On the Consistent Recovery of Joint Distributions from Conditionals

Mahbod Majid[*] **Rattana Pukdee**[*] **Vishwajeet Agrawal**[*] **Burak Varıcı** **Pradeep Ravikumar**
Carnegie Mellon University

Abstract

Self-supervised learning methods that mask parts of the input data and train models to predict the missing components have led to significant advances in machine learning. These approaches learn conditional distributions $p(x_T \mid x_S)$ simultaneously, where x_S and x_T are subsets of the observed variables. In this paper, we examine the core problem of when all these conditional distributions are consistent with some joint distribution, and whether common models used in practice can learn consistent conditionals. We explore this problem in two settings. First, for the complementary conditioning sets where $S \cup T$ is the complete set of variables, we introduce the concept of *path consistency*, a necessary condition for a consistent joint. Second, we consider the case where we have access to $p(x_T \mid x_S)$ for all subsets S and T. In this case, we propose the concepts of *autoregressive* and *swap consistency*, which we show are necessary and sufficient conditions for a consistent joint. For both settings, we analyze when these consistency conditions hold and show that standard discriminative models *may fail to satisfy them*. Finally, we corroborate via experiments that proposed consistency measures can be used as proxies for evaluating the consistency of conditionals $p(x_T \mid x_S)$, and common parameterizations may find it hard to learn true conditionals.

1 INTRODUCTION

In recent years, self-supervised learning has emerged as a powerful paradigm in machine learning, significantly advancing fields such as natural language processing and computer vision. Models like BERT (Devlin et al., 2019) in NLP and Masked Autoencoders (He et al., 2022) in vision employ strategies where parts of the input data are masked or removed, and the model is trained to predict these missing components from the remaining observed data. Furthermore, large language models (LLMs) (Achiam et al., 2023; Jiang et al., 2023; Dubey et al., 2024) have demonstrated remarkable capabilities employing autoregressive strategies, where the model predicts the next word in a sequence given the preceding context. This autoregressive formulation effectively conditions on a subset of the input data to predict future tokens.

Formally, these self-supervised objectives effectively give us access to conditional distributions $p(x_T \mid x_S)$ where x_S is a subset of variables (e.g., unmasked) and x_T is another subset of the remaining variables (e.g., masked). If these conditionals are imperfectly learned, e.g., due to spurious features or general overfitting, they would fail an *external consistency* requirement: they need not be close to ground truth conditionals. One could say, they might nonetheless be learning something useful about the ground truth distributions: as Box has noted, all models are wrong, but some are useful (Box, 1976). The problem arises when we have not one but many conditional distribution models: then we would get models and predictions that lack *internal consistency* – even if they have some external validity, they can contradict one another. Such inconsistencies can arise even under stringent parametric assumptions, such as in conditional Poisson models, where the Poisson parameter (as a function of the conditioning variables) has to be negatively correlated with sufficient statistics for these to be consistent with each other (Inouye et al., 2017). Naturally, when we have more flexible conditional distributions, the conditions under which we can get inconsistencies can get even more subtle (as we detail in the sequel).

In this paper, we thus address the following general question: given a set of conditional distributions, when are they consistent with some joint distribution? And if so, how can we reconstruct this joint distribu-

[*] Equal contribution.

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

tion $p(x)$ from these learned conditionals? Addressing these questions is crucial as extracting a consistent joint distribution allows for probabilistic reasoning, uncertainty quantification, and the generation of new data samples that faithfully represent the underlying data manifold.

We explore these questions in two settings. First, we consider complementary conditioning sets, where we only have access to $p(x_T|x_S)$ where $T \cup S$ equals the complete set of variables. We introduce the concept of path consistency, a necessary condition for consistent recovery of the joint distribution. Path consistency involves reconstructing $p(x)$ by navigating the paths in a graphical model, ensuring that the conditional probabilities are aligned along these paths. We investigate which parameterizations lead to conditionals that satisfy path consistency. Specifically, we show that common discriminative models may not satisfy path consistency if not parameterized properly, highlighting limitations in their ability to reconstruct a consistent joint distribution solely from conditional distributions. Notably, we show that logistic regression models satisfy path consistency if and only if formulated within the degree-2 exponential family distributions. This underlines the important role of the exponential family in recovering a consistent joint distribution which was previously studied for univariate conditional distributions (Yang et al., 2015).

Second, we extend our investigation to a more general setting where $T \cup S$ can be a subset of the complete set of variables. Here, we propose an autoregressive path consistency condition based on the observation that we can reconstruct a joint distribution with an autoregressive path, which we prove to be both sufficient and necessary for consistency. Furthermore, we introduce the concept of swap consistency, which focuses on the interchangeability of variables within the conditionals that provide an equivalent but easier-to-verify condition. However, satisfying these conditions can be challenging; we show that a composition of a linear function and a set-invariant context featurizer does not have enough representational power to ensure consistency (except for a very restricted class of distributions).

In our experiments, we find that both path and swap consistency have a strong correlation with autoregressive consistency, thus providing us with a practical metric to test for consistency without having access to all possible conditionals. Furthermore, we investigate the parameterization of conditionals via more complex neural networks, e.g., multi-layer perceptrons (MLPs). Interestingly, we do not observe a significant improvement over simple logistic parameterization. This finding exposes a significant gap in current modeling approaches, so we pose an open question: How can we parameterize models of conditional distributions consistently that enable us to model rich real-world distributions?

Related Work. *Identifying joint distributions from conditionals* has been extensively studied in statistics. For discrete conditionals $p(x|y)$ and $p(y|x)$, compatibility conditions based on ratio matrices have been established (Arnold and Press, 1989; Arnold et al., 2004; Song et al., 2010). In probabilistic graphical models, the Hammersley-Clifford theorem (Hammersley and Clifford, 1971; Besag, 1974) shows that conditionals consistent with a Markov random field yield a consistent joint distribution. Yang et al. (2015) later extends this result to conditionals where they have shown that if the univariate conditionals are specified by an exponential family then there exists a unique joint distribution that is consistent with the conditionals and it is also specified by an exponential family (Theorem 2). Our work complements this by showing that logistic univariate conditionals are path consistent only when specified by an exponential family, though we later demonstrate that this may be insufficient for general conditioning settings (Theorem 4.9).

Prior work by Besag (1974); Tierney (1994) noted the possibility of recovering joint distributions by varying one variable at a time, similar to our observations in equations (3) and (4). Hobert and Casella (1998); Wang and Ip (2008); Chen (2010) established that when conditionals correspond to a consistent joint distribution, the recovered joint must be path-independent, which aligns with our Proposition 3.3. We build on these results by providing the necessary and sufficient conditions for path consistency (Theorem 3.4, Definition 3.2). We further provide an example of this condition for a general class of discriminative models (Theorem 3.5). While previous work focuses on univariate conditionals, we generalize our result to multivariate conditionals in the complementary conditional setting (Section 3.2) and the general conditional setting (Section 4) with a new notion of autoregressive and swap consistency (Definitions 4.3 and 4.5) where we derived the necessary and sufficient conditions (Theorem 4.6).

Our work also applies to measuring the compatibility between autoencoders $p(x|z)$ and $q(z|x)$, which represents a special case with two conditionals $f_{x_S|x_T}, f_{x_T|x_s}$ where $x_S = x, x_T = z$. Our path consistency result implies condition iv) in the compatibility criterion of Liu et al. (2021) (Theorem 2.3). Our path consistency metric (Equation (25)) can effectively measure consistency levels in this context, with only two possible paths ($x \to z$ and $z \to x$) simplifying the calculation.

2 PRELIMINARIES

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be an instance space and let P be a distribution over $\mathcal{X}$ with the probability density (or mass) function $p(x)$. For each $x \in \mathcal{X}$, we write $x = (x_1, \ldots, x_d)$ where x_i is the i^{th} coordinate of x. We use uppercase letters (e.g., X) to represent random variables and lowercase letters (e.g., x) for their realizations. We also denote a set $\{1, \ldots, d\}$ by $[d]$. For any set $S \subseteq [d]$, we refer x_S to a tuple $(x_i)_{i \in S}$.

In self-supervised learning, a popular training objective is to mask part of the input and train a model to predict the missing component from the remaining. Formally, for each sample x, and some set $S \subseteq [d]$, $T \subseteq [d] \setminus S$, we are training a model to recover x_T from x_S. This objective is equivalent to learning the conditional $p(x_T \mid x_S)$. We can think of this learning process as projecting the true conditional distribution onto some other parametric space of conditional distributions, such as those parameterized by deep neural networks, so that we learn estimates $f_{T|S}(x_T, x_S)$ of $p(x_T \mid x_S)$. For different pairs of $\{S, T\}$, we can learn $f_{T|S}$ functions separately or can have shared components among them to enhance efficiency. With the true conditional $p(x_T \mid x_S)$, there always exists a joint distribution $p(x)$ which is consistent with all of $p(x_T \mid x_S)$ by definition. However, the existence of a joint distribution consistent with these estimates $f_{T|S}$ is *not* guaranteed. This might depend in subtle ways (as we show in the sequel) on the parametric space and the learning algorithm used to learn these conditional distributions. Formally, we define a consistent joint distribution as follows:

Definition 2.1 (Consistent joint distribution)**.** *Let $\mathcal{A} = \{(T, S) \mid S, T \subseteq [d], S \cap T = \emptyset\}$ be a set of conditioning pairs. Conditional distributions $\{q(x_T; x_S)\}_{(T,S)\in\mathcal{A}}$ correspond to a consistent joint distribution when there exists a joint distribution q^* such that*

$$q^*(x_T \mid x_S) = q(x_T; x_S)\,, \tag{1}$$

for all $x \in \mathcal{X}$ and $(T, S) \in \mathcal{A}$.

We use the notation $q(x_T; x_S)$ to refer to a distribution of x_T given x_S. Our goal is to examine the conditions under which $\{f_{T|S}(x_T, x_S)\}_{(T,S)\in\mathcal{A}}$ corresponds to a consistent joint distribution and investigate how to recover that consistent joint distribution.

3 COMPLEMENTARY CONDITIONING SETS

We start with the setting of complementary conditioning sets where our set of conditioning sets $\mathcal{A}$ contains only pairs (T, S) such that $T \cup S = [d]$ and $|S|$ can be of any size. In particular, we first consider the setting when $|S| = d - 1$, where we learn to predict only one variable from the rest, $\mathcal{A}_u := \{(i, [d] \setminus \{i\}) \mid i \in [d]\}$. We define $x_{-i} := x_{[d]\setminus\{i\}}$ and also write $f_{i|-i}$ for $f_{\{i\}|[d]\setminus\{i\}}$. In this special case, previous work has shown that if the conditionals

$$\mathcal{F}_u := \{f_{i|-i}(x_i, x_{-i}) \mid \forall i \in [d]\} \tag{2}$$

correspond to a univariate conditional exponential family, then there exists a consistent joint if and only if the univariate conditional distributions have a specific form and share specific sub-functions with each other (Yang et al., 2015). We aim to generalize this result to arbitrary conditional distributions. Although this turns out to be difficult to characterize at such a high level of generality, we are nonetheless able to identify conditions for a slightly more relaxed notion, which we term *path consistency*.

Before defining this notion, we start by providing a strategy to recover a joint distribution from conditionals and investigate a general necessary condition for the functions $\mathcal{F}_u$ to be consistent. The proofs of all the results in this section are given in Appendix A.

Our key observation is that given conditional distributions $q(x_i \mid x_{-i})$ for all $i \in [d]$, we can always recover the joint distribution $q(x)$ up to a constant factor via a product of the conditional distributions. For example, if $d = 2$, then for any x_1, x_2 and $\bar{x}_1, \bar{x}_2$, we observe that

$$\frac{q(x_1, x_2)}{q(\bar{x}_1, x_2)} = \frac{q(x_1 \,|\, x_2)}{q(\bar{x}_1 \,|\, x_2)}\ , \text{ and } \frac{q(\bar{x}_1, x_2)}{q(\bar{x}_1, \bar{x}_2)} = \frac{q(x_2 \,|\, \bar{x}_1)}{q(\bar{x}_2 \,|\, \bar{x}_1)}\,. \tag{3}$$

Multiplying these two equations, we get

$$\frac{q(x_1, x_2)}{q(\bar{x}_1, \bar{x}_2)} = \frac{q(x_1 \mid x_2)}{q(\bar{x}_1 \mid x_2)} \cdot \frac{q(x_2 \mid \bar{x}_1)}{q(\bar{x}_2 \mid \bar{x}_1)}\,. \tag{4}$$

Since this holds for any $x_1, x_2, \bar{x}_1, \bar{x}_2$, by fixing $\bar{x}_1, \bar{x}_2$, we can recover the joint distribution $q(x_1, x_2)$ up to a constant factor. The strategy here is to go from (x_1, x_2) to $(\bar{x}_1, \bar{x}_2)$ by changing one variable at a time. This allows us to write the ratio of the joint distribution as a product of conditional distributions. This leads to a natural approach to recover a joint distribution (up to a constant) from the given $\mathcal{F}_u$.

Definition 3.1 (Path-Recovered Joint Distribution)**.** *For a set of functions $\mathcal{F}_u$, a constant $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_d)$, and a permutation function $\sigma : [d] \to [d]$ (which represents a path from x to $\bar{x}$ by changing one variable at a time), the corresponding path-recovered joint distribution is given by*

$$h_{\sigma,\bar{x}}(x; \{f_{i|-i}\}) = \prod_{i=1}^{d} \frac{f_{\sigma(i)|-\sigma(i)}(x_{\sigma(i)}, x'_{-i,\sigma})}{f_{\sigma(i)|-\sigma(i)}(\bar{x}_{\sigma(i)}, x'_{-i,\sigma})}\,, \tag{5}$$

where $x'_{-i,\sigma} = \{x_{\sigma(s)} | s > i\} \cup \{\bar{x}_{\sigma(s)} | s < i\}$.

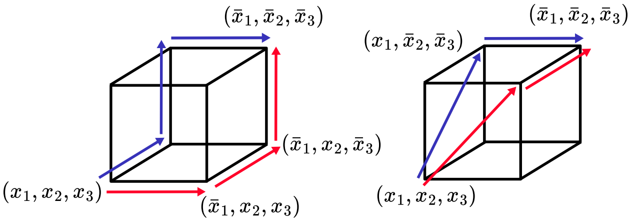
$(\bar{x}_1, \bar{x}_2, \bar{x}_3)$
$(\bar{x}_1, x_2, \bar{x}_3)$
(x_1, x_2, x_3)
$(\bar{x}_1, x_2, x_3)$
$(\bar{x}_1, \bar{x}_2, \bar{x}_3)$
$(x_1, \bar{x}_2, \bar{x}_3)$
(x_1, x_2, x_3)

Figure 1: An illustration of joint distribution recovery with path when $d = 3$. The colored arrows represent different paths from (x_1, x_2, x_3) to $(\bar{x}_1, \bar{x}_2, \bar{x}_3)$ by changing one variable at a time (left) or when we allow changing multiple variables at a time (right). Each path corresponds to a path-recovered joint distribution (Equation (5)). The path consistency (Definition 3.2) ensures that the recovered joint distribution is the same regardless of the path.

We illustrate the path recovery in Figure 1. Since our main goal is to recover a joint distribution that is independent of any permutation σ, it is desirable if the path-recovered joint $h_{\sigma,\bar{x}}$ is also independent of permutation σ. We formally define this as a path consistency condition.

Definition 3.2 (Path Consistency)**.** *Functions* $\mathcal{F}_u = \{f_{i|-i}(x_i, x_{-i}) \mid i \in [d]\}$ *are path consistent if for any* $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_d)$ *and any permutation functions* $\sigma, \sigma' : [d] \to [d]$*, we have*

$$h_{\sigma,\bar{x}}(x; \mathcal{F}_u) = h_{\sigma',\bar{x}}(x; \mathcal{F}_u) \,. \tag{6}$$

In fact, we can show that the path consistency condition is a necessary condition for $\mathcal{F}_u$ to correspond to a consistent joint distribution. Furthermore, if there exists such a consistent joint distribution, path recovery (Equation (5)) will lead to the correct joint distribution (which is also unique), making this a principled approach to recover the joint distribution.

Proposition 3.3 (Correctness of path recovery and Necessity of Path Consistency)**.** *If a set of functions* $\mathcal{F}_u$ *correspond to a consistent joint distribution* q*, then they are also path consistent, so that for any permutation* σ*, the path-recovered joint distribution satisfies*

$$h_{\sigma,\bar{x}}(x; \mathcal{F}_u) = \frac{q(x)}{q(\bar{x})} \,. \tag{7}$$

Consequently, path consistency is a necessary condition for $\mathcal{F}_u$ *to correspond to a consistent joint distribution.*

Moving forward, we will focus on the path consistency condition. We choose to emphasize this condition because it is a desirable property to have as discussed earlier and it is also a necessary condition for a consistent joint. Moreover, path consistency is easier to verify in practice than the consistency condition itself since it only requires checking every permutation σ (of which there are $d!$ permutations), while verifying consistency involves exploring the space of all possible distributions in $\mathcal{X}$. Regardless, both search spaces grow substantially as dimension d increases. Next, we provide a necessary and sufficient condition for any $f_{i|-i}$ to be path consistent, which provides insight into how one can design a parameterization of conditionals that are always path consistent.

Theorem 3.4 (Necessary and Sufficient Condition for Path Consistency)**.** *Functions* $f_{i|-i}(x_i, x_{-i})$ *for* $i \in [d]$ *are path consistent if and only if there exist functions* $h(x)$ *and* $q_i(x_{-i})$ *for* $i \in [d]$ *such that*

$$f_{i|-i}(x_i, x_{-i}) = h(x)\, q_i(x_{-i}) \,. \tag{8}$$

Theorem 3.4 implies that there must be a shared structure h among all $f_{i|-i}$ for the path consistency to hold. In the next section, we investigate the discrete classification setting, and show that if we learn $f_{i|-i}$ separately for each i with discriminative models, then $f_{i|-i}$ may not be consistent.

3.1 Discriminative Models May Not be Path Consistent

Here, we investigate the path consistency of discriminative models for the discrete classification setting where each $x_i \in \{1, \ldots, K\}$ for all $i \in [d]$. We then consider general discriminative models of the form:

$$f_{i|-i}(x_i, x_{-i}) = \frac{\exp\left(W_{x_i,i}^\top \phi_i(x_{-i})\right)}{\sum_{l=1}^k \exp\left(W_{l,i}^\top \phi_i(x_{-i})\right)} \,, \tag{9}$$

where $\phi_i(x_{-i}) \in \mathbb{R}^m$ is a feature map and $W_{l,i} \in \mathbb{R}^m$ are the softmax weight vectors for the l-th possible value of x_r. In deep neural models, $\phi_i(x_{-i})$ is the pre-softmax layer activations, and in logistic regression, $\phi_i(x_{-i}) = x_{-i}$. In the next result, we show that path consistency requires that the neural features ϕ_i and ϕ_j of different $f_{i|-i}$ and $f_{j|-j}$ have to follow specific constraints that also involve their softmax weights.

