In this appendix, we consider the case in which each variable $x_i$, $i = 1, ..., n$, takes values in $\{-1, 1\}$, and the marginal distribution over each $x_i$ induced by $p$ is uniform. This can be viewed as a coarsening of an underlying Gaussian distribution, such that $x_i$ records the sign of a Gaussian variable.

We do not have a complete analysis of our problem for this specification of $p$, and focus on the chain model $1 \rightarrow 2 \rightarrow \cdots \rightarrow n$. In Eliaz et al. (2019), we provided a characterization of the maximal estimated correlation that such a model can generate in a uniform-binary environment. The proof was by induction on $n$. Here we give a constructive proof that emphasizes the analogy with the Gaussian case. Our analysis is based on a few preliminary observations.

**Definition 1** A $n \times n$ matrix $C$ is called “Binary Factorizable” (BF) if it can be written as

$$C = \lim_{M \to \infty} \frac{1}{M} A_M A_M^T$$

Where each $A_M$ is a $n \times M$ matrix whose elements are all $\pm 1$ and each row of $A_M$ is zero mean.

Note that any BF matrix is symmetric, positive semi-definite, and has ones on the diagonal. Note also that any covariance matrix of zero-mean binary random variables must be BF, since we can define the matrix $A_M$ as
a sample covariance matrix, where the sample consists of $M$ i.i.d draws from the underlying distribution. The converse is also true: any BF matrix corresponds to the covariance matrix of zero-mean binary random variables. This can be seen by defining a distribution over $n$ binary variables by randomly picking (with probability $1/M$) one of the columns of $A_M$.

Somewhat surprisingly, however, there exist symmetric, positive semi-definite matrices which are not BF. For example, the reader may recall the following correlation matrix from the example in the Introduction, where it gave the maximal false correlation for $n = 3$ in the Gaussian environment:

$$C = \begin{pmatrix} 1 & b & 0 \\ b & 1 & b \\ 0 & b & 1 \end{pmatrix}$$

with $b = \sqrt{1/2}$. This matrix is not BF. As we will see below, the largest value of $b$ for which $C$ is BF is $1/2$.

**Proposition 1** Suppose all variables take values in $\{-1, 1\}$ and the objective distribution $p$ induces a uniform marginal over each variable. Let the objective (Pearson) coefficient of correlation between $x_1$ and $x_n$, according to $p$, is $r$. Then, the maximal estimated correlation that can be achieved by a linear DAG $G : 1 \rightarrow 2 \rightarrow \cdots \rightarrow n$ is given by:

$$\rho_{1n}^* = \max_{\rho_{ij}, \text{for all } i, j \atop \rho_{ii} = 1 \text{ for all } i \atop \rho_{ii} = r} \prod_{i=1}^{n-1} \rho_{i,i+1}$$

**Proof.** The constraints are self-evident. We only need to show that for a linear DAG defined over uniformly distributed binary variables, the estimated correlation between $x_1$ and $x_n$ is given by the product of the objective pairwise correlations of adjacent variables (as in the Gaussian case). We can show this by viewing $p_G(x_1, \ldots, x_n) = p(x_1)p(x_2 | x_1) \cdots p(x_n | x_{n-1})$ as a Markov chain. The conditional probability $p_G(x_n | x_1)$ is thus given by a matrix product - specifically, the product of all the transition matrices defined
by \( p(x_{i+1} \mid x_i) \). Since all variables are uniformly distributed, the transition matrices are doubly stochastic, which means that they have the same eigenvectors. The top eigenvalue is always 1 and the second eigenvalue gives the correlation. Since all matrices have the same eigenvectors, the eigenvalues just multiply. ■

Note that Proposition 1 is exactly the same as the intermediate result we established at the beginning of Section 4.3 for the Gaussian environment. The only difference is that we replace the requirement that \( \rho \) be positive semi-definite with the requirement that \( \rho \) be BF. As mentioned above, the set of BF matrices is smaller than the set of positive semi-definite matrices. Therefore, we should expect a more stringent upper bound on the maximal false correlation.

**Proposition 2** Suppose all variables take values in \([-1, 1]\) and the objective distribution \( p \) induces a uniform marginal over each variable. Let the objective (Pearson) coefficient of correlation between \( x_1 \) and \( x_n \), according to \( p \) be equal to \( r \). Then, the maximal estimated correlation that can be generated by the DAG \( 1 \to 2 \to \cdots \to n \) is given by:

\[
\rho_{1n}^* = \left( 1 - \frac{1}{n-1} (1 - r) \right)^{n-1}
\]

(1)

**Proof.** From Proposition 1, we know that the maximal estimated correlation is obtained by multiplying elements in a BF correlation matrix \((\rho_{ij})\) such that \( \rho_{1n} = r \). For any \( n \times M \) matrix \( A_M \), let \( a_i^{(M)} \) denote its \( i^{th} \) row. Then, we can rewrite the estimated correlation induced by \( C_M = \frac{1}{M} A_M A_M^T \) as:

\[
\prod_{i=1}^{n-1} \frac{1}{M} a_i^{(M)} (M) a_{i+1}^{(M)}
\]

