My Final College Paper

A Thesis
Presented to
The Division of Mathematics and Natural Sciences
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Your R. Name

May 200x

Approved for the Division
(Mathematics)

Advisor F. Name

Acknowledgements

I want to thank a few people.

Preface

This is an example of a thesis setup to use the reed thesis document class.

List of Abbreviations

Contents

List of Tables

List of Figures

Abstract

The preface pretty much says it all.

Dedication

You can have a dedication here if you wish.

Introduction

Chapter 1

The First

1.1 Background: Graphs, D-separation, Causality

The idea of cause and effect has been studied and discussed in philosophy for centuries, but the formalization of causality in mathematics, statistics, and computer science is much more recent. One framework in particular, Judea Pearl's Structural Causal Models (SCMs), [ref Pearl 1996] is flexible, widely used, and mathematically elegant. However, before we can give SCMs a serious treatment, it is helpful to introduce definitions for the graphical machinery that most analysis of SCMs rely on.

1.1.1 Directed Acylic Graphs

Definition 1.1.1. *A* **graph** *G is a pair $G = (V, E)$ where V denotes a set of nodes (sometimes called vertices) and $E \subseteq \{(i, j)|i, j \in V\}$ denotes a set of edges between nodes.*

Conventions differ, but for our purposes, an edge $(i, j) \in E$ is considered to be a **edge**, read as "an edge from node i to node j". For easy reading, we write $i \to j$ whenever $(i, j) \in E$. In the figures that frequently accompany graphs, an edge (i, j) is depicted as an arrow from node i to node j.

Often times we are interested in how different nodes in a particular graph are or are not connected to one another. In this spirit we define paths.

Definition 1.1.2. *A* **path** *p between nodes $v_1, v_n \in V$ is a sequence $v_1, v_2, ..., v_n \in V$ of distinct nodes such that either $(v_i, v_{i+1}) \in E$ or $(v_i, v_{i+1}) \in E$ edge always exists between v_i and v_{i+1}. When $v_i \to v_{i+1}$ $((i, i-1) \in E)$ for all $i \in \{1, 2, ..., n\}$, we say p is a* **directed path**.

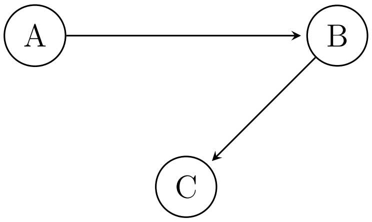


Figure 1.1: Graphs corresponding to $V = \{A, B, C\}$, $E = \{(A, B), (B, C)\}$

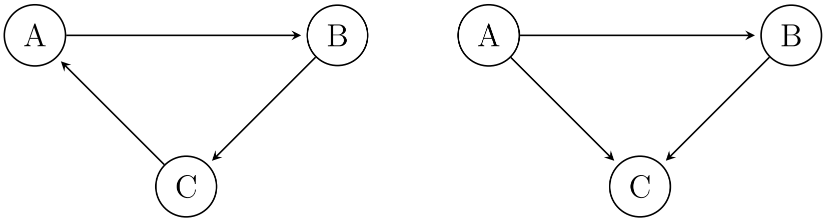


Figure 1.2: Cyclic (left) and acyclic (right) graphs on the nodes A, B, C

For a directed graph G and a particular node $v \in V$, we define several sets of related nodes. The **parents** of v, denoted PA_v, are the nodes with a directed edge ending at v, $\{j \in V | j \to v\}$ and similarly the **children**, denoted CH_v, of v are those nodes which v has a directed edge to, $\{j \in V | v \to j\}$.

Extending these definitions, we define the **ancestors** of v, AN_v, as all the nodes from which a directed path to v exists, $\{j \in V | j \to ... \to v\}$. Likewise we define **descendants** DE_v of v as the nodes for which a directed path from v exists, $\{j \in V | v \to ... \to j\}$ [ref Peters chp 6.1]. With these in place, we progress to the next definition.

Definition 1.1.3. *A directed graph $G = (V, E)$ is said to be an* **directed acyclic graph** *(DAG) if for all nodes $v \in V$, $DE_v \cap AN_v = \emptyset$. That is, a directed graph is a DAG when there is never a directed path from a node to itself.*

Now that we have defined DAGs, we can advance to a relevant application, Bayesian networks.

1.1.2 Bayesian Networks

Before we do so, we review some basic definitions from probability.