Theorem 3.5. *Let* $f_{i|-i}(x_i, x_{-i})$ *be parameterized as in Equation* (9) *for* $i \in [d]$*. If* $f_{i|-i}$ *are path consistent, then for any* $i, j \in \{1, \ldots, d\}$ *and for any possible values* $x_i, x_i', x_j, x_j' \in [K]$*, we must have*

$$\begin{aligned}&(W_{x_i,i} - W_{x_i',i})^\top \left(\phi_i(x_j, x_{-\{i,j\}}) - \phi_i(x_j', x_{-\{i,j\}})\right) \\ &= (W_{x_j,j} - W_{x_j',j})^\top \left(\phi_j(x_i, x_{-\{i,j\}}) - \phi_j(x_i', x_{-\{i,j\}})\right) .\end{aligned} \tag{10}$$

This relation is rather specific and if we learn $f_{i|-i}$ and $f_{j|-j}$ independently, it is unlikely that this condition will hold. To see this more clearly, we consider the case of logistic regression where $\phi_i(x_{-i}) = x_{-i}$ then

Theorem 3.5 implies that the parameterization has to belong to exponential family distribution,

$$f_{i|-i}(x_i, x_{-i}) \propto \exp\Big(\sum_{j \neq i} A_{i,j}\, x_j x_i + B_{i,j} x_j\Big)\,, \quad (11)$$

with $A_{i,j} = A_{j,i}$. We provide examples of models that satisfy Theorem 3.4, Theorem 3.5 in Appendix D. It is an interesting open question whether we can design neural architectures so that these constraints are automatically satisfied, similar to how CNNs architecturally encode translation invariance constraints.

3.2 Complementary Conditioning Sets with Arbitrary Size

Next, we consider a more general setting when $|S|$ may be smaller than $n-1$. Remarkably, our proposed idea of recovering the joint via paths is still applicable in this case. The key difference is that, instead of moving from x to $\bar{x}$ with one variable at a time, we may change more than one variable at a time. We only need to ensure that we can change all x_j to $\bar{x}_j$. To do this, we first define a complete tuple.

Definition 3.6 (Complete tuple)**.** *Let $\mathcal{A} = \{(T,S) \mid T \cup S = [d], T \cap S = \emptyset\}$ be a set of complementary conditioning set pairs. A tuple $\mathcal{T} = (T_1, \dots, T_k)$ is a complete tuple of $\mathcal{A}$ if $\bigcup_{i=1}^k T_i = [d]$ and $(T_i, [d] \setminus T_i) \in \mathcal{A}$ for all $i \in [k]$.*

If $\mathcal{T} = (T_1, \dots, T_k)$ is a complete tuple of $\mathcal{A}$ then for any permutation function $\sigma : [k] \to [k]$, $\mathcal{T}_\sigma = (T_{\sigma(1)}, \dots, T_{\sigma(k)})$ is also a complete tuple of $\mathcal{A}$. We can use any complete tuple of $\mathcal{A}$ to recover a joint distribution with a path along that tuple, and extend our results to this general setting. To this end, we define the general path consistency as follows.

Definition 3.7 (General Path Consistency)**.** *Let $\mathcal{A} = \{(T,S) \mid T \cup S = [d], T \cap S = \emptyset\}$ be a set of complementary conditioning set pairs. For a set of functions $\mathcal{F} = \{f_{T|S}(x_T, x_S) \mid (T,S) \in \mathcal{A}\}$, a constant $\bar{x} = (\bar{x}_1, \dots, \bar{x}_d)$, and a complete tuple of $\mathcal{A}$, $\mathcal{T} = (T_1, \dots, T_k)$, the corresponding path-recovered joint distribution is given by*

$$h_{\bar{x},\mathcal{T}}(x; \mathcal{F}) = \prod_{i=1}^{k} \frac{f_{T_i|S_i}\left(x'_{T_i}, x'_{S_i}\right)}{f_{T_i|S_i}\left(\bar{x}_{T_i}, x'_{S_i}\right)}\,, \quad (12)$$

where for $A \in \{S_i, T_i\}$, $x'_{A_i} = \{\bar{x}_j \mid j \in A_i \cap U_i\} \cup \{x_j \mid j \in A_i \setminus U_i\}$ and $U_i = \bigcup_{s=1}^{i-1} T_s$ is the union of all variables updated in previous steps. Furthermore, we say that $\mathcal{F}$ is path consistent if for any complete tuples $\mathcal{T}_1, \mathcal{T}_2$ of $\mathcal{A}$, we have

$$h_{\bar{x},\mathcal{T}_1}(x; \mathcal{F}) = h_{\bar{x},\mathcal{T}_2}(x; \mathcal{F})\,. \quad (13)$$

Similarly to Proposition 3.3, we can show that general path consistency is a necessary condition for consistency, i.e., if $\mathcal{F}$ corresponds to a consistent joint distribution q, the path recovered distribution satisfies

$$h_{\bar{x},\mathcal{T}}(x; \mathcal{F}) = \frac{q(x)}{q(\bar{x})}\,, \quad \forall x, \bar{x} \in \mathcal{X}\,. \quad (14)$$

Furthermore, Theorem 3.4 can be generalized in this case to show that $\mathcal{F}$ is path consistent if and only if there exists $h(x)$ and $q_S(x_S)$ for any S and $T = [d] \setminus S$,

$$f_{T|S}(x_T, x_S) = h(x)\, q_S(x_S)\,. \quad (15)$$

To summarize, we have extensively explored the path consistency condition, which we have shown to be a necessary condition for the existence of a consistent joint distribution. We also provide a necessary and sufficient condition for path consistency, further demonstrating that when each conditional $f_{i|-i}$ is learned separately, this condition may not hold. Generalizing our path consistency results further – bringing them closer to achieving full consistency of conditionals in the setting of complementary conditioning sets – remains an open question for future research.

4 GENERAL CONDITIONING SETS

In this section, we extend our analysis to the setting where the conditioning pairs in $\mathcal{A}$ include subsets (T,S) such that $T \cup S \subseteq [d]$. First, we consider the case where the set of conditioning pairs contains all possible pairs in the form $(\{i\}, S)$ for some set S. We denote this as

$$\mathcal{A}_p := \{(\{i\}, S) \mid \forall i \in [d], \forall S \subseteq [d] \setminus \{i\}\}\,. \quad (16)$$

We also let $\mathcal{F}_p := \{f_{T|S}(x_T, x_S) \mid \forall (T,S) \in \mathcal{A}_p\}$. Note that $\{(i, -i) \mid i \in [d]\} \subseteq \mathcal{A}_p$. Hence, the necessary conditions for the existence of a consistent joint distribution discussed in Section 3 continue to hold in this setting. The following proposition clarifies why we are focusing on this set $\mathcal{A}_p$ (all proofs for this section are given in Appendix B).

Proposition 4.1. *For any conditioning set $\mathcal{A}$ such that $\mathcal{A}_p \subseteq \mathcal{A}$, the conditionals $\mathcal{F} = \{f_{T|S}(x_T, x_S) \mid (T,S) \in \mathcal{A}\}$ are consistent provided the following two conditions hold:*

(a) The conditionals in $\mathcal{F}_p$ are consistent with a unique joint distribution q.

(b) The conditionals in $\mathcal{F}_{\mathcal{A} \setminus \mathcal{A}_p} = \{f_{T|S}(x_T, x_S) \mid (T,S) \in \mathcal{A} \setminus \mathcal{A}_p\}$ are consistent with q. This can be verified using only $\mathcal{F}_p$.

Our key observation is that assuming the first condition holds, the second condition can be verified simply as follows. When $\mathcal{F}_p$ corresponds to a consistent joint distribution q, this, in turn, can be used to recover the marginals $q(x_S)$ for any $S \subseteq [d]$. This also implies that we can recover the conditionals $q(x_T \mid x_S) = q(x_{S\cup T})/q(x_S)$ for any $(S,T) \in \mathcal{A} \setminus \mathcal{A}_p$. Given these, we can simply check whether $f_{T|S}(x_T, x_S) = q(x_T \mid x_S)$. Given this proposition, in the sequel, we focus on $\mathcal{A}_p$ and investigate the conditions under which $\mathcal{F}_p$ corresponds to a consistent joint distribution.

Our next key observation is that with access to conditionals of the form $q(x_i \mid x_S)$, we can utilize an autoregressive model to recover the joint distribution. Specifically, if we have access to the sequence of conditionals $q(x_1)$, $q(x_2 \mid x_1)$, $\ldots$, $q(x_d \mid x_1, \ldots, x_{d-1})$, we can reconstruct the joint distribution by:

$$q(x) = q(x_1) \prod_{i=2}^{d} q(x_i \mid x_{<i}) \,, \tag{17}$$

where $x_{<i} = (x_1, x_2, \ldots, x_{i-1})$. This leads to another approach to recover a joint distribution.

Definition 4.2 (Joint Distribution Recovery with Autoregressive Model)**.** *For a set of functions $\mathcal{F}_p$ and a permutation function $\sigma : [d] \to [d]$, the joint distribution recovered via an autoregressive model with respect to σ is given by*

$$g_\sigma(x; \mathcal{F}_p) = \prod_{i=1}^{d} f_{\sigma(i)|\sigma(<i)}(x_{\sigma(i)}, x_{\sigma(<i)}) \,, \tag{18}$$

where $f_{\sigma(i)|\sigma(<i)} = f_{\{\sigma(i)\}|\{\sigma(1),\ldots,\sigma(i-1)\}}$ and $x_{\sigma(<i)} = (x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(i-1)})$.

Since we are recovering a joint distribution that is independent of a permutation σ, it would also be desirable if $g_\sigma(x; \mathcal{F}_p)$ is independent of any permutation σ. We define a notion of path consistency for the autoregressive joint distribution recovery.

Definition 4.3 (Autoregressive Path Consistency)**.** *A set of functions $\mathcal{F}_p$ is said to be* autoregressive path consistent *if for any two permutations σ and σ', the autoregressive recovered joint distributions are equal:*

$$g_\sigma(x; \mathcal{F}_p) = g_{\sigma'}(x; \mathcal{F}_p) \,. \tag{19}$$

As with the path consistency notion specified in Definition 3.2, the autoregressive consistency defined above is also a necessary condition for consistency. But, perhaps surprisingly, it is also *a sufficient* condition for consistency.

Theorem 4.4. *A set of functions $\mathcal{F}_p$ is consistent if and only if it is autoregressive path consistent. In particular, when $\mathcal{F}_p$ is consistent, the joint distribution is given by $q(x) = g_\sigma(x; \mathcal{F}_p)$ for any permutation σ.*

This theorem implies that autoregressive path consistency enables a practical algorithm for testing consistency: We can check whether the joint distributions recovered from different permutations are equivalent. Specifically, we can compute the joint distributions g_σ corresponding to all permutations σ and verify their equality. However, can we provide a simpler condition that is easier to check? Toward this goal, we introduce the concept of swap consistency.

Definition 4.5 (Swap Consistency)**.** *A set of functions $\mathcal{F} = \{f_{i|S}(x_i, x_S) \mid (i,S) \in \mathcal{A}\}$ for some set $\mathcal{A} \subset [d] \times 2^{[d]}$ is said to be* swap consistent *if*

$$f_{i|S\cup\{j\}} f_{j|S} = f_{j|S\cup\{i\}} f_{i|S} \tag{20}$$

for all $\{i, j, S\}$ such that $(i,S), (j,S), (i, S\cup\{j\}), (j, S\cup\{i\}) \in \mathcal{A}$.

If $\mathcal{F}$ is consistent with a joint p, then LHS and RHS of (20) represent two ways of factorizing $p(x_i, x_j|x_S)$ as $p(x_i, x_j|x_S) = p(x_i|x_{S\cup\{j\}})p(x_j|x_S) = p(x_j|x_{S\cup\{i\}})p(x_i|x_S)$. Hence, swap consistency can be seen as a necessary condition for the consistency of any $\mathcal{F}$. Furthermore, when $\mathcal{F} = \mathcal{F}_p$, it becomes equivalent to autoregressive consistency since any permutation σ is equivalent to a sequence of permutations, each differing by a swap of consecutive terms. Consequently, the swap consistency also becomes sufficient for consistency.

Theorem 4.6. *If the set of conditionals $\mathcal{F} = \{f_{i|S}(x_i, x_S) | (i,S) \in \mathcal{A}\}$ is consistent with some joint, then $\mathcal{F}$ is swap consistent. Furthermore, when $\mathcal{F} = \mathcal{F}_p$, that is, $\mathcal{F}$ includes all conditionals, swap consistency also becomes sufficient for consistency.*

As noted above, swap consistency is a necessary condition for consistency of arbitrary sets $\mathcal{F}$, and hence can be used as a direct necessary if not sufficient check for consistency of these as well.

4.1 On Parameterization of Conditionals

So far, we have investigated the question of when any given set of conditionals is consistent. Next, we consider this question in the discrete setting where each variable $x_j \in \mathcal{X} = \{1, \ldots K\}$ for a discrete set of size K. Our key observation is that conditional distributions could be viewed as set functions over a particular class of sets. Specifically, $f_{i|S}(x_i, x_S)$ takes as argument the tuple (x, i, S), which could be mapped to the following set:

$$C(x, i, S) = \{(i, j, x_j) \mid j \in S\} \cup \{(i, j) \mid j \notin S\} \,. \tag{21}$$

Next, we define the class of such sets $\mathcal{C} = \cup_{x\in\mathcal{X}, i,j\in[d]} C(x, i, S) \subseteq 2^{[d]\times[d]\times(\mathcal{X}+1)}$.

Proposition 4.7. *The set of functions $\{f_{i|S}(x_i, x_S) \mid S \subseteq [d], i \in [d], i \notin S\}$ for a discrete set $\mathcal{X}$ with $|\mathcal{X}| = K$ can be equivalently written as a set function $\rho : \mathcal{C} \to \mathbb{R}^K$ where:*

$$\rho(C(x, i, S))_v = f_{i|S}(x_i = v, x_S)\,.$$

Since ρ is a function acting over a set, we can use the result of Zaheer et al. (2017, Theorem 2) which showed that any $f : \mathcal{C} \to \mathbb{R}$ with domain as a class of sets can be decomposed as $f(C) = \rho(\sum_{c \in C} \phi(c))$, for suitable transformations ϕ and ρ. This can be readily extended to functions mapping $\mathcal{C} \mapsto \mathbb{R}^K$. Using an embedding to represent each element $c \in \mathcal{X} \times [d] \times [d]$ via $\phi(c) \in \mathbb{R}^D$, we get the following:

Proposition 4.8. *For any set of conditionals $\{f_{i|S}(x_i, x_S) \mid S \subset [d], i \in [d], i \notin S\}$ there exists some $g : \mathbb{R}^D \to \mathbb{R}^K_+$ and $w_{i,j,k}, w_{i,j} \in \mathbb{R}^D$ for $i, j \in [d]$, $k \in \mathcal{X}$ such that*

$$f_{i|S}(x_i = v, x_S) = g(\psi_i(x_S))_v\,, \tag{22}$$

where $\psi_i(x_S) = \sum_{j \in S} w_{i,j,x_j} + \sum_{j \notin S} w_{i,j}$.

Any set of conditionals, whether consistent or not, can be parameterized as described in Proposition 4.8. We aim to characterize the conditions on g under which the resulting conditionals are consistent, or equivalently the set of joint distributions induced by such a parameterized family of g. We show that when g is a logistic parameterization, i.e., a linear layer followed by a sigmoid, it results in a degenerate bag-of-words distribution.

Theorem 4.9. *Let $g(z)_v = \frac{\exp(w_v^\top z)}{\sum_{k=1}^2 \exp(w_k^\top z)}$ for some parameters $w_k \in \mathbb{R}^D$ for each $k \in \{1, 2\}$. The set of conditionals $\{f_{i|S}(x_i, x_S) \mid S \subset [d], i \in [d], i \notin S\}$ are consistent if and only if $p(x)$ can be factorized as*

$$p(x) = \prod_{\{v\} \in V} p(x_v) \prod_{\{u,v\} \in E} p(x_u, x_v) \prod_{C \in \mathcal{C}} p(x_C)\,, \tag{23}$$

where $V, E, \mathcal{C}$ are partitions of the index set $[d]$ into subsets of size 1, 2 and $\geqslant 3$, respectively, and $p(x_C)$ is a degenerate distribution that only depends on the count of the words: $p(x) \propto \exp(f(sum(x, 1), sum(x, 2))$ for some function $f : \mathbb{Z}^2 \to \mathbb{R}$, and $sum(x, k) = \sum_{i=1}^d \mathbb{I}(x_i = k)$.

Theorem 4.9 implies that the probability distribution that can be consistently parameterized with a logistic function g is very limited where the joint distribution must decompose into isolated components: either independent variables, pairwise interactions, or larger clusters that depend only on aggregate counts of categories rather than specific configurations. We expect to obtain richer distributions as we make g more complex. We pose characterizing joint distributions entailed by richer classes of g (e.g., MLPs, transformers, RNNs) as an open question for future research.

5 EXPERIMENTS

The primary goal of our experimental study is to evaluate the performance and consistency of conditional distributions modeled via neural network parameterizations, and test the consistency measures we proposed. Specifically, we aim to

1. Explore whether a deep neural network such as a multi-layer perceptron (MLP) can be used to model g given in Proposition 4.8.
2. Investigate whether path consistency (Definition 3.1) and swap consistency (Definition 4.5) can serve as effective proxies for assessing the consistency of a model.

Data generating process. We follow a similar data generating process to Jiang et al. (2024) (which also learn conditional distributions) and focus on binary variables $\mathcal{X} = \{0, 1\}^d$. To create a controlled environment where the true conditional distributions are known, we construct a joint distribution $p(x)$ using Bayesian networks. Specifically, we generate three DAGs with $d \in \{10, 25, 50\}$ variables and $32, 84, 116$ edges, respectively. For each DAG, the entries of the conditional probability table, i.e., $p(X_i \mid X_{\mathrm{pa}(i)})$'s where $\mathrm{pa}(i)$ denotes the parents of node i, are sampled from the uniform distribution $U(0, 1)$ for each parent configuration.

Learning the conditionals. Since our goal is to learn conditional distributions $p(x_i|x_S)$ for varying i, S, the input to the model is a tuple (i, S, x_S) and output is a distribution over $x_i = \{0, 1\}$. As customary to the related literature (e.g., masked autoencoders), we sample tuples (x, i, S) as follows:

1. Sample an x from the true joint p.
2. Sample the size of a mask m from a uniform distribution over $[d]$.
3. Sample a set S' with size $|S'| = m$ with a uniform probability, and sample $i \in [d] \setminus S'$ with a uniform probability.