As we discussed following the definition of BF matrices, the dot product between the \( i^{th} \) and \( j^{th} \) rows of \( A_M \) is proportional to the empirical correlation of \( x_i \) and \( x_j \) in a sample consisting of \( M \) \( i.i.d \) draws from the underlying distribution.
Given a matrix $A_M$ that gives an objective correlation of $\rho_{1n} = r$, we can always attempt to improve the estimated correlation by optimizing all other rows of the matrix $a_2, \ldots, a_{n-1}$. This implies that for any $M$:

$$\rho_{1n}^* \leq \max_{a_2, \ldots, a_{n-1} \in \{-1,1\}^M, a_1 = a_1^{(M)}, a_n = a_n^{(M)}} \prod_{i=1}^{n-1} \frac{1}{M} a_i^T a_{i+1}$$  \hspace{1cm} (2)

This is an upper bound for two reasons. First, we are not enforcing the constraint that the binary vectors $a_i$ are zero mean. Second, if $C = \frac{1}{M} A_M A_M^T$ for some finite $M$, then $C$ is BF.

For binary vectors $a_i, a_j \in \{-1,1\}^M$, the dot product $\frac{1}{M} a_i^T a_j$ is a monotone function of the proportion $q$ of components for which the two vectors agree: $\frac{1}{M} a_i^T a_j = 2q - 1$. Thus, maximizing the dot product between two binary vectors is equivalent to minimizing the number of components on which they disagree. This means that the R.H.S of (2) is a form of a shortest path on a lattice: we are given two points in $\{-1,1\}^M$ ($a_1$ and $a_n$), and seek a set of intermediate points on this lattice that are as close as possible to each other. By analogy, in the third step of our proof for the Gaussian case, we were also given two vectors in a high-dimensional space (an $n$-dimensional unit sphere) and searched for a set of intermediate points on the sphere such that the intermediate points are as close as possible to one another (in terms of spherical distance).

To solve this “shortest path on a lattice” problem, we divide the $M$ indices into two disjoint groups: $M_1$ indices $k$ for which $a_1(k) = a_n(k)$ and $M_{-1}$ indices $k$ for which $a_1(k) \neq a_n(k)$. For any of the $M_1$ indices for which $a_1(k) = a_n(k)$, setting $a_i(k) = a_1(k)$ for all $i$ can only increase the objective function (since this can only increase the dot product between consecutive vectors).

For the remaining $M_{-1}$ indices $k$ for which $a_1(k) \neq a_n(k)$, denote by $m_i$ the number of indices $k$ for which $a_i(k) = a_1(k)$ and $a_i(k) \neq a_n(k)$. Assuming $m_i > m_j$, the dot product between $a_i$ and $a_j$ can be written as follows:

$$a_i^T a_j = M - 2(m_i - m_j)$$
This enables us to rewrite (2) as:

\[
\rho_{1n}^* \leq \max_{m_2, \ldots, m_{n-1}} \prod_{i=1}^{n-1} \frac{1}{M} (M - 2(m_{i-1} - m_i)) \tag{3}
\]

The R.H.S. of (3) should be maximized subject to the constraint that \( m_i \in \{0, 1, \ldots, M-1\} \), but we can get an upper bound by maximizing over real-valued \( m_i \).

Taking the logarithm of the R.H.S of (3) and differentiating with respect to \( m_i \) yields that at an optimum, \( m_i \) should be linearly spaced between \( m_1 \) and \( m_n \):

\[
m_i - m_{i+1} = \frac{M-1}{n-1}
\]

Thus, the optimal shortest path is a set of binary vectors whose components agree with \( x_1 \) and \( x_n \) whenever they coincide, and the rest of the indices agree with \( x_1 \) with a fraction that decreases linearly with \( i \).

Now, for large \( M \), \( M-1/M \) converges to the probability that \( x_1 \neq x_n \), namely \( \frac{1 - r}{2} \), such that

\[
\frac{1}{M} a_i^T a_{i+1} \to \left( 1 - \frac{1}{n-1} (1 - r) \right)
\]

Since there are \( n-1 \) such dot products, we take their product, thus obtaining the R.H.S of (1).

To show that the upper bound is tight, given two uniform binary random variables \( x_1, x_n \) that satisfy \( E(x_1 x_n) = r \), consider a set of variables \( x_i \), whose distribution conditional on \( x_1, x_n \) is defined as follows:

- If \( x_1 = x_n \), then \( x_i = x_1 = x_n \).
- If \( x_1 \neq x_n \), then \( x_i = x_1 \) with probability \( 1 - \frac{i}{n} \) and \( x_i = x_n \) with probability \( \frac{i}{n} \).

By construction, a vector of \( M \) random samples from \( x_i \) and \( x_{i-1} \) will generate a normalized dot product \( \frac{1}{M} a_i^T a_{i+1} \) that converges to \( \left( 1 - \frac{1}{n-1} (1 - r) \right) \) when \( M \to \infty \), thus attaining the upper bound.
It is also worth noting that in Eliaz et al. (2019), we implement the upper bound by taking the $n$ variables to be the sign of the Gaussian variables we used in the implementation of the upper bound of our the main theorem. ■

Let us illustrate the upper bound. For $n = 3$ and $r = 0$, the maximal estimated correlation between $x_1$ and $x_3$ using the chain model $1 \rightarrow 2 \rightarrow 3$ is $\frac{1}{4}$ (compared with the value $\frac{1}{2}$ in the Gaussian case). Finally, for any $r$, the maximal estimated correlation converges to $e^{r-1}$ as $n \to \infty$ (compared with 1 in the Gaussian case).