Definition 1.1.4. *For random variables $X_1, X_2, ..., X_n$, the* ***joint distribution*** *with associated joint probability function $P(x_1, x_2, ..., x_n)$ gives the probability of every possible combination of values for $X_1, X_2, ..., X_n$. For any subset $\{Y_1, ..., Y_k\} \subseteq \{X_1, X_2, ..., X_n\}$, the probability distribution for $Y_1, Y_2, ..., Y_k$, is called the* ***marginal distribution****. Finally, for a subset*

1.1.3 D-Separation

Graphical structure can allow us to reason about causal and statistical relationships. One of the essential tools is d-separation. To develop the intuition, consider the very basic DAG $A \to B \to C$. There is exactly one directed path from A to C, and the path passes through B. So to get to C from A you need to pass through B. In this way, we say that B blocks the directed path from A to C. We extend this notion [ref Pearl 2008], [ref Peters 2017].

Definition 1.1.5. *Given a directed graph G, a path p from node v_1 to node v_n is said to be* **blocked** *by a set S (with $S \cap \{v_1, v_n\} = \emptyset$) if:*

1. *$v_j \in S$ and p contains a* **chain***: $v_{j-1} \to v_j \to v_{j+1}$ or $v_{j-1} \leftarrow v_j \leftarrow v_{j+1}$, or*

2. *$v_j \in S$ and p contains a* **fork***: $v_{j-1} \leftarrow v_j \rightarrow v_{j+1}$, or*

3. *$v_j \notin S$ and $DE_{v_j} \cap S = \emptyset$ and p contains a* **collider***: $v_{j-1} \rightarrow v_j \leftarrow v_{j+1}$. As we will see, colliders play an especially important role in the analysis of causal DAGs.*

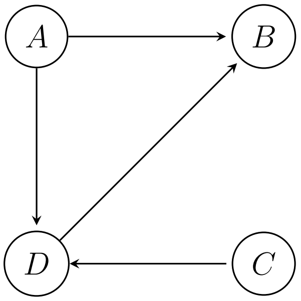
Example 1. *Consider a path $A \rightarrow B \rightarrow C \leftarrow D \rightarrow E$ between nodes A and E. There are many sets that block this path. For one, there is a collider on the path. This means that any set which does not include C or any elements of DE_C would successfully block the path. However, we could also pick $S = \{B\}$, $S = \{D\}$, or $S = \{B, D\}$ as blocking sets since the path contains a chain at B and a fork at D. Importantly, if we have either of these nodes in our blocking set, we could include C as well because the first blocking condition would be satisfied, even though the second would not. In fact, the only $S \subseteq \{B, C, D\}$ that would not block the path is $S = \{C\}$.*

With blocking defined, we address d-separation.

Definition 1.1.6. *For a DAG G, two sets of nodes $A, B \in G$ are said to be* **d-separated** *by a set S if all paths between A and B are blocked by S. We denote d-separation by the symbol $\perp\!\!\!\perp_G$, with the statement "In DAG G, A is d-separated from B by S" expressed as $A \perp\!\!\!\perp_G B|S$.*

Remark 1.1.1. *If $A \perp\!\!\!\perp_G B|S$ then $B \perp\!\!\!\perp_G A|S$*

Example 2. *Consider the DAG G below.*



Which nodes can be d-separated, and by which blocking sets? Nodes which are connected by an edge cannot be d-separated. To see why, take A and B. Although some paths between A and B can be blocked, the path $A \rightarrow B$ contains no nodes other than A and B and therefore cannot be blocked. Since there is no set that can block the path, A and B are not able to be d-separated. So consider A and C, two nodes which are not connected by an edge. There are two paths between A and C, $A \rightarrow B \leftarrow D \leftarrow C$ and $A \rightarrow D \leftarrow C$. Both of these paths include a collider. Of the four possible blocking sets, $\emptyset, \{B\}, \{D\}, \{B, D\}$, all but $\emptyset$ include a collider node for paths that are not blocked by other means. Then $A \perp\!\!\!\perp_G C|\emptyset$ and no other sets will work. What about B and C? Again there are two paths: $B \leftarrow D \leftarrow C$ and $B \leftarrow A \rightarrow D \leftarrow C$. The second path contains a collider, and therefore is blocked by $\emptyset$, but the first path cannot be. To block $B \leftarrow D \leftarrow C$ we must have D in our blocking set. So, $B \perp\!\!\!\perp_G C|D$, but we can also include A since doing so will not unblock a path with a collider. Then $B \perp\!\!\!\perp_G C|D, A$.

1.1.4 Structural Causal Models

Many questions in the natural and social sciences involve understanding the way that one set of factors (e.g. genes, medicines, social policies) influence others (e.g. disease risk, mortality, childhood poverty rates). In these cases, the language and of classical statistics is often inadequate. The tools of hypothesis testing, (linear) regression, interval estimation, etc. can tell us much about relationships in the data, but except in the highly constrained setting of randomized controlled trials, very little about cause and effect. These limitations are not just theoretical. Many longstanding controversies and debates within the sciences come down to disagreements over which factors associated with a particular outcome are causing the outcome.