This process specifies a tuple (x, i, S) which is drawn from some distribution denoted by $\mathcal{D}$. Then, our objective is given by

$$\min_\theta \mathbb{E}_{(x,i,S)\sim\mathcal{D}}[-\log f_{i|S}(x_i \mid x_S; \theta)]\,. \tag{24}$$

To evaluate the goodness of fit of the learned conditionals, we compare the total variation (TV) distance

Table 1: d_{TV} between the learned model and the true conditionals when g is a 3 layer MLP with hidden size 64. Δ represents the percent reduction in d_{TV} as the embedding dimension D increases from 2 to 512.

d/D	2	8	32	128	512	Δ
10	0.079	0.068	0.062	0.062	0.059	25%
25	0.121	0.110	0.107	0.104	0.100	17%
50	0.111	0.102	0.098	0.098	0.097	13%

Table 2: d_{TV} between the learned model and the true conditionals when g is a L-layer MLP with hidden size 128 ($L = 0$ indicates g is a just a sigmoid). Δ represents the percent reduction in d_{TV} as the number of hidden layers increase from 0 to 8.

d/L	0	3	8	Δ
10	0.089	0.052	0.049	47%
25	0.129	0.097	0.098	17%
50	0.122	0.089	0.089	31%

between the model and true conditional distributions,

$$d_{\mathrm{TV}}(p, f_\theta) = \mathbb{E}_{(x,i,S)\sim\mathcal{D}}\left[\left|f_{i|S}(x_i|x_S;\theta) - p(x_i|x_S)\right|\right].$$

Model & training details. We parameterize $f_{i|S}$ following Proposition 4.8, that is, the weights $w_{i,j,k}$, $w_{i,j}$ for $i \in [d]$, $j \in [d]$, $k \in [2]$. We parameterize g with an MLP with l hidden layers, ReLU activation, skip connection, and batch normalization in each layer. We train each model for 20000 steps with a batch size of 256, and use Adam optimizer with a learning rate of 0.01 for the first 2000 steps and 0.001 for the next 18000 steps.

MLPs fail to learn the true conditionals. We present the total variation (TV) distance between the learned model and the true conditionals. To determine whether g parameterized through an MLP can learn the true conditionals, we varied two key factors: the embedding dimension D (Table 1) and the number of hidden layers L (Table 2). We found that as the embedding dimension D increases from 2 to 512, we observe only slight improvements in TV distance, which diminish as the number of variables d increases. In addition, we observe a similar trend as we increase the model complexity via the number of hidden layers where the TV distance only decreases slightly with a diminishing return in d. These results suggest that MLPs struggle to learn conditional distributions, especially as the number of variables grows.

Table 3: Pearson correlation coefficient of $\mathcal{E}_{PC}$ and $\mathcal{E}_{SC}$ with $\mathcal{E}_{AC}$ when g is a sigmoid ($L = 0$) or an MLP with 1 hidden layer.

d	10	25	50	10	25	50
L	0	0	0	1	1	1
Swap/AR	0.97	0.99	0.99	0.97	0.96	0.98
Path/AR	0.89	0.95	0.92	0.56	0.89	0.97

Table 4: d_{TV} between the learned model and the true conditionals when f is parameterized with a transformer where L denote the number of layers and $d_{\mathrm{emb}} = d_{\mathrm{model}} = 128, d_{\mathrm{head}} = 4$.

d/L	1	4	9
10	0.121	0.112	0.127
25	0.173	0.178	0.176
50	0.189	0.190	0.179

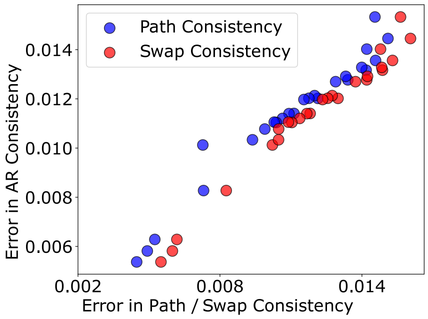

Figure 2: Measures of Path and Swap Consistency ($\mathcal{E}_{\mathrm{PC}}$, $\mathcal{E}_{\mathrm{SC}}$) show high correlation with a measure of Autoregressive Consistency ($\mathcal{E}_{\mathrm{AC}}$) for a model where g is a 1 layer MLP, with $D = 32$, and number of variables $d = 50$.

Transformer underperforms compared to our set-invariant parameterization. Further, we investigate the performance of a transformer model on the same task in Table 4 where we provide full details of the implementation in Appendix C. We found that the TV distance of the transformer models is significantly higher even when using up to 9 layers, compared to our models which follow the parameterization given by Proposition 4.8, even when g is just a sigmoid. One possible explanation could be that our parameterization suggests n^2 embeddings, one for each pair of i, j whereas transformer only has n embeddings, one for each position i.

Path consistency and swap consistency are effective consistency proxies. Since autoregressive consistency is a necessary and sufficient condition for consistency, it serves as a ground truth consistency measure. Next, we define the following metrics to measure each of the proposed consistency measures. For a given distribution we collect all the models f_θ for $(\theta_1, \ldots \theta_t)$ gathered every 500 training steps and compute various consistency metrics.

d	$\mathcal{E}_{\text{SC}}$	$\mathcal{E}_{\text{PC}}$	$-\log p(x_i \mid x_{[d]\setminus i})$	$-\log p(x_i \mid x_{[d]\setminus i,j})$
8	0.43 ± 0.14	0.48 ± 0.29	5.31 ± 1.38	5.98 ± 1.41
16	0.29 ± 0.11	0.45 ± 0.23	3.57 ± 1.05	3.88 ± 1.17
32	0.23 ± 0.10	0.37 ± 0.15	2.92 ± 0.88	3.08 ± 0.97
64	0.16 ± 0.08	0.25 ± 0.09	2.46 ± 0.67	2.62 ± 0.62
128	0.09 ± 0.05	0.18 ± 0.07	2.27 ± 0.99	2.34 ± 0.99
256	0.09 ± 0.06	0.19 ± 0.04	2.21 ± 0.80	2.24 ± 0.75

Table 5: Measures of Path and Swap Consistency of a pretrained BERT model with different values of the context length d. We also report the negative log likelihood for each d.

- Path consistency: We take the mean of standard deviation (std) of $\log h_\sigma$ as σ is varied. Path consistency would imply a small value of this measure,

$$\mathcal{E}_{\text{PC}}(\theta) = \mathbb{E}_{x,\bar{x}\sim P}\left[\text{Std}_\sigma[\log h_{\sigma,\bar{x}}(x; f_\theta)]/d\right]. \quad (25)$$

- Autoregressive consistency: Similarly we take the mean of std of $\log g_\sigma$ which should be small if autoregressive consistency holds,

$$\mathcal{E}_{\text{AC}}(\theta) = \mathbb{E}_{x\sim P}\left[\text{Std}_\sigma[\log g_\sigma(x; f_\theta)]/d\right]. \quad (26)$$

- Swap consistency: We report the mean of the absolute difference between log of LHS and RHS of Equation (20) with respect to a distribution $\mathcal{D}'$ over (x, i, j, S) similar to the way we defined $\mathcal{D}$ over (x, i, S),

$$\mathcal{E}_{\text{SC}}(\theta) = \mathbb{E}_{(x,i,j,S)\sim\mathcal{D}'}[\Delta(f_\theta, x, i, j, S)], \quad (27)$$

$$\Delta(f, x, i, j, S) = \Big| \log\big(f(x_i|x_{S\cup\{j\}})f(x_j|x_S)\big) - \log\big(f(x_j|x_{S\cup\{i\}})f(x_i|x_S)\big)\Big|/2. \quad (28)$$

We note that the metrics defined above directly capture meaningful quantitative metrics for the consistency of our conditionals (rather than being mere yes/no criteria). A detailed discussion of the equations above is provided in Appendix E. In Table 3, we show that $\mathcal{E}_{\text{PC}}$ and $\mathcal{E}_{\text{SC}}$ are highly correlated with $\mathcal{E}_{\text{AC}}$. This result suggests that our measures of path consistency and swap consistency are not only theoretically desirable but are also a practical and useful proxy for measuring how consistent are the learned conditionals.

Experiments with BERT. We demonstrate that our proposed metrics are applicable to modern self-supervised models by calculating the consistency metrics on a pretrained BERT model (Devlin et al., 2019). In particular, we consider bert-base-uncased with 110 million parameters and evaluate our consistency metrics on random sentences of varying length from BookCorpus dataset (Zhu et al., 2015), which was one of the two datasets which BERT was trained on.

We start with the procedure to estimate $\mathcal{E}_{\text{PC}}$. For a given context length d, we sample $n = 1024$ random sentences of length at least d from the corpus and select the first d tokens to get n sentences $x^1, \ldots, x^n$ of length d. Then, we create a random shuffle of these sentences to get reference data points $x'^1, \ldots, x'^n$ to create pairs $(x^1, x'^1), \ldots, (x^n, x'^n)$. Finally, for each pair, we use $k = 30$ random permutations σ to compute $\mathcal{E}_{\text{PC}}$ according to (25). For $\mathcal{E}_{\text{SC}}$, for each $x^1, \ldots x^n$, we use $k = 30$ random pairs of positions $(i, j) \in [d]$. Then, use (28) and (27) to calculate $\mathcal{E}_{\text{SC}}$.

We also report mean NLL (negative log likelihood) of x_i given the $x_{[d]\setminus i}$ (logit for token at position i given the sentence with position i masked) and x_i given another masked token at position j, i.e., given $x_{[d]\setminus\{i,j\}}$ (logit for token at position i given the sentence with positions i, j masked). We observe that both the proposed consistency metrics and NLL become smaller as d gets larger. Regardless, the value of the consistency metric is still positive for $d = 256$, which suggests that modern self-supervised models such as BERT may not be consistent, and learning more consistent representations remains an important research problem.

6 CONCLUSION

In this paper, we addressed the question: given a set of conditional distributions, when are they consistent with a joint distribution? We introduced consistency conditions which are necessary (and sufficient) for a consistent joint distribution which can also serve as a good proxy to evaluate the consistency level in practice. We examined when these conditions hold and demonstrated that standard discriminative models may fail to satisfy them. Furthermore, our experiments showed that even more complex parameterizations fail to achieve consistency. This highlights the inherent difficulty of this problem. We pose an open question on how can we parameterize models of conditional distributions to enable learning of rich joint distributions?

Acknowledgements

We acknowledge the support of NSF via IIS-2211907, ONR via N00014-23-1-2368, and AFRL and DARPA via FA8750-23-2-1015 and Bloomberg Data Science Fellowship.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arnold, B. C., Castillo, E., and Sarabia, J. M. (2004). Compatibility of partial or complete conditional probability specifications. *Journal of statistical planning and inference*, 123(1):133–159.

Arnold, B. C. and Press, S. J. (1989). Compatible conditional distributions. *Journal of the American Statistical Association*, 84(405):152–156.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Chen, H. Y. (2010). Compatibility of conditionally specified models. *Statistics & probability letters*, 80(7-8):670–677.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hammersley, J. M. and Clifford, P. (1971). Markov fields on finite graphs and lattices. *Unpublished manuscript*, 46.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

Hobert, J. P. and Casella, G. (1998). Functional compatibility, markov chains, and gibbs sampling with improper posteriors. *Journal of Computational and Graphical Statistics*, 7(1):42–60.

Inouye, D. I., Yang, E., Allen, G. I., and Ravikumar, P. (2017). A review of multivariate distributions for count data derived from the poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(3):e1398.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jiang, Y., Rajendran, G., Ravikumar, P. K., Aragam, B., and Veitch, V. (2024). On the origins of linear representations in large language models. In *Proc. International Conference on Machine Learning*, Vienna, Austria.

Liu, C., Tang, H., Qin, T., Wang, J., and Liu, T.-Y. (2021). On the generative utility of cyclic conditionals. *Advances in Neural Information Processing Systems*, 34:30242–30256.

Song, C.-C., Li, L.-A., Chen, C.-H., Jiang, T. J., and Kuo, K.-L. (2010). Compatibility of finite discrete conditional distributions. *Statistica Sinica*, pages 423–440.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728.

Wang, Y. J. and Ip, E. H. (2008). Conditionally specified continuous distributions. *Biometrika*, 95(3):735–746.

Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. (2015). Graphical models via univariate exponential family distributions. *The Journal of Machine Learning Research*, 16(1):3813–3847.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. (2017). Deep sets. *Advances in neural information processing systems*, 30.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/**No**/Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/**No**/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
   (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**/No/Not Applicable] We have provided all details of the experimental setting and the models in Section 5 and Appendix C.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**/No/Not Applicable] Please refer to Section 5 and Appendix C.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/**Not Applicable**]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/**Not Applicable**]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/**Not Applicable**]
   (b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/**Not Applicable**]
   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

On the Consistent Recovery of Joint Distributions from Conditionals: Supplementary Material

A PROOFS FOR COMPLEMENTARY CONDITIONING SETS

In the following proofs, for brevity, we will occasionally have a slight abuse of notation in which we can swap the order of the inputs of a function, similar to how we do it in probability. For example, if we have a function $h : (x_1, x_2, x_3) \to \mathbb{R}$, we can write h as

$$h(x_1, x_2, x_3) = h(x_1, x_{\{2,3\}}) = h(x_2, x_1, x_3)\,, \tag{29}$$

where we use the index x_i to note that the value is for the i^{th} input of h and use the set notation x_S to refer to $\{x_i \mid i \in S\}$.

A.1 Proof of Proposition 3.3

Proposition 3.3 (Correctness of path recovery and Necessity of Path Consistency)**.** *If a set of functions $\mathcal{F}_u$ correspond to a consistent joint distribution q, then they are also path consistent, so that for any permutation σ, the path-recovered joint distribution satisfies*

$$h_{\sigma,\bar{x}}(x; \mathcal{F}_u) = \frac{q(x)}{q(\bar{x})}\,. \tag{7}$$

Consequently, path consistency is a necessary condition for $\mathcal{F}_u$ to correspond to a consistent joint distribution.

Proof. Let $\mathcal{F}_u$ correspond to a consistent joint distribution q then for any $i \in [d]$ and $x \in \mathcal{X}$,

$$f_{i|-i}(x_i, x_{-i}) = q(x_i \mid x_{-i})\,. \tag{30}$$

Therefore, for any x_{-i} we have

$$\frac{f_{i|-i}(x_i, x_{-i})}{f_{i|-i}(\bar{x}_i, x_{-i})} = \frac{q(x_i \mid x_{-i})}{q(\bar{x}_i \mid x_{-i})} = \frac{q(x_i, x_{-i})}{q(\bar{x}_i, x_{-i})}\,. \tag{31}$$

We can substitute this back to the definition of the path recovery,

$$h_{\sigma,\bar{x}}(x; \{f_{i|-i}\}) = \prod_{i=1}^{d} \frac{f_{\sigma(i)|-\sigma(i)}(x_{\sigma(i)}, x'_{-i,\sigma})}{f_{\sigma(i)|-\sigma(i)}(\bar{x}_{\sigma(i)}, x'_{-i,\sigma})} \tag{32}$$

$$= \prod_{i=1}^{d} \frac{q(x_{\sigma(i)}, x'_{-i,\sigma})}{q(\bar{x}_{\sigma(i)}, x'_{-i,\sigma})} \tag{33}$$

$$= \frac{q(x)}{q(\bar{x})}\,. \tag{34}$$

The final step is from the definition of $x'_{-l,\sigma} = \{x_{\sigma(s)} | s > l\} \cup \{\bar{x}_{\sigma(s)} | s < l\}$ where we can rewrite it as

$$x'_{-l,\sigma} = \bigcup_{i=1}^{l-1} \{\bar{x}_{\sigma(i)}\} \cup \bigcup_{i=l+1}^{d} \{x_{\sigma(i)}\}\,. \tag{35}$$

Therefore,

$$\{\bar{x}_{\sigma(l)}\} \cup x'_{-l,\sigma} = \{\bar{x}_{\sigma(l)}\} \cup \bigcup_{i=1}^{l-1} \{\bar{x}_{\sigma(i)}\} \cup \bigcup_{i=l+1}^{d} \{x_{\sigma(i)}\} \tag{36}$$

$$= \bigcup_{i=1}^{l} \{\bar{x}_{\sigma(i)}\} \cup \bigcup_{i=l+1}^{d} \{x_{\sigma(i)}\} \tag{37}$$

$$= \bigcup_{i=1}^{l} \{\bar{x}_{\sigma(i)}\} \cup \bigcup_{i=l+2}^{d} \{x_{\sigma(i)}\} \cup \{x_{\sigma(l+1)}\} \tag{38}$$

$$= x'_{-(l+1),\sigma} \cup \{x_{\sigma(l+1)}\} \,. \tag{39}$$

As a consequence, for any l

$$\frac{q(x_{\sigma(l+1)}, x'_{-(l+1),\sigma})}{q(\bar{x}_{\sigma(l)}, x'_{-l,\sigma})} = 1 \,, \tag{40}$$

so that these terms cancel each other out and we are left with the numerator of the first term and the denominator of the last term which are $q(x)$ and $q(\bar{x})$ respectively. □

A.2 Proof of Theorem 3.4

Theorem 3.4 (Necessary and Sufficient Condition for Path Consistency)**.** *Functions $f_{i|-i}(x_i, x_{-i})$ for $i \in [d]$ are path consistent if and only if there exist functions $h(x)$ and $q_i(x_{-i})$ for $i \in [d]$ such that*

$$f_{i|-i}(x_i, x_{-i}) = h(x)\, q_i(x_{-i}) \,. \tag{8}$$

Proof. ⇒) Suppose that $f_{i|-i}$ are path consistent. We first show that $f_{i|-i}$ must be of the form $f_{i|-i}(x_i, x_{-i}) = h(x)q_i(x_{-i})$ for some functions h, q_i. For a fixed constant $\bar{x}$, consider a permutation σ such that $\sigma(1) = i$. Applying the path recovery we have

$$h_{\sigma,\bar{x}}(x; \mathcal{F}_u) = \frac{f_{i|-i}(x_i, x_{-i})}{f_{i|-i}(\bar{x}_i, x_{-i})} \prod_{j=2}^{d} \frac{f_{\sigma(j)|-\sigma(j)}(x_{\sigma(j)}, x'_{-j,\sigma})}{f_{\sigma(j)|-\sigma(j)}(\bar{x}_{\sigma(j)}, x'_{-j,\sigma})} \,. \tag{41}$$

Note that the right-hand side terms do not depend on x_i apart from $f_{i|-i}(x_i, x_{-i})$ because other terms contain $\bar{x}_i$ instead. We define

$$\frac{1}{q_i(x_{-i})} = \frac{1}{f_{i|-i}(\bar{x}_i, x_{-i})} \prod_{j=2}^{d} \frac{f_{\sigma(j)|-\sigma(j)}(x_{\sigma(j)}, x'_{-j,\sigma})}{f_{\sigma(j)|-\sigma(j)}(\bar{x}_{\sigma(j)}, x'_{-j,\sigma})} \,. \tag{42}$$

