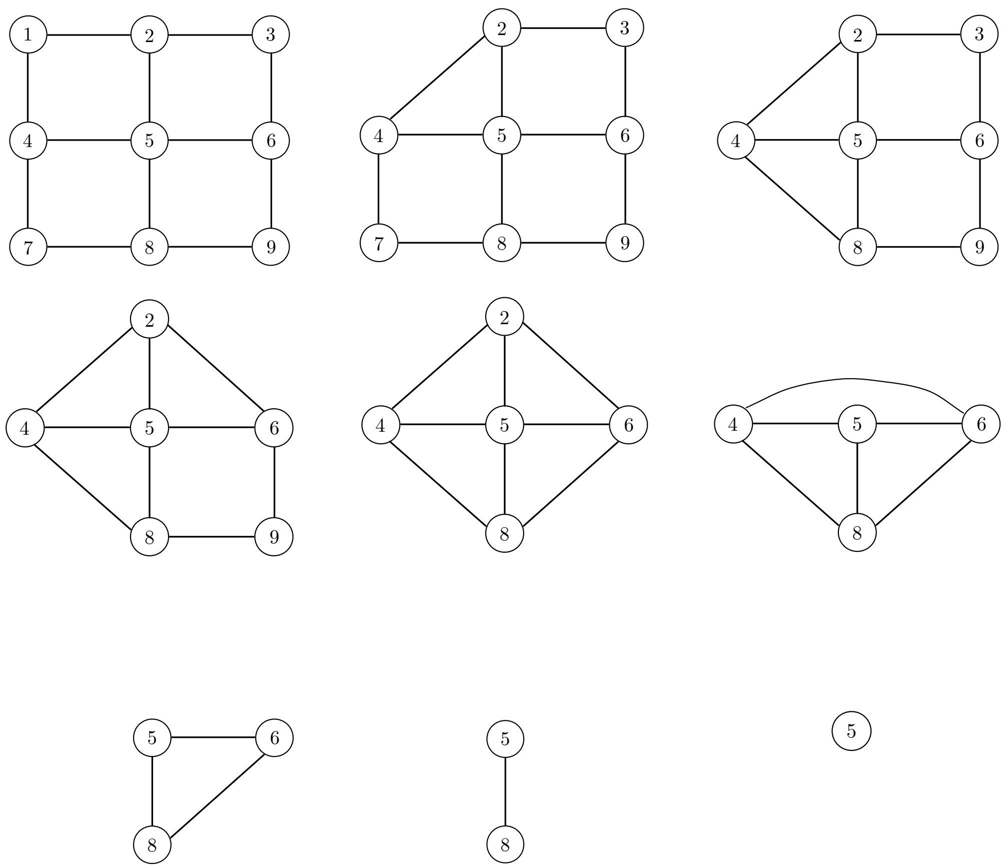

Figure 4: Sequence of graph elimination following the ordering $\{5, 4, 8, 6, 2, 9, 3, 7, 1\}$.

(b) Using intuition from the previous example ($n = 3$), give a reasonable ($\ll n^2$) upper bound on the treewidth of the $n \times n$ grid.

**Solution:** If we eliminate the nodes one row at a time, from left to right, then it is easy to see that at each step the largest elimination clique is of size at most $n + 1$. Thus the treewidth of the $n \times n$ grid is at most $n$.