We have $h_{\sigma,\bar{x}}(x; \mathcal{F}_u) = \frac{f_{i|-i}(x_i, x_{-i})}{q_i(x_{-i})}$, that is, $f_{i|-i}(x_i, x_{-i}) = h_{\sigma,\bar{x}}(x; \mathcal{F}_u)q_i(x_{-i})$. We can repeat the same argument for any $i = 1, \dots, d$,

$$f_{i|-i}(x_i, x_{-i}) = h_{\sigma^{(i)},\bar{x}}(x; \mathcal{F}_u)q_i(x_{-i}) \,, \tag{43}$$

where $\sigma^{(i)}$ is a permutation such that $\sigma^{(i)}(1) = i$. The path consistency condition ensures that we have the same $h(x)$ for all i that is for any permutation $\sigma^{(i)}, \sigma^{(j)}$ we have

$$h_{\sigma^{(i)},\bar{x}}(x; \mathcal{F}_u) = h_{\sigma^{(j)},\bar{x}}(x; \mathcal{F}_u) \,, \tag{44}$$

and we denote this as $h(x)$. Therefore, there exists a function $h(x)$ and $q_i(x_{-i})$ such that for all $i \in [d]$,

$$f_{i|-i}(x_i, x_{-i}) = h(x)\, q_i(x_{-i}) \,. \tag{45}$$

⇐) Suppose that $f_{i|-i}(x_i, x_{-i}) = h(x)\, q_i(x_{-i})$ for some functions h, q_i for all $i \in [d]$. We will show that $\mathcal{F}_u$ are path consistent. For any $\bar{x}$ and any permutation functions σ, σ', we have

$$h_{\sigma,\bar{x}}(x; \mathcal{F}_u) = \prod_{i=1}^{d} \frac{f_{\sigma(i)|-\sigma(i)}(x_{\sigma(i)}, x'_{-i,\sigma})}{f_{\sigma(i)|-\sigma(i)}(\bar{x}_{\sigma(i)}, x'_{-i,\sigma})} \tag{46}$$

$$= \prod_{i=1}^{d} \frac{h(x_{\sigma(i)}, x'_{-i,\sigma})q_{\sigma(i)}(x_{-i,\sigma'})}{h(\bar{x}_{\sigma(i)}, x'_{-i,\sigma})q_{\sigma(i)}(x_{-i,\sigma'})} \tag{47}$$

$$= \prod_{i=1}^{d} \frac{h(x_{\sigma(i)}, x'_{-i,\sigma})}{h(\bar{x}_{\sigma(i)}, x'_{-i,\sigma})} \tag{48}$$

$$= \frac{h(x)}{h(\bar{x})} \,. \tag{49}$$

The last equality follows from the cancellation between h between path from x to $\bar{x}$ which hold from an observation that

$$\frac{h(x_{\sigma(l+1)}, x'_{-(l+1),\sigma})}{h(\bar{x}_{\sigma(l)}, x'_{-l,\sigma})} = 1 \,. \tag{50}$$

We refer to Equation (36) for more details on the derivation. Finally, we can see that the recovered $h_{\sigma,\bar{x}}(x; \mathcal{F}_u)$ does not depend on the permutation σ and, therefore, must be the same for all σ. This implies that $f_{i|-i}$ are path consistent. □

A.3 Proofs for Section 3.1

Theorem 3.5. *Let $f_{i|-i}(x_i, x_{-i})$ be parameterized as in Equation (9) for $i \in [d]$. If $f_{i|-i}$ are path consistent, then for any $i, j \in \{1, \dots, d\}$ and for any possible values $x_i, x'_i, x_j, x'_j \in [K]$, we must have*

$$\begin{aligned}&(W_{x_i,i} - W_{x'_i,i})^\top \left(\phi_i(x_j, x_{-\{i,j\}}) - \phi_i(x'_j, x_{-\{i,j\}})\right)\\ &= (W_{x_j,j} - W_{x'_j,j})^\top \left(\phi_j(x_i, x_{-\{i,j\}}) - \phi_j(x'_i, x_{-\{i,j\}})\right) .\end{aligned} \tag{10}$$

Proof. From Equation (9) we have that

$$f_{i|-i}(x_i, x_{-i}) = \frac{\exp\left(W_{x_i,i}^\top \phi_i(x_{-i})\right)}{\sum_{l=1}^{k} \exp\left(W_{l,i}^\top \phi_i(x_{-i})\right)} \,. \tag{51}$$

Since $f_{i|-i}$ are path consistent, Theorem 3.4 implies that there exists a function h and q_i such that

$$\frac{\exp\left(W_{x_i,i}^\top \phi_i(x_{-i})\right)}{\sum_{l=1}^{k} \exp\left(W_{l,i}^\top \phi_i(x_{-i})\right)} = h(x) q_i(x_{-i}) \tag{52}$$

$$\frac{h(x)}{\exp\left(W_{x_i,i}^\top \phi_i(x_{-i})\right)} = q_i(x_{-i}) \sum_{l=1}^{k} \exp\left(W_{l,i}^\top \phi_i(x_{-i})\right) \,. \tag{53}$$

Note that the terms on the right-hand side are independent of x_i. This implies that as we change the value of x_i, the value of the right-hand side would be the same. Formally, for any $x_i, x'_i \in [K]$,

$$\frac{h(x_i, x_{-i})}{\exp\left(W_{x_i,i}^\top \phi_i(x_{-i})\right)} = \frac{h(x'_i, x_{-i})}{\exp\left(W_{x'_i,i}^\top \phi_i(x_{-i})\right)} \,, \tag{54}$$

$$\frac{h(x_i, x_{-i})}{h(x'_i, x_{-i})} = \frac{\exp\left(W_{x_i,i}^\top \phi_i(x_{-i})\right)}{\exp\left(W_{x'_i,i}^\top \phi_i(x_{-i})\right)} \,. \tag{55}$$

Let $x_{-\{i,j\}}$ denote the rest of x without x_i and x_j. With some abuse of notation, we can rewrite the above equation as

$$\frac{h(x_i, x_j, x_{-\{i,j\}})}{h(x'_i, x_j, x_{-\{i,j\}})} = \frac{\exp\left(W_{x_i,i}^\top \phi_i(x_j, x_{-\{i,j\}})\right)}{\exp\left(W_{x'_i,i}^\top \phi_i(x_j, x_{-\{i,j\}})\right)} \,. \tag{56}$$

Conversely, by swapping i and j we have

$$\frac{h(x_i, x_j, x_{-\{i,j\}})}{h(x_i, x'_j, x_{-\{i,j\}})} = \frac{\exp\left(W_{x_j,j}^\top \phi_j(x_i, v)\right)}{\exp\left(W_{x'_j,j}^\top \phi_j(x_i, x_{-\{i,j\}})\right)} \,. \tag{57}$$

These equations give us the ratio between $h(x)$ as we change one variable x_i to x_i' in terms of $W_{x_i,i}, W_{x_i',i}$ and $\phi_i(x_{-i})$. Our result comes from an observation that we can change from $(x_i, x_j, x_{-\{i,j\}})$ to $(x_i', x_j', x_{-\{i,j\}})$ in two different ways: i) $(x_i, x_j, x_{-\{i,j\}}) \to (x_i', x_j, x_{-\{i,j\}}) \to (x_i', x_j', x_{-\{i,j\}})$, and ii) $(x_i, x_j, x_{-\{i,j\}}) \to (x_i, x_j', x_{-\{i,j\}}) \to (x_i', x_j', x_{-\{i,j\}})$. This corresponds to

$$\frac{h(x_i, x_j, x_{-\{i,j\}})}{h(x_i', x_j', x_{-\{i,j\}})} = \frac{h(x_i, x_j, x_{-\{i,j\}})}{h(x_i', x_j, x_{-\{i,j\}})} \frac{h(x_i', x_j, x_{-\{i,j\}})}{h(x_i', x_j', x_{-\{i,j\}})} = \frac{h(x_i, x_j, x_{-\{i,j\}})}{h(x_i, x_j', x_{-\{i,j\}})} \frac{h(x_i, x_j', x_{-\{i,j\}})}{h(x_i', x_j', x_{-\{i,j\}})} . \tag{58}$$

By substituting (56) and (57) into this equation, we have

$$\exp\left(W_{x_i,i}^\top \phi_i(x_j, x_{-\{i,j\}}) + W_{x_j,j}^\top \phi_j(x_i', x_{-\{i,j\}}) + W_{x_j',j}^\top \phi_j(x_i, x_{-\{i,j\}}) + W_{x_i',i}^\top \phi_i(x_j', x_{-\{i,j\}})\right) \tag{59}$$

$$= \exp\left(W_{x_i',i}^\top \phi_i(x_j, x_{-\{i,j\}}) + W_{x_j',j}^\top \phi_j(x_i', x_{-\{i,j\}}) + W_{x_j,j}^\top \phi_j(x_i, x_{-\{i,j\}}) + W_{x_i,i}^\top \phi_i(x_j', x_{-\{i,j\}})\right) . \tag{60}$$

Rearranging the terms gives us

$$(W_{x_i,i} - W_{x_i',i})^\top \left(\phi_i(x_j, x_{-\{i,j\}}) - \phi_i(x_j', x_{-\{i,j\}})\right) = (W_{x_j,j} - W_{x_j',j})^\top \left(\phi_j(x_i, x_{-\{i,j\}}) - \phi_j(x_i', x_{-\{i,j\}})\right) . \tag{61}$$

□

Next, we consider a special case of logistic regression where $\phi_i(x_{-i}) = x_{-i}$. This lets us better see the impact of Theorem 3.5.

Corollary A.1. *Let $f_{i|-i}(x_i, x_{-i})$ be parameterized as in Equation (9). If $f_{i|-i}$ are path consistent and $\phi_i(x_{-i}) = x_{-i} \in [K]^{d-1}$, (logistic regression) then we must have,*

$$f_{i|-i}(x_i, x_{-i}) \propto \exp\left(\sum_{j \neq i} A_{i,j}\, x_j x_i + B_{i,j} x_j\right) , \tag{62}$$

for some constants $A_{i,j} \in \mathbb{R}$ where $A_{i,j} = A_{j,i}$.

Proof. Recall that we have a parameterization,

$$f_{i|-i}(x_i, x_{-i}) = \frac{\exp\left(W_{x_i,i}^\top \phi_i(x_{-i})\right)}{\sum_{l=1}^k \exp\left(W_{l,i}^\top \phi_i(x_{-i})\right)} , \tag{63}$$

where $\phi_i(x_{-i}) \in \mathbb{R}^{d-1}$ and $W_{l,i} \in \mathbb{R}^{d-1}$. For the sake of notational convenience, we add a dummy i^{th} column with value zero to ϕ_i and $W_{l,i}$ so that they are now a vector in $\mathbb{R}^d$. Note that this does not change any output of $f_{i|-i}(x_i, x_{-i})$ but allows us to write an index more concisely. As a result, we observe that

$$\phi_i(x_j, x_{-\{i,j\}}) - \phi_i(x_j', x_{-\{i,j\}}) = (x_j - x_j')e_j , \tag{64}$$

where e_j denotes a vector in $\mathbb{R}^d$ with value 1 on the j^{th} coordinate and 0 elsewhere. Applying Theorem 3.5, for any i, j and any $x_i \neq x_j, x_i' \neq x_j'$, we have

$$(W_{x_i,i} - W_{x_i',i})^\top \left(\phi_i(x_j, x_{-\{i,j\}}) - \phi_i(x_j', x_{-\{i,j\}})\right) = (W_{x_j,j} - W_{x_j',j})^\top \left(\phi_j(x_i, x_{-\{i,j\}}) - \phi_j(x_i', x_{-\{i,j\}})\right), \tag{65}$$

$$(W_{x_i,i} - W_{x_i',i})^\top (x_j - x_j')e_j = (W_{x_j,j} - W_{x_j',j})^\top (x_i - x_i')e_i, \tag{66}$$

$$(W_{x_i,i} - W_{x_i',i})_j (x_j - x_j') = (W_{x_j,j} - W_{x_j',j})_i (x_i - x_i'), \tag{67}$$

$$\frac{(W_{x_i,i} - W_{x_i',i})_j}{(x_i - x_i')} = \frac{(W_{x_j,j} - W_{x_j',j})_i}{(x_j - x_j')} . \tag{68}$$

We can see that the left-hand side is independent of x_j, x_j' while the right-hand side is also independent of x_i, x_i'. Since this has to hold for any x_i, x_j, there exists a constant $A_{i,j}$ for which

$$\frac{(W_{x_i,i} - W_{x_i',i})_j}{(x_i - x_i')} = \frac{(W_{x_j,j} - W_{x_j',j})_i}{(x_j - x_j')} = A_{i,j} . \tag{69}$$

Therefore, we have

$$(W_{x_i,i} - W_{x'_i,i})_j = A_{i,j}(x_i - x'_i)\,, \tag{70}$$

and setting $x'_i = c$ for some constant c,

$$(W_{x_i,i})_j = (W_{c,i})_j + A_{i,j}(x_i - c)\,. \tag{71}$$

Substitute back to Equation (63), we have that

$$f_{i|-i}(x_i, x_{-i}) \propto \exp(W_{x_i,i}^\top \phi_i(x_{-i})) \tag{72}$$

$$f_{i|-i}(x_i, x_{-i}) \propto \exp(\sum_{j\neq i}(W_{x_i,i})_j x_j) \tag{73}$$

$$f_{i|-i}(x_i, x_{-i}) \propto \exp(\sum_{j\neq i}((W_{c,i})_j + A_{i,j}(x_i - c))x_j) \tag{74}$$

$$f_{i|-i}(x_i, x_{-i}) \propto \exp(\sum_{j\neq i} A_{i,j}x_i x_j + B_{i,j}x_j)\,, \tag{75}$$

where $A_{i,j}, B_{i,j}$ are constant and that $A_{i,j} = A_{j,i}$. □

A.4 Proof for Section 3.2

Proposition A.2 (Correctness of path recovery)**.** *Let $\mathcal{A} = \{(T, S) \mid T \cup S = [d], T \cap S = \emptyset\}$ be a set of complementary conditioning set pairs. If a set of functions $\mathcal{F} = \{f_{T|S}(x_T, x_S) \mid (T, S) \in \mathcal{A}\}$ correspond to a consistent joint distribution q then for any complete tuple of $\mathcal{A}$, $\mathcal{T} = (T_1, \ldots, T_k)$, and any $x, \bar{x}$, the path-recovered joint distribution satisfies*

$$h_{\bar{x},\mathcal{T}}(x; \mathcal{F}) = \frac{q(x)}{q(\bar{x})}\,. \tag{76}$$

Consequently, path consistency is a necessary condition for $\mathcal{F}$ to correspond to a consistent joint distribution.

Proof. Let $\mathcal{F}$ correspond to a consistent joint distribution q, then for any $(T, S) \in \mathcal{A}$ and for any $x \in \mathcal{X}$,

$$f_{T|S}(x_T, x_S) = q(x_T \mid x_S)\,. \tag{77}$$

Recall that for a constant $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_d)$, and a complete tuple of $\mathcal{A}$, $\mathcal{T} = (T_1, \ldots, T_k)$, the corresponding path-recovered joint distribution is given by

$$h_{\bar{x},\mathcal{T}}(x; \mathcal{F}) = \prod_{i=1}^{k} \frac{f_{T_i|S_i}\left(x'_{T_i}, x'_{S_i}\right)}{f_{T_i|S_i}\left(\bar{x}_{T_i}, x'_{S_i}\right)} = \prod_{i=1}^{k} \frac{q(x'_{T_i} \mid x'_{S_i})}{q(\bar{x}_{T_i} \mid x'_{S_i})} = \prod_{i=1}^{k} \frac{q(x'_{T_i}, x'_{S_i})}{q(\bar{x}_{T_i}, x'_{S_i})}\,, \tag{78}$$

where $x'_{S_i} = \{\bar{x}_j \mid j \in S_i \cap U_i\} \cup \{x_j \mid j \in S_i \setminus U_i\}$, $x'_{T_i} = \{\bar{x}_j \mid j \in T_i \cap U_i\} \cup \{x_j \mid j \in T_i \setminus U_i\}$ and $U_i = \bigcup_{s=1}^{i-1} T_s$ is the union of all variables updated in previous steps. Intuitively, x'_{S_i}, x'_{T_i} change all x_j that have been updated in the previous step to $\bar{x}_j$. By definition, we have

$$\begin{aligned}\bar{x}_{T_i} \cup x'_{S_i} &= \bar{x}_{T_i} \cup \{\bar{x}_j \mid j \in S_i \cap U_i\} \cup \{x_j \mid j \in S_i \setminus U_i\} &(79)\\ &= \{\bar{x}_j \mid j \in (S_i \cap U_i) \cup T_i\} \cup \{x_j \mid j \in S_i \setminus U_i\}\,. &(80)\end{aligned}$$

Since $S_i \cup T_i = [d]$, we have

$$(S_i \cap U_i) \cup T_i = (([d] \setminus T_i) \cap \bigcup_{s=1}^{i-1} T_s) \cup T_i = \bigcup_{s=1}^{i} T_s = U_{i+1}\,, \tag{81}$$

and

$$S_i \setminus U_i = ([d] \setminus T_i) \setminus \bigcup_{s=1}^{i-1} T_s = [d] \setminus U_{i+1}\,. \tag{82}$$

Therefore,

$$\bar{x}_{T_i} \cup x'_{S_i} = \{\bar{x}_j \mid j \in U_{i+1}\} \cup \{x_j \mid j \in [d] \setminus U_{i+1}\}\,. \tag{83}$$

On the other hand, we use the fact that $T_{i+1} \cup S_{i+1} = [d]$ to show that

$$x'_{T_{i+1}} \cup x'_{S_{i+1}} = \{\bar{x}_j \mid j \in S_{i+1} \cap U_{i+1}\} \cup \{x_j \mid j \in S_{i+1} \setminus U_{i+1}\} \cup \{\bar{x}_j \mid j \in T_{i+1} \cap U_{i+1}\} \cup \{x_j \mid j \in T_{i+1} \setminus U_{i+1}\} \tag{84}$$

$$= \{\bar{x}_j \mid j \in U_{i+1}\} \cup \{x_j \mid j \in [d] \setminus U_{i+1}\} \tag{85}$$

$$= \bar{x}_{T_i} \cup x'_{S_i} \,. \tag{86}$$

Therefore,

$$\frac{q(x'_{T_{i+1}} \cup x'_{S_{i+1}})}{q(\bar{x}_{T_i} \cup x'_{S_i})} = 1 \,, \tag{87}$$

and that these terms would cancel each other out and we are left with

$$h_{\bar{x},\mathcal{T}}(x;\mathcal{F}) = \frac{q(x'_{T_1}, x'_{S_1})}{q(\bar{x}_{T_k}, x'_{S_k})} = \frac{q(x)}{q(\bar{x})} \,, \tag{88}$$

as required. □

Theorem A.3 (Sufficient and Necessary Condition for Path Consistency)**.** *Let $\mathcal{A} = \{(T,S) \mid T \cup S = [d], T \cap S = \emptyset\}$ be a set of complementary conditioning set pairs. A set of functions $\mathcal{F} = \{f_{T|S}(x_T, x_S) \mid (T,S) \in \mathcal{A}\}$ are path consistent if and only if there exist functions $h(x)$ and $q_S(x_S)$ for any S that $(T,S) \in \mathcal{A}$ such that*

$$f_{T|S}(x_T, x_S) = h(x)\, q_S(x_S) \,. \tag{89}$$

Proof. The proof strategy is similar to the proof of Theorem 3.4. ⇒) Assume that $\mathcal{F}$ is path consistent. Then consider a complete tuple $\mathcal{T} = \{T_1, \ldots, T_k\}$ such that $S_1 = [d] \setminus T_1 = S$. Also, for simplicity, we write $T = T_1$. From path consistency, there exists a unique $h(x)$ such that

$$h(x) = h_{\bar{x},\mathcal{T}}(x;\mathcal{F}) = \prod_{i=1}^{k} \frac{f_{T_i|S_i}\left(x'_{T_i}, x'_{S_i}\right)}{f_{T_i|S_i}\left(\bar{x}_{T_i}, x'_{S_i}\right)} = \frac{f_{T|S}\left(x_T, x_S\right)}{f_{T|S}\left(\bar{x}_T, x_S\right)} \prod_{i=2}^{k} \frac{f_{T_i|S_i}\left(x'_{T_i}, x'_{S_i}\right)}{f_{T_i|S_i}\left(\bar{x}_{T_i}, x'_{S_i}\right)} \,. \tag{90}$$

This follows from the fact that $x'_{T_1} = x_{T_1}$ and $x'_{S_1} = x_S$. Now, we observe that

$$\frac{1}{f_{T|S}\left(\bar{x}_T, x_S\right)} \prod_{i=2}^{k} \frac{f_{T_i|S_i}\left(x'_{T_i}, x'_{S_i}\right)}{f_{T_i|S_i}\left(\bar{x}_{T_i}, x'_{S_i}\right)} \,, \tag{91}$$

depends only on x_S, since the term that depends on x_T is replaced by $\bar{x}_T$. Denote

$$\frac{1}{q_S(x_S)} = \frac{1}{f_{T|S}\left(\bar{x}_T, x_S\right)} \prod_{i=2}^{k} \frac{f_{T_i|S_i}\left(x'_{T_i}, x'_{S_i}\right)}{f_{T_i|S_i}\left(\bar{x}_{T_i}, x'_{S_i}\right)} \,, \tag{92}$$

and we have

$$h(x) = \frac{f_{T|S}\left(x_T, x_S\right)}{q_S(x_S)} \,. \tag{93}$$

Therefore, $f_{T|S}\left(x_T, x_S\right) = h(x) q_S(x_S)$. The path consistency condition ensures that $h(x)$ is well-defined and is the same for all complete tuples $\mathcal{T}$.

⇐) Assume that for a set of function $\mathcal{F} = \{f_{T|S}(x_T, x_S) \mid (T,S) \in \mathcal{A}\}$, there exists a function h, q_S where for any $(T,S) \in \mathcal{A}$,

$$f_{T|S}(x_T, x_S) = h(x) q_S(x_S) \,. \tag{94}$$

We will show that $\mathcal{F}$ is path consistent. For any complete tuple $\mathcal{T} = (T_1, \ldots, T_k)$, we have

$$h_{\bar{x},\mathcal{T}}(x;\mathcal{F}) = \prod_{i=1}^{k} \frac{f_{T_i|S_i}\left(x'_{T_i}, x'_{S_i}\right)}{f_{T_i|S_i}\left(\bar{x}_{T_i}, x'_{S_i}\right)} \tag{95}$$

$$= \prod_{i=1}^{k} \frac{h(x'_{T_i}, x'_{S_i}) q_{S_i}(x'_{S_i})}{h(\bar{x}_{T_i}, x'_{S_i}) q_{S_i}(x'_{S_i})} \tag{96}$$

$$= \prod_{i=1}^{k} \frac{h(x'_{T_i}, x'_{S_i})}{h(\bar{x}_{T_i}, x'_{S_i})} \tag{97}$$

$$= \frac{h(x)}{h(\bar{x})} \, . \tag{98}$$

This implies that $h_{\bar{x},\mathcal{T}}(x; \mathcal{F})$ is independent of $\mathcal{T}$ and therefore $\mathcal{F}$ is path consistent. The final step is due to the fact that

$$x'_{T_{i+1}} \cup x'_{S_{i+1}} = \bar{x}_{T_i} \cup x'_{S_i} \, , \tag{99}$$

so that terms in between cancel with each other. We refer to (84) for the details of the derivation. □

B PROOFS FOR GENERAL CONDITIONING SETS

B.1 Proof of Proposition 4.1

Proposition 4.1. *For any conditioning set $\mathcal{A}$ such that $\mathcal{A}_p \subseteq \mathcal{A}$, the conditionals $\mathcal{F} = \{f_{T|S}(x_T, x_S) \mid (T, S) \in \mathcal{A}\}$ are consistent provided the following two conditions hold:*

(a) The conditionals in $\mathcal{F}_p$ are consistent with a unique joint distribution q.

(b) The conditionals in $\mathcal{F}_{\mathcal{A} \setminus \mathcal{A}_p} = \{f_{T|S}(x_T, x_S) \mid (T, S) \in \mathcal{A} \setminus \mathcal{A}_p\}$ are consistent with q. This can be verified using only $\mathcal{F}_p$.

Proof. By definition, $\mathcal{F}$ corresponds to a consistent joint distribution if there exists a joint distribution q such that for any $(T, S) \in \mathcal{A}$,

$$f_{T|S}(x_T, x_S) = q(x_T \mid x_S) \, . \tag{100}$$

This implies that $\mathcal{F}_p \subseteq \mathcal{F}$ must be consistent with joint distribution q. To show that q is unique, we observe that we can use an autoregressive recovery to recover q

$$q(x) = q(x_1) \prod_{i=2}^{d} q(x_i \mid x_{<i}) \, , \tag{101}$$

$$q(x) = f_{1|\emptyset}(x_1) \prod_{i=2}^{d} f_{i|<i}(x_i, x_{<i}) \, , \tag{102}$$

where we write $f_{i|<i}$ in short for $f_{\{i\}|\{1,\ldots i-1\}}$ and $x_{<i}$ in short for $\{x_1, \ldots, x_{i-1}\}$. Note that the right-hand side only depends on $f_{i|S} \in \mathcal{F}_p$. Therefore, q must be unique for each $\mathcal{F}_p$ and this concludes a). Next, we can see that the first part of b) is straightforward from the definition. We will show that this condition can be verified with access to $\mathcal{F}_p$. For any $f_{T|S} \in \mathcal{F}_{\mathcal{A} \setminus \mathcal{A}_p}$, from consistency with q, we have

$$f_{T|S}(x_T, x_S) = q(x_T \mid x_S) = \frac{q(x_{T \cup S})}{q(x_S)} \, . \tag{103}$$

The final step is from the fact that q is a valid distribution. Similar to above, if $\mathcal{F}_p$ is consistent with q, we can write $q(x_T \mid x_S)$ and $q(x_S)$ as a product of functions in $\mathcal{F}_p$ with an autoregressive recovery. For any $S = \{s_1, \ldots, s_k\}$,

$$q(x_S) = q(x_{s_1}) \prod_{i=2}^{d} q(x_{s_i} \mid x_{<s_i}) \tag{104}$$

$$q(x_S) = f_{1|\emptyset}(x_{s_1}) \prod_{i=2}^{d} f_{\{s_i\}|\{s_j|j<i\}}(x_{s_i}, x_{<s_i}) \tag{105}$$

where we write $x_{<s_i} = \{x_{s_j} \mid j < i\}$ for brevity. For a set $S = \{s_1, \ldots, s_k\}$ where for any $i < j$, $s_i < s_j$, we denote

$$g(x, S, \mathcal{F}_p) = f_{1|\emptyset}(x_{s_1}) \prod_{i=2}^{d} f_{\{s_i\}|\{s_j|j<i\}}(x_{s_i}, x_{<s_i}) \tag{106}$$

as an autoregressive recovery for $q(x_S)$ starting from the variables x_i with the smallest index i to the largest index. We can see that every term in the right-hand side of the equation is in the form of $f_{i|S}$ so it is a product of members of $\mathcal{F}_p$. Also, if $\mathcal{F}_p$ is consistent with q, then we must have

$$g(x, S, \mathcal{F}_p) = q(x_S)\,. \tag{107}$$

Finally, we conclude that if condition a) holds, then condition b) holds only when for any $f_{T|S} \in \mathcal{F}_{\mathcal{A}\setminus\mathcal{A}_p}$,

$$f_{T|S}(x_T, x_S) = q(x_T \mid x_S) = \frac{q(x_{T\cup S})}{q(x_S)} = \frac{g(x, S\cup T, \mathcal{F}_p)}{g(x, S, \mathcal{F}_p)}\,. \tag{108}$$

Since the final term depends only on $\mathcal{F}_p$, we can verify whether the condition b) holds using only $\mathcal{F}_p$.

□

B.2 Proofs of Theorem 4.4 and Theorem 4.6

Recall that Theorem 4.4 and Theorem 4.6 are given by

Theorem 4.4. *A set of functions $\mathcal{F}_p$ is consistent if and only if it is autoregressive path consistent. In particular, when $\mathcal{F}_p$ is consistent, the joint distribution is given by $q(x) = g_\sigma(x; \mathcal{F}_p)$ for any permutation σ.*

Theorem 4.6. *If the set of conditionals $\mathcal{F} = \{f_{i|S}(x_i, x_S)|(i, S) \in \mathcal{A}\}$ is consistent with some joint, then $\mathcal{F}$ is swap consistent. Furthermore, when $\mathcal{F} = \mathcal{F}_p$, that is, $\mathcal{F}$ includes all conditionals, swap consistency also becomes sufficient for consistency.*

In this section, we will prove these theorems with the following plan

1. We will first show that autoregressive path consistency is a necessary condition for a consistent joint.
2. We will show that swap consistency is also a necessary condition for a consistent joint.
3. We will show that when $\mathcal{F} = \mathcal{F}_p$, swap consistency and autoregressive path consistency are equivalent.
4. Finally, when $\mathcal{F} = \mathcal{F}_p$, autoregressive path consistency and swap consistency are sufficient condition for a consistent joint.

Note that, these four results altogether imply Theorem 4.4 and Theorem 4.6.

Proposition B.1. *If $\mathcal{F}_p$ corresponds to a consistent joint distribution then it has to be autoregressive path consistent.*

Proof. Assume that $\mathcal{F}_p$ corresponds to a consistent joint distribution, then there exists a joint distribution q such that for any $f_{i|S} \in \mathcal{F}_p$,

$$f_{i|S}(x_i, x_S) = q(x_i \mid x_S). \tag{109}$$

From its definition, an autoregressive path recovery with respect to a permutation σ is given by

$$g_\sigma(x; \mathcal{F}_p) = \prod_{i=1}^{d} f_{\sigma(i)|\sigma(<i)}(x_{\sigma(i)}, x_{\sigma(<i)})\,, \tag{110}$$

where $f_{\sigma(i)|\sigma(<i)} = f_{\{\sigma(i)\}|\{\sigma(1),\ldots,\sigma(i-1)\}}$ and $x_{\sigma(<i)} = (x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(i-1)})$. Substituting (109), we have

$$g_\sigma(x; \mathcal{F}_p) = \prod_{i=1}^{d} q(x_{\sigma(i)} \mid x_{\sigma(<i)}) = q(x)\,, \tag{111}$$

which is independent of the permutation σ. The final Equation holds since $q(x)$ is a joint distribution so we can recover q with an autoregressive path. Therefore, we can conclude that if $\mathcal{F}_p$ corresponds to a consistent joint distribution then it is also autoregressive path consistent. □

Proposition B.2. *If $\mathcal{F}$ corresponds to a consistent joint distribution then it has to be swap consistent.*

Proof. Assume that $\mathcal{F}$ corresponds to a consistent joint distribution, then there exists a joint distribution q such that for any $f_{i|S} \in \mathcal{F}_p$,

$$f_{i|S}(x_i, x_S) = q(x_i \mid x_S)\,. \tag{112}$$

The proof idea is based on the property that we can factorize a conditional distribution in different ways, where for any set S and $i, j \in [d] \setminus S$, we have

$$q(x_i, x_j, x_S) = q(x_i \mid x_j, x_S) q(x_j \mid x_S) q(x_S)\,, \tag{113}$$

and

$$q(x_i, x_j, x_S) = q(x_j \mid x_i, x_S) q(x_i \mid x_S) q(x_S)\,. \tag{114}$$

As a result, we have

$$q(x_i \mid x_j, x_S) q(x_j \mid x_S) = q(x_j \mid x_i, x_S) q(x_i \mid x_S)\,. \tag{115}$$

Substituting (112), we obtain

$$f_{i|S\cup\{j\}}(x_i, x_S \cup \{x_j\}) f_{j|S}(x_j, x_S) = f_{j|S\cup\{i\}}(x_j, x_S \cup \{x_i\}) f_{i|S}(x_j, x_S)\,, \tag{116}$$

which implies that $\mathcal{F}$ is swap consistent, by definition. □

Proposition B.3. *$\mathcal{F}_p$ is autoregressive path consistent if and only if it is swap consistent.*

Proof. $\Rightarrow$) Assume that $\mathcal{F}_p$ is autoregressive path consistent. For any set $S \subseteq [d]$ and $i, j \in [d] \setminus S$, we will show that $\mathcal{F}_p$ satisfies the condition for swap consistency for (S, i, j). First, we write $S = \{s_1, s_2, \ldots, s_k\}$ and $T = [d] \setminus (S \cup \{i, j\}) = \{t_1, t_2, \ldots, t_{d-k-2}\}$. We consider a permutation $\sigma : [d] \to [d]$,

$$\sigma(m) = \begin{cases} s_m & \text{if } 1 \leqslant m \leqslant k\,, \\ i & \text{if } m = k+1\,, \\ j & \text{if } m = k+2\,, \\ t_{m-k-2} & \text{if } k+3 \leqslant m \leqslant d\,. \end{cases} \tag{117}$$

Equivalently, in one-line notation,

$$\sigma = (s_1, s_2, \ldots, s_k,\ i,\ j,\ t_1, t_2, \ldots, t_{d-k-2})\,. \tag{118}$$

Alternately, define σ' by swapping the i and j,

$$\sigma' = (s_1, s_2, \ldots, s_k,\ j,\ i,\ t_1, t_2, \ldots, t_{d-k-2})\,. \tag{119}$$

By autoregressive path consistency, we must have,

$$g_\sigma(x; \mathcal{F}_p) = g_{\sigma'}(x; \mathcal{F}_p)\,, \tag{120}$$

$$\prod_{l=1}^{d} f_{\sigma(l)|\sigma(<l)}(x_{\sigma(l)}, x_{\sigma(<l)}) = \prod_{l=1}^{d} f_{\sigma'(l)|\sigma'(<l)}(x_{\sigma'(l)}, x_{\sigma'(<l)})\,. \tag{121}$$

Note that the terms for $l \leqslant k$ and $l \geqslant k+3$ are equal on both sides, i.e.,

$$f_{\sigma(l)|\sigma(<l)}(x_{\sigma(l)}, x_{\sigma(<l)}) = f_{\sigma'(l)|\sigma'(<l)}(x_{\sigma'(l)}, x_{\sigma'(<l)})\,, \tag{122}$$

which follows by the definition of σ and σ'. As a result, these terms cancel each other and we are left with

$$\prod_{l=k+1}^{k+2} f_{\sigma(l)|\sigma(<l)}(x_{\sigma(l)}, x_{\sigma(<l)}) = \prod_{l=k+1}^{k+2} f_{\sigma'(l)|\sigma'(<l)}(x_{\sigma'(l)}, x_{\sigma'(<l)})\,, \tag{123}$$

$$f_{i|S}(x_i, x_S) f_{j|S\cup\{i\}}(x_j, x_S \cup \{x_i\}) = f_{j|S}(x_j, x_S) f_{i|S\cup\{j\}}(x_i, x_S \cup \{x_j\})\,. \tag{124}$$

Since this holds for any (S, i, j), we conclude that $\mathcal{F}_p$ is swap consistent.

$\Leftarrow$) Assume that $\mathcal{F}_p$ is swap consistent. We will show that it is also autoregressive path consistent. Our strategy is to show that swap consistency implies that for any permutation σ, if we swap any output of two consecutive indices e.g. $\sigma'(k+1) = \sigma(k), \sigma'(k) = \sigma(k+1)$ and keep the rest of the permutation to be the same, then this new permutation σ' would still lead to the same autoregressive recovery. In particular, for any σ, define σ' as

$$\sigma'(m) = \begin{cases} \sigma(m) & \text{if } m \notin \{k, k+1\}\,, \\ \sigma(k+1) & \text{if } m = k\,, \\ \sigma(k) & \text{if } m = k+1\,. \end{cases} \tag{125}$$

Here, we have that $x_\sigma(m) = x'_\sigma(m)$ for any $m \notin \{k, k+1\}$ and that $x_{<\sigma(m)} = x_{<\sigma'(m)}$ for any $m \neq k+1$. Then,

$$g_{\sigma'}(x; \mathcal{F}_p) = \prod_{l=1}^{d} f_{\sigma'(l)|\sigma'(<l)} \tag{126}$$

$$= \Big(\prod_{l \neq k, k+1} f_{\sigma'(l)|\sigma'(<l)}\Big) \cdot f_{\sigma'(k)|\sigma'(<k)} \cdot f_{\sigma'(k+1)|\sigma'(<k+1)} \tag{127}$$

$$= \Big(\prod_{l \neq k, k+1} f_{\sigma(l)|\sigma(<l)}\Big) \cdot f_{\sigma(k+1)|\sigma(<k)} \cdot f_{\sigma(k)|\sigma(<k)\cup\{\sigma(k+1)\}}\,, \tag{128}$$

where we omitted the input of each function f for simplicity. Note that, by swap consistency, we know that

$$f_{\sigma(k+1)|\sigma(<k)} \cdot f_{\sigma(k)|\sigma(<k)\cup\{\sigma(k+1)\}} = f_{\sigma(k)|\sigma(<k)} \cdot f_{\sigma(k+1)|\sigma(k)\cup\{\sigma(<k)\}} \tag{129}$$

$$= f_{\sigma(k)|\sigma(<k)} \cdot f_{\sigma(k+1)|\sigma(<k+1)}\,. \tag{130}$$

Substituting into (128), we have

$$g_{\sigma'}(x; \mathcal{F}_p) = \Big(\prod_{l \neq k, k+1} f_{\sigma(l)|\sigma(<l)}\Big) \cdot f_{\sigma(k)|\sigma(<k)} \cdot f_{\sigma(k+1)|\sigma(<k+1)} = \prod_{l=1}^{d} f_{\sigma(l)|\sigma(<l)} = g_\sigma(x; \mathcal{F}_p)\,. \tag{131}$$

We can conclude that σ, σ' leads to the same autoregressive recovery. Therefore, swap consistency implies that swapping the output of any two consecutive indices in the permutation does not change the output of the autoregressive recovery of a new permutation. However, we note that we can achieve any permutation from any starting permutation by applying a sequence of these swaps. Thus, every permutation must have the same autoregressive recovery and $\mathcal{F}_p$ must be autoregressive path consistent. □

Proposition B.4. *If $\mathcal{F}_p$ is autoregressive path consistent then $\mathcal{F}_p$ corresponds to a consistent joint distribution.*

Proof. Our proof is divided into two parts; first, we will show that the output of the autoregressive recovery $q(x) = g_\sigma(x; \mathcal{F}_p)$ is a valid distribution, second, we will show that q with consistent with any conditionals $f_{i|S}(x_i, x_S)$. Let σ be an identity permutation where $\sigma(i) = i$. The autoregressive path recovery is given by

$$g_\sigma(x; \mathcal{F}_p) = \prod_{l=1}^{d} f_{l|<l}(x_l, x_{<l})\ . \tag{132}$$

We denote this as $q(x)$. We will show that q is a valid joint distribution by showing that $q(x) > 0$ for any x and the sum of $q(x)$ over all possible values of x is 1 (the similar result also holds for the continuous setting where we replace the sum with an integration). The first point is trivial since $f_{l|<l}(x_l, x_{<l}) \geqslant 0$ for any l since $f_{l|<l}$ is a conditional distribution. On the second point,

$$\sum_x q(x) = \sum_{x_1, x_2, \ldots, x_d} f_{1|\emptyset}(x_1)\, f_{2|1}(x_2, x_1) \cdots f_{d|<d}(x_d, x_{<d}) \tag{133}$$

$$= \sum_{x_1} f_{1|\emptyset}(x_1) \Bigg(\sum_{x_2} f_{2|1}(x_2, x_1) \cdots \Bigg(\sum_{x_{d-1}} f_{d-1|<d-1}(x_{d-1}, x_{<d-1}) \Big(\sum_{x_d} f_{d|<d}(x_d, x_{<d}) \Big) \cdots \Bigg) \cdots \Bigg)\,. \tag{134}$$

Since $f_{l|<l}$ are conditional distributions, we know that for any fixed $\bar{x}_{<l}$, we have $\sum_{x_l} f_{l|<l}(x_l, \bar{x}_{<l}) = 1$. Therefore, we can start from the deepest term in the bracket, $\sum_{x_d} f_{d|<d}(x_d, x_{<d}) = 1$, and then work backward. Now, we have

$$\sum_x q(x) = \sum_{x_1} f_{1|\emptyset}(x_1) \Big(\sum_{x_2} f_{2|1}(x_2, x_1) \cdots \Big(\sum_{x_{d-1}} f_{d-1|<d-1}(x_{d-1}, x_{<d-1}) \Big) \cdots \Big) . \tag{135}$$

Using the same argument, the summation over x_{d-1} for each fixed $\bar{x}_{<d-1}$ would be 1 and we can keep doing this until we reach the first term. Thus, we must have

$$\sum_x q(x) = 1 , \tag{136}$$

which concludes that $q(x)$ is a valid joint distribution. Next, we will show that q is consistent with any conditional $f_{i|S} \in \mathcal{F}_p$, that is, we need to show that for any x_i, x_S

$$f_{i|S}(x_i, x_S) = q(x_i, x_S) . \tag{137}$$

First, we note that the conditional $q(x_i|x_S)$ is defined as

$$q(x_i|x_S) = \frac{q(x_{S\cup\{i\}})}{q(x_S)} , \tag{138}$$

where the marginal $q(x_S)$ is computed by

$$q(x_S) = \sum_{x_{[d]\setminus S}} q(x_S, x_{[d]\setminus S}) . \tag{139}$$

Denote the set $S = \{s_1, s_2, \ldots, s_k\}$. Construct a permutation $\sigma : [d] \to [d]$ such that i) the first k elements are the elements of S, i.e. $\sigma(i) = s_i$ for all $i \in [k]$, ii) the next element is i, $\sigma(k+1) = i$, iii) the remaining elements are the indices not in $S \cup \{i\}$, denoted by $T = [d] \setminus (S \cup \{i\})$, in any order. With this permutation, we can express $q(x_{S\cup\{i\}})$ as

$$\begin{aligned} q(x_{S\cup\{i\}}) &= \sum_{x_T} q(x) && (140)\\ &= \sum_{x_T} \Big(\prod_{l=1}^{d} f_{l|<l}(x_l, x_{<l}) \Big) && (141)\\ &= \sum_{x_T} \Big(\prod_{l=1}^{d} f_{\sigma(l)|<\sigma(<l)}(x_{\sigma(l)}, x_{\sigma(<l)}) \Big) , && (142) \end{aligned}$$

where x_T denotes the variables x_j for $j \in T$, and $\sigma(< l) = \{\sigma(1), \sigma(2), \ldots, \sigma(l-1)\}$. Here, we use the autoregressive path consistency property to ensure that the term above equals to $q(x)$ defined earlier with an identity permutation. Breaking down the product inside the summation, we have

$$\prod_{l=1}^{d} f_{\sigma(l)|<\sigma(<l)} = \Big(\prod_{l=1}^{k} f_{\sigma(l)|<\sigma(<l)} \Big) \cdot f_{i|S} \cdot \Big(\prod_{l=k+2}^{d} f_{\sigma(l)|<\sigma(<l)} \Big) , \tag{143}$$

where we drop the input of each function for simplicity. We note that the first term only involves variables in S while the second term is $f_{i|S}$ so that we can factorize the first and the second term from the sum,

$$\begin{aligned} q(x_{S\cup\{i\}}) &= \sum_{x_T} \left(\prod_{l=1}^{d} f_{\sigma(l)|<\sigma(<l)}(x_{\sigma(l)}, x_{\sigma(<l)}) \right) && (144)\\ &= \left(\prod_{l=1}^{k} f_{\sigma(l)|<\sigma(<l)} \right) \cdot f_{i|S} \cdot \sum_{x_T} \left(\prod_{l=k+2}^{d} f_{\sigma(l)|<\sigma(<l)} \right) . && (145) \end{aligned}$$

Similarly to when we were showing that the sum of $q(x)$ over all possible x is one, we will show that

$$\sum_{x_T} \Big(\prod_{l=k+2}^{d} f_{\sigma(l)|<\sigma(<l)} \Big) = 1 \,. \tag{146}$$

We can see this by expanding it as

$$\sum_{x_{\sigma(k+2)}} f_{\sigma(k+2)|\sigma(<k+2)}(x_{\sigma(k+2)}, x_{\sigma(<k+2)}) \Bigg(\cdots \Bigg(\cdots \sum_{x_{\sigma(d)}} f_{\sigma(d)|\sigma(<d)}(x_{\sigma(d)}, x_{\sigma(<d)}) \Bigg) \cdots \Bigg) \,, \tag{147}$$

where we can use the same argument that for any l, and any fixed $\bar{x}_{\sigma(<l)}$,

$$\sum_{x_{\sigma(l)}} f_{\sigma(l)|\sigma(<l)}(x_{\sigma(l)}, \bar{x}_{\sigma(<l)})) = 1 \,. \tag{148}$$

Finally, we conclude that

$$q(x_{S\cup\{i\}}) = \left(\prod_{l=1}^{k} f_{\sigma(l)|<\sigma(<l)}(x_{\sigma(l)}, x_{\sigma(<l)}) \right) \cdot f_{i|S}(x_i, x_S) \,. \tag{149}$$

Similarly, to compute the marginal $q(x_S)$, we sum over both x_i and x_T:

$$\begin{aligned} q(x_S) &= \sum_{x_i} \sum_{x_T} q(x) && (150)\\ &= \sum_{x_i} \left(\prod_{l=1}^{k} f_{\sigma(l)|<\sigma(<l)}(x_{\sigma(l)}, x_{\sigma(<l)}) \right) \cdot f_{i|S}(x_i, x_S) && (151)\\ &= \left(\prod_{l=1}^{k} f_{\sigma(l)|<\sigma(<l)}(x_{\sigma(l)}, x_{\sigma(<l)}) \right) \cdot \sum_{x_i} f_{i|S}(x_i, x_S) && (152)\\ &= \left(\prod_{l=1}^{k} f_{\sigma(l)|<\sigma(<l)}(x_{\sigma(l)}, x_{\sigma(<l)}) \right) . && (153) \end{aligned}$$

Again, the final line holds from the fact that for a fixed $\bar{x}_S$, $\sum_{x_i} f_{i|S}(x_i, \bar{x}_S) = 1$. Now, we can compute the conditional probability $q(x_i|x_S)$:

$$q(x_i|x_S) = \frac{q(x_{S\cup\{i\}})}{q(x_S)} = f_{i|S}(x_i, x_S) \,.$$

Thus, we have shown that $q(x_i|x_S) = f_{i|S}(x_i, x_S)$ for all i and S, confirming that the constructed q satisfies the required consistency condition. □

B.3 Proof of Proposition 4.7

Proposition 4.7. *The set of functions* $\{f_{i|S}(x_i, x_S) \mid S \subseteq [d], i \in [d], i \notin S\}$ *for a discrete set* $\mathcal{X}$ *with* $|\mathcal{X}| = K$ *can be equivalently written as a set function* $\rho : \mathcal{C} \to \mathbb{R}^K$ *where:*

$$\rho(C(x, i, S))_v = f_{i|S}(x_i = v, x_S) \,.$$

Proof. Recall the definition of C in (21) as :

$$C(x, i, S) := \{(i, j, x_j) | j \in S\} \cup \{(i, j) | j \notin S\} \,. \tag{154}$$

First, observe that $C(x^1, i, S) = C(x^2, i, S)$ for all x_1, x_2, such that $(x_1)_S = (x_2)_S$. Thus, $C(x, i, S)$ only depends on (x_S, i, S). Hence, let us use $C(x_S, i, S)$ for $C(x, i, S)$.

Next, define $g(x_S, i, S) \in \mathbb{R}^K$ as

$$g(x_S, i, S)_v := f_{i|S}(x_i = v, x_S)\,. \tag{155}$$

We can readily see that g uniquely defines $\{f_{i|S}\}$. Now we will prove that $C(x_S, i, S)$ uniquely defines (x_S, i, S), i.e. there exists C^{-1} such that $C^{-1}C(x_S, i, S) = (x_S, i, S)$. Suppose on the contrary that there exists two tuples $(x^1_{S^1}, i^1, S^1), (x^2_{S^2}, i^2, S^2)$ such that $C(x^1_{S^1}, i^1, S^1) = C(x^2_{S^2}, i^2, S^2)$. Since $\{i : (i, j) \in C(x_S, i, S)\} = \{i\}$, $\{j : (i, j) \in C(x_S, i, S)\} = [d]/S$, thus $i^1 = i^2$ and $S^1 = S^2$. Finally $\{(j, x_j) : (i, j, x_j) \in C(x_S, i, S)\} = x_S$, thus $x^1_S = x^2_S$. Hence there exists $\rho = g \cdot C^{-1}$ such that $\rho(C(x, i, S))_v = \rho(C(x, i, S))_v = f_{i|S}(x_i = v, x_S)$.

□

B.4 Proof of Proposition 4.8

Proposition 4.8. *For any set of conditionals $\{f_{i|S}(x_i, x_S) \mid S \subset [d], i \in [d], i \notin S\}$ there exists some $g : \mathbb{R}^D \to \mathbb{R}^K_+$ and $w_{i,j,k}, w_{i,j} \in \mathbb{R}^D$ for $i, j \in [d]$, $k \in \mathcal{X}$ such that*

$$f_{i|S}(x_i = v, x_S) = g(\psi_i(x_S))_v\,, \tag{22}$$

where $\psi_i(x_S) = \sum_{j \in S} w_{i,j,x_j} + \sum_{j \notin S} w_{i,j}$.

Proof. From Proposition 4.7, there exists $\rho : \mathcal{C} \to \mathbb{R}^K$ where $\rho(C(x, i, S))_v = f_{i|S}(x_i = v, x_S)$. Then, using the result in Zaheer et al. (2017, Theorem 2), ρ can be decomposed as $\rho(C) = g(\sum_{c \in C} \phi(c))$ for some function g. Letting $\phi(c) = w_c \in \mathbb{R}^D$, we have $\rho(C) = g(\sum_{c \in C} w_c)$. Since $C(x_S, i, S) = \{(i, j, x_j) | j \in S\} \cup \{(i, j) | j \notin S\}$, we have $\rho(C(x_S, i, S)) = g(\sum_{j \in S} w_{i,j,x_j} + \sum_{j \notin S} w_{i,j})$ which concludes the proof. □

B.5 Proof of Theorem 4.9

Theorem B.5. *Let $g(z)_v = \exp(w_v^\top z)/\sum_{k=1}^K \exp(w_k^\top z)$ for some parameters $w_k \in \mathbb{R}^D$ for each $k \in [K]$ where $K = 2$. The set of conditionals $\{f_{i|S}(x_i, x_S) \mid S \subset [d], i \in [d], i \notin S\}$ parameterized by*

$$f_{i|S}(x_i = k, x_S) = g(\psi_i(x_S))_k\,, \tag{156}$$

where

$$\psi_i(x_S) = \sum_{j \in S} z_{i,j,x_k} + \sum_{j \notin S} z_{i,j}\,, \tag{157}$$

correspond to a consistent joint distribution only when the joint distribution can be factorized into independent subsets of variables of sizes 1, 2, or at least 3 where for subsets of size at least 3, the distribution has to be simple and depends only on the counts of the different values in the subset. Formally, $p(x)$ can be factorized as

$$p(x) = \prod_{\{v\} \in V} p(x_v) \prod_{\{u,v\} \in E} p(x_u, x_v) \prod_{C \in \mathcal{C}} p(x_C)\,. \tag{158}$$

Here $V, E, \mathcal{C}$ are partitions of the index set $[d]$ into a subset of size 1,2 and at least 3 respectively

$$[d] = \bigcup_{v \in V} \{v\} \cup \bigcup_{\{u,v\} \in E} \{u, v\} \cup \bigcup_{C \in \mathcal{C}} C\,. \tag{159}$$

For any $C \in \mathcal{C}$, the distribution $p(x_C)$ has a specific simple form given by

$$p(x_C) = f_C(\mathrm{count}(x_C, 1), \ldots \mathrm{count}(x_C, K))\,, \tag{160}$$

for some function f_C and $\mathrm{count}(x_C, k) = \sum_{c \in C} \mathbb{I}(x_c = k)$. As a result, the variables within C are interchangeable.

Proof. Assume that the set of conditionals is parameterized as in equations (156), (157). corresponds to a consistent joint distribution p. Then for any i, S we have

$$f_{i|S}(x_i, x_S) = p(x_i \mid x_S)\,, \tag{161}$$

We will show that $p(x)$ can be factorized as in the Theorem above. We break down our proof into 2 parts,

1. **Step 1:** By using the swap consistency condition and solving the equations with this specific parameterization, we can show that for any pair $i, j \in [d]$, one of the following must hold
   (a) x_i, x_j are independent to each other
   (b) x_i, x_j are independent of other variables x_S
   (c) x_i, x_j can be exchanged.
2. **Step 2:** We will show that the conditions from Step 1 implies that we can factorize $p(x)$ as desired.

Step 1: Deriving the conditions for any pair (i, j)

We first start with writing $p(x_i \mid x_S)$ in terms of our parameterization. Recall that we parameterize

$$\psi_i(x_S) = \sum_{j \in S} z_{i,j,x_k} + \sum_{j \notin S} z_{i,j}\,, \tag{162}$$

so we have

$$w_k^\top \psi_i(x_S) = w_k^\top \big(\sum_{j \in S} z_{i,j,x_j} + \sum_{j \notin S} z_{i,j} \big) \tag{163}$$

$$= \sum_{j \in S} w_k^\top z_{i,j,x_j} + \sum_{j \notin S} w_k^\top z_{i,j} \tag{164}$$

$$= \sum_{j \in S} w_k^\top (z_{i,j,x_j} - z_{i,j}) + \sum_{j=1}^{d} w_k^\top z_{i,j}\,. \tag{165}$$

Let $v_{i,j,x_j}^k = w_k^\top (z_{i,j,x_j} - z_{i,j})$ and $v_i^k = \sum_{j=1}^d w_k^\top z_{i,j}$, we also denote $\phi_i(x_S)_k = w_k^\top \psi_i(x_S)$. Then, we have

$$\phi_i(x_S)_k = w_k^\top \psi_i(x_S) = \sum_{j \in S} v_{i,j,x_j}^k + v_i^k\,. \tag{166}$$

Thus $p(x_i|x_S)$ is given as

$$p(x_i|x_S) = \frac{\exp(\phi_i(x_S)_{x_i})}{\sum_{k=1}^K \exp(\phi_i(x_S)_k)} = \frac{\exp(\sum_{j \in S} v_{i,j,x_j}^{x_i} + v_i^{x_i})}{\sum_{k=1}^K \exp(\sum_{j \in S} v_{i,j,x_j}^k + v_i^k)}\,. \tag{167}$$

For consistency, using the swap consistency condition (Theorem 4.6), we require that

$$p(x_i, x_j|x_S) = p(x_j|x_S, x_i)p(x_i|x_S) = p(x_i|x_S, x_j)p(x_i|x_S)\,. \tag{168}$$

Substituting the parameterization of the conditionals above into these equations, we have

$$\text{LHS} : p(x_j \mid x_S, x_i) \times p(x_i \mid x_S) = \left[\frac{\exp\left(\phi_j(x_S)_{x_j} + v_{j,i,x_i}^{x_j}\right)}{\sum_{k=1}^K \exp\left(\phi_j(x_S)_k + v_{j,i,x_i}^k\right)} \right] \times \left[\frac{\exp\left(\phi_i(x_S)_{x_i}\right)}{\sum_{k=1}^K \exp\left(\phi_i(x_S)_k\right)} \right], \tag{169}$$

$$\text{RHS} : p(x_i \mid x_S, x_j) \times p(x_j \mid x_S) = \left[\frac{\exp\left(\phi_i(x_S)_{x_i} + v_{i,j,x_j}^{x_i}\right)}{\sum_{k=1}^K \exp\left(\phi_i(x_S)_k + v_{i,j,x_j}^k\right)} \right] \times \left[\frac{\exp\left(\phi_j(x_S)_{x_j}\right)}{\sum_{k=1}^K \exp\left(\phi_j(x_S)_k\right)} \right]. \tag{170}$$

By the equality of these two expressions, we get

$$\left[\frac{\exp\left(\phi_j(x_S)_{x_j} + v_{j,i,x_i}^{x_j}\right)}{\sum_{k=1}^K \exp\left(\phi_j(x_S)_k + v_{j,i,x_i}^k\right)} \right] \times \left[\frac{\exp\left(\phi_i(x_S)_{x_i}\right)}{\sum_{k=1}^K \exp\left(\phi_i(x_S)_k\right)} \right] \tag{171}$$

$$= \left[\frac{\exp\left(\phi_i(x_S)_{x_i} + v_{i,j,x_j}^{x_i}\right)}{\sum_{k=1}^K \exp\left(\phi_i(x_S)_k + v_{i,j,x_j}^{k}\right)}\right] \times \left[\frac{\exp\left(\phi_j(x_S)_{x_j}\right)}{\sum_{k=1}^K \exp\left(\phi_j(x_S)_k\right)}\right]. \tag{172}$$

Rearranging:

$$\frac{\exp\left(v_{j,i,x_i}^{x_j}\right)}{\left(\sum_{k=1}^K \exp\left(\phi_j(x_S)_k + v_{j,i,x_i}^{k}\right)\right) \times \left(\sum_{k=1}^K \exp\left(\phi_i(x_S)_k\right)\right)} \tag{173}$$

$$= \frac{\exp(v_{i,j,x_j}^{x_i})}{\left(\sum_{k=1}^K \exp(\phi_i(x_S)_k + v_{i,j,x_j}^{k})\right) \times \left(\sum_{k=1}^K \exp(\phi_j(x_S)_k)\right)} \tag{174}$$

Cross-multiplying the denominator:

$$\frac{\left(\sum_{k=1}^K \exp\left(\phi_j(x_S)_k + v_{j,i,x_i}^{k}\right)\right) \times \left(\sum_{k=1}^K \exp\left(\phi_i(x_S)_k\right)\right)}{\exp\left(v_{j,i,x_i}^{x_j}\right)} \tag{175}$$

$$= \frac{\left(\sum_{k=1}^K \exp(\phi_i(x_S)_k + v_{i,j,x_j}^{k})\right) \times \left(\sum_{k=1}^K \exp(\phi_j(x_S)_k)\right)}{\exp(v_{i,j,x_j}^{x_i})} \tag{176}$$

Simplifying:

$$\left(\sum_{k=1}^K \exp\left(\phi_j(x_S)_k + v_{j,i,x_i}^{k} - v_{j,i,x_i}^{x_j}\right)\right) \times \left(\sum_{k=1}^K \exp(\phi_i(x_S)_k)\right) \tag{177}$$

$$= \left(\sum_{k=1}^K \exp((\phi_i(x_S)_k + v_{i,j,x_j}^{k} - v_{i,j,x_j}^{x_i}))\right) \times \left(\sum_{k=1}^K \exp(\phi_j(x_S)_k)\right) \tag{178}$$

$$\frac{\left(\sum_{k=1}^K \exp\left(\phi_j(x_S)_k + v_{j,i,x_i}^{k} - v_{j,i,x_i}^{x_j}\right)\right)}{\sum_{k=1}^K \exp(\phi_j(x_S)_k)} = \frac{\left(\sum_{k=1}^K \exp\left(\phi_i(x_S)_k + v_{i,j,x_j}^{k} - v_{i,j,x_j}^{x_i}\right)\right)}{\sum_{k=1}^K \exp(\phi_i(x_S)_k)} \tag{179}$$

Expanding $\phi_j(x_S), \phi_i(x_S)$ the conditions for consistency are summarized as follows.

For all $i, j \in [d], i \neq j$**,** $x_i, x_j \in [K]$**,** $S \subseteq [d]/\{i,j\}$**,** $x_S \in [K]^{|S|}$**, we require**

$$\frac{\sum_{k=1}^K \exp\left(\sum_{m\in S} v_{j,m,x_m}^{k} + v_j^k + v_{j,i,x_i}^{k} - v_{j,i,x_i}^{x_j}\right)}{\sum_{k=1}^K \exp(\sum_{m\in S} v_{j,m,x_m}^{k} + v_j^k)} = \frac{\sum_{k=1}^K \exp\left(\sum_{m\in S} v_{i,m,x_m}^{k} + v_i^k + v_{i,j,x_j}^{k} - v_{i,j,x_j}^{x_i}\right)}{\sum_{k=1}^K \exp(\sum_{m\in S} v_{i,m,x_m}^{k} + v_i^k)}. \tag{180}$$

For the rest of the proof, we consider the binary variable setting, $K = 2$**.**

Notations : *In the rest of the proof, the following notations will be used frequently.*

$$u_{i,j,k} = v_{i,j,k}^2 - v_{i,j,k}^1 \qquad u_i = v_i^2 - v_i^1 \tag{181}$$

$$y_k = \mathbb{I}(x_k = 1) \qquad y'_k = \mathbb{I}(x_k = 2) = 1 - y_k \tag{182}$$

$$\alpha_{j,k} = u_{j,k,1} - u_{j,k,2} \qquad \nu_{j,S} = \sum_{k\in S} u_{j,k,2} + u_j \tag{183}$$

Using $x_i = 1, x_j = 1$, the conditions for consistency, Equation (179) becomes

$$\frac{\exp\left(\phi_j(x_S)_2 + v_{j,i,1}^2 - v_{j,i,1}^1\right) + \exp\left(\phi_j(x_S)_1\right)}{\exp(\phi_j(x_S)_1) + \exp(\phi_j(x_S)_2)} = \frac{\exp\left(\phi_i(x_S)_2 + v_{i,j,1}^2 - v_{i,j,1}^1\right) + \exp((\phi_i(x_S)_1)}{\exp(\phi_i(x_S)_1) + \exp(\phi_i(x_S)_2)}, \tag{184}$$

$$\frac{\exp\left(\phi_j(x_S)_2 - \phi_j(x_S)_1 + v_{j,i,1}^2 - v_{j,i,1}^1\right) + 1}{\exp(\phi_j(x_S)_2 - \phi_j(x_S)_1) + 1} = \frac{\exp\left(\phi_i(x_S)_2 - \phi_i(x_S)_1 + v_{i,j,1}^2 - v_{i,j,1}^1\right) + 1}{\exp(\phi_i(x_S)_2 - \phi_i(x_S)_1) + 1}. \tag{185}$$

Next, let us use the notations in (181) and (182), and substitute ϕ from Equation (166) into Equation (185), which gives

$$\frac{\exp\left(\sum_{k\in S} u_{j,k,1}y_k + u_{j,k,2}y'_k + u_{j,i,1} + u_j\right) + 1}{\exp(\sum_{k\in S} u_{j,k,1}y_k + u_{j,k,2}y'_k + u_j) + 1} = \frac{\exp\left(\sum_{k\in S} u_{i,k,1}y_k + u_{i,k,2}y'_k + u_i + u_{i,j,1}\right) + 1}{\exp(\sum_{k\in S} u_{i,k,1}y_k + u_{i,k,2}y'_k + u_i) + 1}\,, \tag{186}$$

$$\frac{\exp\left(\sum_{k\in S}(u_{j,k,1} - u_{j,k,2})y_k + u_{j,k,2} + u_j + u_{j,i,1}\right) + 1}{\exp(\sum_{k\in S}(u_{j,k,1} - u_{j,k,2})y_k + u_{j,k,2} + u_j) + 1} = \frac{\exp\left(\sum_{k\in S}(u_{i,k,1} - u_{i,k,2})y_k + u_{i,k,2} + u_i + u_{i,j,1}\right) + 1}{\exp(\sum_{k\in S}(u_{i,k,1} - u_{i,k,2})y_k + u_{i,k,2} + u_i) + 1}\,. \tag{187}$$

Finally, let us use the notation (183), $\alpha_{j,k} = u_{j,k,1} - u_{j,k,2}$ and $\nu_{j,S} = \sum_{k\in S} u_{j,k,2} + u_j$, to rewrite

$$\frac{\exp\left(\sum_{k\in S}\alpha_{j,k}y_k + \nu_{j,S} + u_{j,i,1}\right) + 1}{\exp\left(\sum_{k\in S}\alpha_{j,k}y_k + \nu_{j,S}\right) + 1} = \frac{\exp\left(\sum_{k\in S}\alpha_{i,k}y_k + \nu_{i,S} + u_{i,j,1}\right) + 1}{\exp\left(\sum_{k\in S}\alpha_{i,k}y_k + \nu_{i,S}\right) + 1}\,. \tag{188}$$

Subtracting 1 from both sides and rearranging, we get

$$\frac{\exp(\sum_{k\in S}\alpha_{j,k}y_k + \nu_{j,S})(\exp(u_{j,i,1}) - 1)}{\exp(\sum_{k\in S}\alpha_{j,k}y_k + \nu_{j,S}) + 1} = \frac{\exp(\sum_{k\in S}\alpha_{i,k}y_k + \nu_{i,S})(\exp(u_{i,j,1}) - 1)}{\exp(\sum_{k\in S}\alpha_{i,k}y_k + \nu_{i,S}) + 1}\,, \tag{189}$$

$$\frac{\exp(u_{j,i,1}) - 1}{\exp(-\sum_{k\in S}\alpha_{j,k}y_k - \nu_{j,S}) + 1} = \frac{\exp(u_{i,j,1}) - 1}{\exp(-\sum_{k\in S}\alpha_{i,k}y_k - \nu_{i,S}) + 1}\,. \tag{190}$$

We can perform similar calculations as above for different values of x_i, x_j in Equation (184), to obtain the following conditions.

1. $x_i = 1$ and $x_j = 1$:

$$\frac{\exp(u_{j,i,1}) - 1}{\exp(-\sum_{k\in S}\alpha_{j,k}y_k - \nu_{j,S}) + 1} = \frac{\exp(u_{i,j,1}) - 1}{\exp(-\sum_{k\in S}\alpha_{i,k}y_k - \nu_{i,S}) + 1}\,. \tag{191}$$

2. $x_i = 1$ and $x_j = 2$:

$$\frac{\exp(-u_{j,i,1}) - 1}{\exp(\sum_{k\in S}\alpha_{j,k}y_k + \nu_{j,S}) + 1} = \frac{\exp(u_{i,j,2}) - 1}{\exp(-\sum_{k\in S}\alpha_{i,k}y_k - \nu_{i,S}) + 1}\,. \tag{192}$$

3. $x_i = 2$ and $x_j = 1$:

$$\frac{\exp(u_{j,i,2}) - 1}{\exp(-\sum_{k\in S}\alpha_{j,k}y_k - \nu_{j,S}) + 1} = \frac{\exp(-u_{i,j,1}) - 1}{\exp(\sum_{k\in S}\alpha_{i,k}y_k + \nu_{i,S}) + 1}\,. \tag{193}$$

4. $x_i = 2$ and $x_j = 2$:

$$\frac{\exp(-u_{j,i,2}) - 1}{\exp(\sum_{k\in S}\alpha_{j,k}y_k + \nu_{j,S}) + 1} = \frac{\exp(-u_{i,j,2}) - 1}{\exp(\sum_{k\in S}\alpha_{i,k}y_k + \nu_{i,S}) + 1}\,. \tag{194}$$

Setting 1: $u_{i,j,1} = 0 \implies (i,j)$ **are independent.**

If $u_{j,i,1} = 0$, then from Equation (191), we would have $u_{i,j,1} = 0$, and from Equation (192), we have $u_{i,j,2} = 0$, which then from Equation (194) implies $u_{j,i,2} = 0$ as well. Then, we would have

$$p(x_i|x_S, x_j) = p(x_i|x_S) \quad \text{and} \quad p(x_j|x_S, x_i) = p(x_j|x_S)\,. \tag{195}$$

In other words, i, j are independent. We denote this by $i \perp j$.

Setting 2: (i,j) **are not independent,** $i \not\perp j$.

Now, we focus on pairs (i,j) that are not independent, and $u_{i,j,1}, u_{j,i,1} \neq 0$. Denoting

$$\mu_{i,j} = \frac{\exp(u_{j,i,1}) - 1}{\exp(u_{i,j,1}) - 1}\,, \tag{196}$$

Equation (191) gives

$$\exp\Big(-\sum_{k\in S}\alpha_{j,k}y_k - \nu_{j,S}\Big) + 1 = \mu_{i,j}\cdot\Big(\exp\Big(-\sum_{k\in S}\alpha_{i,k}y_k - \nu_{i,S}\Big) + 1\Big)\,. \tag{197}$$

Substituting $y_k = 0$ in Equation (197) for all k we get

$$\exp(-\nu_{j,S}) + 1 = \mu_{i,j}(\exp(-\nu_{i,S}) + 1)\,. \tag{198}$$

We have

$$\exp(-\nu_{i,S}) = \frac{\exp(-\nu_{j,S}) + 1}{\mu_{i,j}} - 1\,. \tag{199}$$

Substituting $y_k = 1$ in Equation (197) for any k, we get

$$\exp(-\alpha_{j,k} - \nu_{j,S}) + 1 = \mu_{i,j}(\exp(-\alpha_{i,k} - \nu_{i,S}) + 1)\,. \tag{200}$$

Substituting $\nu_{i,S}$ in terms of $\nu_{j,S}$ from Equation (198), this is simplified as

$$\exp(-\alpha_{j,k} - \nu_{j,S}) + 1 = \mu_{i,j}\Big(\exp(-\alpha_{i,k})(\frac{\exp(-\nu_{j,S}) + 1}{\mu_{i,j}} - 1) + 1\Big)\,, \tag{201}$$

$$\exp(-\alpha_{j,k} - \nu_{j,S}) + 1 - \mu_{i,j} = \exp(-\alpha_{i,k})(\exp(-\nu_{j,S}) + 1 - \mu_{i,j})\,, \tag{202}$$

$$\exp(-\alpha_{i,k}) = \frac{\exp(-\alpha_{j,k} - \nu_{j,S}) + 1 - \mu_{i,j}}{\exp(-\nu_{j,S}) + 1 - \mu_{i,j}}\,. \tag{203}$$

Substituting $y_k = 1, y_l = 1$ in Equation (197) for any k, l we get

$$\exp(-\alpha_{j,k} - \alpha_{j,l} - \nu_{j,S}) + 1 = \mu_{i,j}(\exp(-\alpha_{i,k} - \alpha_{i,l} - \nu_{i,S}) + 1)\,. \tag{204}$$

Substituting $\nu_{i,S}$ from Equation (198) we get

$$\exp(-\alpha_{j,k} - \alpha_{j,l} - \nu_{j,S}) + 1 = \mu_{i,j}\exp(-\alpha_{i,k} - \alpha_{i,l})\Big(\frac{\exp(-\nu_{j,S}) + 1}{\mu_{i,j}} - 1\Big) + \mu_{i,j}\,, \tag{205}$$

$$\exp(-\alpha_{j,k} - \alpha_{j,l} - \nu_{j,S}) + 1 - \mu_{i,j} = \exp(-\alpha_{i,k} - \alpha_{i,l})(\exp(-\nu_{j,S}) + 1 - \mu_{i,j})\,. \tag{206}$$

Substituting $\alpha_{i,k}, \alpha_{i,l}$ from Equation (203) and letting $\bar{\mu}_{i,j} = 1 - \mu_{i,j}$, we have

$$\exp(-\alpha_{j,k} - \alpha_{j,l} - \nu_{j,S}) + \bar{\mu}_{i,j} = \exp(-\alpha_{i,k} - \alpha_{i,l})(\exp(-\nu_{j,S}) + \bar{\mu}_{i,j}) \tag{207}$$

$$= \frac{(\exp(-\alpha_{j,k} - \nu_{j,S}) + \bar{\mu}_{i,j})(\exp(-\alpha_{j,l} - \nu_{j,S}) + \bar{\mu}_{i,j})}{\exp(-\nu_{j,S}) + \bar{\mu}_{i,j}} \tag{208}$$

$$(\exp(-\nu_{j,S}) + \bar{\mu}_{i,j})(\exp(-\alpha_{j,k} - \alpha_{j,l} - \nu_{j,S}) + \bar{\mu}_{i,j}) = (\exp(-\alpha_{j,k} - \nu_{j,S}) + \bar{\mu}_{i,j})(\exp(-\alpha_{j,l} - \nu_{j,S}) + \bar{\mu}_{i,j}) \tag{209}$$

$$\bar{\mu}_{i,j}\exp(-\nu_{j,S})(\exp(-\alpha_{j,k} - \alpha_{j,l}) + 1) = \bar{\mu}_{i,j}\exp(-\nu_{j,S})(\exp(-\alpha_{j,k}) + \exp(-\alpha_{j,l})) \tag{210}$$

$$\bar{\mu}_{i,j}(\exp(-\alpha_{j,k} - \alpha_{j,l}) + 1) = \bar{\mu}_{i,j}(\exp(-\alpha_{j,k}) + \exp(-\alpha_{j,l})) \tag{211}$$

$$\bar{\mu}_{i,j}(1 - \exp(-\alpha_{j,k}) - \exp(-\alpha_{j,l}) + \exp(-\alpha_{j,k} - \alpha_{j,l})) = 0 \tag{212}$$

$$\bar{\mu}_{i,j}(1 - \exp(-\alpha_{j,k}))(1 - \exp(-\alpha_{j,l})) = 0 \tag{213}$$

$$(1 - \mu_{i,j})(1 - \exp(-\alpha_{j,k}))(1 - \exp(-\alpha_{j,l})) = 0\,. \tag{214}$$

This implies $\mu_{i,j} = 1$ or $\alpha_{j,k} = 0$ or $\alpha_{j,l} = 0$. $\mu_{i,j} = 1$ is equivalent to $u_{j,i,1} = u_{i,j,1}$. Substituting $\alpha_{j,k} = 0$ in Equation (203) we also get $\alpha_{i,k} = 0$. Since this holds for all pairs $k, l \in S, k \neq l$ and for all S, at least one of the following holds:

$$u_{j,i,1} = u_{i,j,1} \quad \text{or} \quad \alpha_{j,k} = \alpha_{i,k} = 0 \text{ for all } k \in [d]/\{i,j,l\} \text{ for some } l \in [d]\,. \tag{215}$$

We used the consistency conditions in Equation (191) to derive the above. If we use conditions in Equation (194), with the exact same calculation, we can also derive at least one of the following to hold true:

$$u_{j,i,2} = u_{i,j,2} \quad \text{or} \quad \alpha_{j,k} = \alpha_{i,k} = 0 \text{ for all } k \in [d]/\{i,j,l\} \text{ for some } l \in [d]\,. \tag{216}$$

Combining the two, we can say at least one of the following holds true

$$1). \quad u_{j,i,1} = u_{i,j,1} \text{ and } u_{j,i,2} = u_{i,j,2} \tag{217}$$

$$2). \quad \alpha_{j,k} = \alpha_{i,k} = 0 \text{ for all } k \in [d]/\{i,j,l\} \text{ for some } l \in [d]\,. \tag{218}$$

Case 1: $u_{j,i,1} \neq u_{i,j,1}$**,** $\exists\ k$ **which** $\alpha_{j,k}, \alpha_{i,k} \geqslant 0$**, and** $\forall l \in [d] \setminus \{i,j,k\}$**,** $\alpha_{j,l}, \alpha_{i,l} = 0$

Using Equations (198), and (200), we have

$$\mu_{ij} \exp(-\nu_{i,S})(\exp(-\alpha_{i,k}) - \exp(-\alpha_{j,k})) = (\mu_{ij} - 1)(\exp(-\alpha_{j,k}) - 1)\,. \tag{219}$$

Since this equation holds for all S, $\nu_{i,S}$ has to be independent of S, and thus $\nu_{i,S} = u_i$ and $u_{i,l,2} = 0$ for all $l \in [d]/\{i,j\}$. Then, using Equation (198), we also have $\nu_{j,S} = u_j$ and $u_{j,l,2} = 0$ for all $l \in [d]/\{i,j\}$.

Next, revisiting the consistency conditions Equations (191), and (192) by substituting $\nu_{j,S} = u_j, \nu_{i,S} = u_i$, and $\alpha_{j,k'} = 0$ for $k' \neq k$, we have

1. $x_i = 1$ and $x_j = 1$:

$$\frac{\exp(u_{j,i,1}) - 1}{\exp(-\alpha_{j,k} - u_j) + 1} = \frac{\exp(u_{i,j,1}) - 1}{\exp(-\alpha_{i,k} - u_i) + 1}\,, \tag{220}$$

$$\frac{\exp(u_{j,i,1}) - 1}{\exp(-u_j) + 1} = \frac{\exp(u_{i,j,1}) - 1}{\exp(-u_i) + 1}\,. \tag{221}$$

2. $x_i = 1$ and $x_j = 2$:

$$\frac{\exp(-u_{j,i,1}) - 1}{\exp(\alpha_{j,k} + u_j) + 1} = \frac{\exp(u_{i,j,2}) - 1}{\exp(-\alpha_{i,k} - u_i) + 1}\,, \tag{222}$$

$$\frac{\exp(-u_{j,i,1}) - 1}{\exp(u_j) + 1} = \frac{\exp(u_{i,j,2}) - 1}{\exp(-u_i) + 1}\,. \tag{223}$$

Eliminating $u_{i,j,1}, u_{j,i,1}, u_{i,j,2}, u_{j,i,2}$ (dividing Equation (221) from Equation (220), and Equation (223) from Equation (222)) we get

1. $k = 1$ and $l = 1$:

$$\frac{\exp(-\alpha_{j,k} - u_j) + 1}{\exp(-u_j) + 1} = \frac{\exp(-\alpha_{i,k} - u_i) + 1}{\exp(-u_i) + 1}\,. \tag{224}$$

2. $k = 1$ and $l = 2$:

$$\frac{\exp(\alpha_{j,k} + u_j) + 1}{\exp(u_j) + 1} = \frac{\exp(-\alpha_{i,k} - u_i) + 1}{\exp(-u_i) + 1}\,. \tag{225}$$

Since the RHS of both equations are the same, equating the LHS, we get

$$\frac{\exp(-\alpha_{j,k} - u_j) + 1}{\exp(-u_j) + 1} = \frac{\exp(\alpha_{j,k} + u_j) + 1}{\exp(u_j) + 1} \tag{226}$$

$$\exp(u_j + \alpha_{j,k}) = \exp(u_j) \tag{227}$$

$$\alpha_{j,k} = 0 \,. \tag{228}$$

Similarly we also get $\alpha_{i,k} = 0$.

Thus, if $i \not\perp j$ and $\mu_{i,j} \neq 1$, we have $\alpha_{i,k} = \alpha_{j,k} = 0$ for all $k \neq i, j$. This implies that for all $S \subseteq [d]/\{i, j\}$

$$p(x_i|x_S) = p(x_i) \,, \tag{229}$$
$$p(x_j|x_S) = p(x_j) \,, \tag{230}$$
$$p(x_i|x_S, x_j) = p(x_i|x_j) \,, \tag{231}$$
$$p(x_j|x_S, x_i) = p(x_j|x_i) \,. \tag{232}$$

or $i \perp k$ and $j \perp k$ for all $k \neq i, j$.

Case 2: $u_{j,i,1} = u_{i,j,1}$ **and** $u_{j,i,2} = u_{i,j,2}$

$u_{j,i,1} = u_{i,j,1}$ immediately gives us the following

$$\nu_{i,S} = \nu_{j,S} \,, \tag{233}$$
$$\alpha_{i,k} = \alpha_{j,k} \,. \tag{234}$$

From expression of ν we then have

$$\sum_{k \in S} u_{j,k,2} + u_j = \sum_{k \in S} u_{i,k,2} + u_i \,, \tag{235}$$

for all S, which immediately giving us

$$u_i = u_j \,, \tag{236}$$
$$u_{j,k,2} = u_{i,k,2} \,. \tag{237}$$

Combined with definition of $\alpha_{j,k}$ we also get

$$u_{j,k,1} = u_{i,k,1} \,. \tag{238}$$

Thus variables i, j are interchangeable. This can be summarized as

$$p(x_i|x_S) = p(x_j|x_S) \,, \tag{239}$$

for all $S \subseteq [d]/\{i, j\}$, and

$$p(x_i|x_S, x_j) = p(x_j|x_S, x_i) \,. \tag{240}$$

Step 2: Deriving the form of joint $p(x)$ from the identified conditions for any pair (i, j)

To summarize, for any pair i, j we have one of the following conditions that may hold.

$C1$. **Independence**: $i \perp j$

$$p(x_i|x_S, x_j) = p(x_i|x_S) \tag{241}$$
$$p(x_j|x_S, x_i) = p(x_j|x_S) \tag{242}$$

Because $p(x_i|x_S)p(x_j|x_S, x_i) = p(x_j|x_S)p(x_i|x_S, x_j)$, it follow that $p(x_i|x_S, x_j) = p(x_i|x_S)$ if and only if $p(x_j|x_S, x_i) = p(x_j|x_S)$. Thus, both conditions are equivalent.

$C2$. **Independence of other variables**: $\forall k \neq i, j$ we have $i \perp k$ and $j \perp k$.

$C3$. **Exchange equivalence**: i and j can be exchanged

$$p(x_i|x_S) = p(x_j|x_S) \tag{243}$$
$$p(x_i|x_S, x_j) = p(x_j|x_S, x_i) \tag{244}$$

Since $p(x_i|x_S)p(x_j|x_S, x_i) = p(x_j|x_S)p(x_i|x_S, x_j)$, we have $p(x_i|x_S) = p(x_j|x_S)$ if and only if $p(x_i|x_S, x_j) = p(x_j|x_S, x_i)$. Thus, both conditions are equivalent.

To show that $p(x)$ can be factorized as in the Theorem statement, we rely on a series of lemmas and definitions which we state here.

Lemma B.6. *Exchange equivalence ($C3$) forms an equivalence class.*

Proof. 1) Reflexivity is trivially satisfied when $i = j$. 2) Symmetry: Conditions are symmetric in i, j. 3) Transitivity: Let (i, j) and (j, k) be exchange equivalent. Then $p(x_i|x_S) = p(x_j|x_S) = p(x_k|x_S)$, thus $p(x_i|x_S) = p(x_k|x_S)$. □

Definition B.7 (Equivalence class)**.** *Let $\langle i\rangle$ refer to the equivalence class of i with respect to exchange equivalence relation. Any two equivalence classes (not necessarily distinct) $\langle i\rangle$, $\langle j\rangle$ be called independent denoted by $\langle i\rangle \perp \langle j\rangle$ if for all $k \in \langle i\rangle$ and $l \in \langle j\rangle$, such that $k \neq l$, we have $k \perp l$.*

Lemma B.8. *If $i \perp j$, and $\langle j\rangle = \langle k\rangle$ where $k \neq i$, then $i \perp k$.*

Proof. $i \perp j$ implies $p(x_j|x_S, x_i) = p(x_j|x_S)$. $\langle j\rangle = \langle k\rangle$ implies $p(x_j|x_S) = p(x_k|x_S)$ and $p(x_j|x_S, x_i) = p(x_k|x_S, x_i)$. Thus we get $p(x_k|x_S, x_i) = p(x_k|x_S)$, and thus $i \perp k$. □

The above lemma lets us lift the independence relation to the level of equivalence classes.

Lemma B.9. *For $i \neq j$, $i \perp j$ if and only if $\langle i\rangle \perp \langle j\rangle$.*

Proof. ($\Leftarrow$) is trivial since it is simply the definition of $\langle i\rangle \perp \langle j\rangle$. ($\Rightarrow$) From Lemma B.8, $i \perp j'$ for all $j' \neq i$ such that $\langle j'\rangle = \langle j\rangle$. Again from Lemma B.8, $j' \perp i$ implies $j' \perp i'$ for $i' \neq j'$ such that $\langle i'\rangle = \langle i\rangle$. Together this gives, $i' \perp j'$ for all $i' \in \langle i\rangle, j' \in \langle j\rangle, j' \neq i'$, which is the required condition for $\langle i\rangle \perp \langle j\rangle$. □

Lemma B.10. *If $|\langle i\rangle| \geqslant 2$, then for all $j \notin \langle i\rangle$, we have that $\langle i\rangle \perp \langle j\rangle$.*

Proof. Suppose there exists j such that $\langle i\rangle \not\perp \langle j\rangle$. Since $|\langle i\rangle| \geqslant 2$, there exists $k, l \in \langle i\rangle$ such that $k \not\perp j$ and $l \not\perp j$. Since $\langle k\rangle \neq \langle j\rangle$ and $k \not\perp j$, the second condition ($C2$) must hold, i.e. j is independent of all other variables, but that implies $j \perp l$, leading to a contradiction. □

Thus by Lemma B.10 for any $i : |\langle i\rangle| \geqslant 2$, $\langle i\rangle$ is independent of all other variables and the joint can be factorized as

$$p(x) = p(x_{\langle i\rangle})p(x_{[d]/\langle i\rangle})\,. \tag{245}$$

Using this factorization for all equivalence classes, $i : |\langle i\rangle| \geqslant 2$ we get

$$p(x) = p(x_{[d]/C}) \prod_{\langle i\rangle \subseteq C} p(x_{\langle i\rangle})\,, \tag{246}$$

where

$$C = \bigcup_{i:|\langle i\rangle| \geqslant 2} \langle i\rangle\,. \tag{247}$$

Let us denote $V \subseteq [d]/C$ is the set of variables which are independent of the rest, thus factorizing p as

$$p(x) = p(x_{[d]/\{C \cup V\}}) \prod_{v \in V} p(x_v) \prod_{\langle i\rangle \subseteq C} p(x_{\langle i\rangle})\,. \tag{248}$$

Now since any variable $i \in [d]/\{C \cup V\}$, it cannot be independent to all of the rest, i.e. there exists some $j \in [d]/\{C \cup V\}$ such that $i \not\perp j$. For this pair (i, j), from condition ($C2$) they have to be independent from the rest of variables. Thus $p(x_i, x_j)$ can be factored out. Since this holds for any variable in $[d]/\{C \cup V\}$, p can be further factorized as

$$p(x) = \prod_{v \in V} p(x_v) \prod_{\{u,v\} \in E} p(x_u, x_v) \prod_{\langle i\rangle \subseteq C} p(x_{\langle i\rangle})\,, \tag{249}$$

where E is a partition of $[d]/\{C \cup V\}$. To get a unique factorization we can extract out $\langle i \rangle \subseteq C$ where $|\langle i \rangle| = 2$ and place them as part of E. Let $\mathcal{C} = \{\langle i \rangle : |\langle i \rangle| \geqslant 3\}$, where $\mathcal{C}$ is the set of equivalence classes containing three or more variables. The joint can be written as

$$p(x) = \prod_{v \in V} p(x_v) \prod_{\{u,v\} \in E} p(x_u, x_v) \prod_{C \in \mathcal{C}} p(x_C)\,, \tag{250}$$

where $V \cup E \cup \mathcal{C}$ is a partition of $[d]$ into independent subsets of variables of sizes 1, 2 and greater than or equal to 3, respectively.

Further for any $C \in \mathcal{C}$, $p(X_C)$ has the form

$$p(x_C) = f_C(\text{count}(x_C, 1), \text{count}(x_C, 2))\,, \tag{251}$$

for some function f_C, where $\text{count}(x_C, k) = \sum_{c \in C} \mathbb{I}(x_c = k)$. This comes from the fact that variables in C are exchangeable, i.e. given two permutations of variables in C, $x_{i_1}, \ldots x_{i_k}$ and $x_{j_1}, \ldots x_{j_k}$, the probabilities are equal, i.e.

$$p(x_{i_1}, \ldots x_{i_k}) = p(x_{j_1}, \ldots x_{j_k})\,. \tag{252}$$

From Zaheer et al. (2017, Theorem 2), any permutation invariant function can be written as a sum, i.e. there exists f_C, ϕ_C such that

$$p(x_C) = f_C\Big(\sum_{c \in C} \phi_C(x_c)\Big)\,. \tag{253}$$

Since any x_c takes two values $x_c = 1$ or $x_c = 2$, let $\phi_C(1) = \lambda_1$ and $\phi_C(2) = \lambda_2$, we get

$$p(x_C) = f_C\Big(\sum_{c \in C} \lambda_1 \mathbb{I}(x_c = 1) + \lambda_2 \mathbb{I}(x_c = 2)\Big)\,. \tag{254}$$

This is a special case of a more general form

$$p(x_C) = f_C\Big(\sum_{c \in C} \mathbb{I}(x_c = 1), \sum_{c \in C} \mathbb{I}(x_c = 2)\Big) \tag{255}$$

$$= f_C(\text{count}(x_C, 1), \text{count}(x_C, 2))\,. \tag{256}$$

□

C DETAILS OF EXPERIMENTS

C.1 Model and Dataset Statistics

d	# edges	# params	max params
10	32	221	1023
25	84	610	3.3×10^7
50	116	647	1.1×10^{15}

Table 6: # edges denote the total number of edges in the DAG, # params denote the total number of parameters in the CPT, i.e. $\Pi_{i=1}^{d} 2^{|pa(i)|}$. Max params denote the number of parameters ($2^d - 1$) required to specify an arbitrary joint distribution over d variables.

C.2 Experiments for Transformer

C.2.1 Transformer Model

We use an encoder-only transformer model that takes a sequence of inputs $U = (u_i, \ldots u_n)$ where $u_i \in \mathbb{R}^D$ and transforms them to produce the sequence of outputs $V = (v_i, \ldots v_n)$ where $v_i \in \mathbb{R}^D$. This is achieved by a sequence of intermediate transformations resulting in sequences $U_1, \ldots U_k$ by positionwise feedforward network and self-attention.

n	d_{context}	# layers	d_{hidden}	# params
10	2	0	-	620
25	2	0	-	3800
50	2	0	-	15100
10	512	8	128	3.4×10^5
25	512	8	128	1.1×10^6
50	512	8	128	4.0×10^6

Table 7: Statistics for the smallest (g is a sigmoid) and largest (g is an MLP with d_{hidden} neurons in each of the 8 hidden layers) model used in our experiments.

Positionwise Feedforward: In positionwise feedforward network, the current sequence $U_j = (u_1^j, \dots u_n^j)$ is transformed to $U_{j+1} = (u_1^{j+1}, \dots u_n^{j+1})$ as

$$u_i^{j+1} = f_{\theta^j}(u_i^j), \quad \text{for } i = 1 \dots n,$$

where f is a feedforward network (a 2-layer ReLU network followed by layer normalization and skip connection) with parameters θ_j.

Multi head self-attention : In the multi-head self-attention mechanism, the transformed output is given as

$$u_i^{j+1} = \text{Concat}([u'_1 \dots u'_h]), \tag{257}$$

$$u'_s = \sum_{k=1}^{n} \alpha_{ik}^s W_j^s u_k^j, \quad \text{for } s = 1 \dots h, \tag{258}$$

where W_j^s for $s = 1 \dots h$ are learned projection matrices for j'th layer. h is the number of heads. α_{ik}^s is given as,

$$\alpha_{ik}^s = \tilde{\alpha}_{ik}^s / \sum_{k=1}^{n} \tilde{\alpha}_{ik}^s, \tag{259}$$

$$\tilde{\alpha}_{ik}^s = \exp((Q_j^s u_k^j)^\top (K_j^s u_i^j)), \tag{260}$$

where Q_j^s, K_j^s are learned projection matrices for j'th layer and heads $s = 1 \dots h$.

C.2.2 Adapting transformer for our setting

Given an input (i, S, X_S), we create the input sequence of length $|S| + 1$ to be fed to transformer as

$$(\phi((j_1, X_{j_1})), \dots \phi((j_k, X_{j_k})), \phi(i))$$

where $S = \{j_1, \dots j_k\}$ and ϕ is a learned embedding lookup table. As is common in the usage of transformer for language modeling, we add the embeddings for position $j_1, \dots j_k$ and words $X_{j_1} \dots X_{j_k}$ as

$$\phi((i, v)) = p_i + z_v \text{ for } v = 1 \dots K, \tag{261}$$

and

$$\phi(i) = p_i + z_0, \tag{262}$$

where $p_1, \dots p_d$ are learned positional embeddings for the variables $1 \dots d$ and $v_1, \dots v_K$ are embeddings for the values any x_i takes that ranges from $1 \dots K$. z_0 is an additional embedding referring to the absence of the value or to mark the variable i for which the value is to be predicted. In language modeling, this is referred to as a masked token.

Finally, the last element of the output sequence is used as a prediction for X_i, after projecting back to $\mathbb{R}^K$ space, i.e. let $v_i \in \mathbb{R}^D$ refer to the last element of the output sequence, then X_i is predicted as

$$P(X_i = k) = \frac{\exp(w_k^\top v_i)}{\sum_{k=1}^{K} \exp(w_k^\top v_i)}, \tag{263}$$

where $w_1, \dots w_K$ are also learned embeddings.

D Examples of models that satisfy Theorem 3.4, Theorem 3.5

Theorem 3.4 implies that $f_{i|-i}(x_i, x_{-i})$ must be able to factorize in the form $h(x)q_i(x_{-i})$ where $h(x)$ is a common function for all $f_{i|-i}(x_i, x_{-i})$. One example of an architecture that satisfies this condition is to explicitly parameterize h and each q_i with separate neural networks. However, this approach requires d neural networks to model $p(x_i|x_{-i})$ and exponentially many neural networks, to model $p(x_T|x_S)$ even when $S \cup T = [d]$. For computational efficiency, we require some shared components between different conditionals. In our paper, we provide empirical evidence that MLPs are not consistent and pose the problem of designing an efficient and consistent architecture as an open question.

Theorem 3.5 provides a necessary condition for path consistency for parameterizations of the form of softmax layer on a linear layer W on top of a feature map ϕ. We do not make any assumptions on ϕ so this result is quite general and includes deep neural networks. An example of a model that satisfies this condition is an exponential family in Equation (11). In general, if we have all the feature maps ϕ_i, then having one linear head W_1 is sufficient to recover all W_j up to an additive factor by solving the Equation (10).

E Quantitative Metrics

In this section, we explain why our proposed metrics are meaningful measures of consistency metric.

1. **Path consistency** (Equation (25)):

$$\mathcal{E}_{\text{PC}}(\theta) = \mathbb{E}_{x,\bar{x}\sim P}\left[\text{Std}_\sigma[\log h_{\sigma,\bar{x}}(x; f_\theta)]/d\right]. \tag{264}$$

The term inside the expectation is the standard deviation of $\log(h_\sigma(x; f_\theta))$ over a random path σ. As described in Section 3, h_σ in Equation (5) is the constructed ratio of the joint distribution of x and x' given σ. This ratio is independent of σ when the model is consistent (Proposition 3.3). Thus, our metric which calculates the standard deviation of this ratio over different paths σ becomes zero. Moreover, when these values are not exactly zero, which implies inconsistency, it still tells us how much the recovered ratio of the joint distribution can vary with the chosen path. Hence, these values being small imply that the conditionals are more robust to the different paths, and thus are more consistent.

2. **Autoregressive path consistency** (Equation (26)):

$$\mathcal{E}_{\text{AC}}(\theta) = \mathbb{E}_{x\sim P}\left[\text{Std}_\sigma[\log g_\sigma(x; f_\theta)]/d\right]. \tag{265}$$

The term inside the expectation is the standard deviation of $\log(g_\sigma(x; f_\theta))$ over a random path σ. As described in Section 4, $g_\sigma(x)$ is the joint distribution at x recovered by an autoregressive path recovery with a permutation σ. When the conditionals are consistent, $g_\sigma(x)$ would be independent of σ which implies that our metric which calculates the standard deviation of $\log(g_\sigma(x))$ would also be zero. In the same manner, our metric would have a smaller value whenever the conditionals are more consistent.

3. **Swap consistency metric** (Equation (27)):

$$\mathcal{E}_{\text{SC}}(\theta) = \mathbb{E}_{(x,i,j,S)\sim\mathcal{D}'}[\Delta(f_\theta, x, i, j, S)], \tag{266}$$

$$\begin{aligned}\Delta(f, x, i, j, S) = |&\log(f(x_i|x_{S\cup\{j\}})f(x_j|x_S)) \\ &- \log(f(x_j|x_{S\cup\{i\}})f(x_i|x_S))|.\end{aligned} \tag{267}$$

The term inside the expectation is the standard deviation of the recovered $P(x_i, x_j \mid x_S)$ over two different paths $S \to i \to j$ and $S \to j \to i$. This is based on the observation that in Equation (266), we are calculating $\Delta(f, x, i, j, S) = \left|\log\left(f(x_i \mid x_{S\cup\{j\}})f(x_j \mid x_S)\right) - \log\left(f(x_j \mid x_{S\cup\{i\}})f(x_i \mid x_S)\right)\right|$. Here, each term is a different way to recover the conditional $P(x_i, x_j \mid x_S)$. The standard deviation interpretation follows from the fact that the standard deviation of two values a, b is given by $|a - b|/2$. Similarly, this becomes a meaningful quantitative metric for swap consistency.