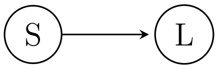
A paradigmatic example of such a controversy is over the relationship between cigarette smoking and lung cancer. Today, it is uncontroversial that smoking cigarettes (especially habitually over a long period of time) increases a person's risk of contracting lung cancer. This is a causal relationship, and we might express our belief in this relationship as a ***causal model*** in which a person's risk of contracting lung cancer is a *function* of a person's smoking habits (for simplicity, we pretend that smoking is the only such causal factor).

So, let L be a random variable denoting whether or not a person contracts lung cancer and S be a random variable denoting smoking. Under our causal model in which smoking is the only structural cause of lung cancer, we express L as a function of S:

$$L := f(S, N_L),$$

$$S := N_S.$$

Here, the use of := denotes an assignment rather that an algebraic relationship [ref Peters] and N_L is a noise term capturing random effects on lung cancer risk, and N_S a noise variable for cigarette smoking. Together, these two assignments are a structural causal model. Now that we have an SCM, we can express the structure as a DAG. Since L is assigned as a function of smoking but S is not caused by any factors in our model, we would express our model as a directed edge going from S to L:



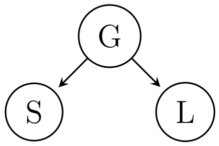
However, historically, this causal structure was not the only one suggested to explain the association between lung cancer and smoking. One particularly illustrious detractor of this theory was British statistician R. A. Fisher, a foundational figure in modern statistics (and a heavy smoker) [ref Fisher]. In Fisher's view, there was no causal effect of smoking on lung cancer, rather, an underlying genetic factor was responsible for both a predisposition to smoking and a higher risk of developing lung cancer. So Fisher's causal model would have looked like:

$$L := f(G, N_L),$$

$$S := f(G, N_S),$$

$$G := N_G.$$

Where G is the genetic factor. As before, the structure of this model can be expressed as a graph:



In this way, every SCM entails a directed graph. Although it is entirely possible for an SCM to entail a graph that has cycles, we focus our attention on the case where the graph entailed is a DAG. Importantly, these two (graphical) models imply different observational joint distributions. Without having specified the functions, little can be said about the joint distribution entailed by the model, however, one important property is built in to every SCM.

Definition 1.1.7 (Markov property)**.** *In any SCM C with entailed graph G, if $A \perp\!\!\!\perp_G B|S \implies A \perp\!\!\!\perp B|S$ for all disjoint sets of nodes A, B, S. That is, every d-separation in the graph corresponds to a conditional independence in the joint distribution. This fact is implied by the definition of SCMs and is called the* **Markov Property***.*

[This seems important enough to reproduce a proof but I will have to spend a little more time than I have tonight]

This result is a powerful tool that allows us to test how well differing causal models comport with the observational data. Returning to Fisher's suggestion that an underlying genetic cause was responsible for the association of smoking with lung cancer, consider that by the Markov property $S \perp\!\!\!\perp L|G$. Let's assume we have a population of identical twins, some of whom were smoking discordant. Since we know that identical twins are genetically identical we could assess whether or not smoking and lung cancer are independent within twin pairs. If not, then something must be wrong with the model.

Chapter 2

Bareinboim et al 2014

2.1 Introduction

Bareinboim, Tian, and Pearl's 2014 paper "Recovering from Selection Bias in Causal and Statistical Inference" provides a formulation of selection bias explicitly in terms of causal DAGs and d-separation, making it an especially important resource for developing our study of the topic. In particular, the paper provides graphical conditions (in terms of d-separations) under which a *conditional* distribution $P(y|x)$ can be obtained despite selection bias. When this can be done, we say that the $P(y|x)$ can be "recovered", from a causal graph containing a selection mechanism. However, the content of the paper is highly abstracted and does not attempt to provide real-life cases in which their techniques are useful. For this reason, we will summarize the main results of the paper and interpret them in the context of some realistic examples.

The paper is broken into three sections, the first two of which are the most substantial and the focus of our attention. The first gives conditions for when a distribution can be recovered if only the biased distribution is accessible, the second gives conditions for the case where population level data is available for some variables, and the last gives a brief discussion of recovering causal effects.

2.2 A Famous Case of Selection Bias

One classic example of selection bias is the 1936 presidential election poll published in an American magazine called The Literary Digest. The election was between Franklin Delano Roosevelt, the democratic incumbent, and Alfred Landon, a republican. The Literary Digest conducted a poll, drawing from their readership and by telephone and using records of automobile ownership. The sample size was extremely large, with over 2 million survey respondents. Their prediction, that Landon would comfortably win the election, proved embarrassing. Not only did Landon lose, but he lost in one of the largest electoral landslides in American history: 523 electoral votes went to Roosevelt and 8 went to the Digest's predicted winner. Shortly after, the magazine ceased publication.

The precise details of what caused this failure are not fully known. One view that has become the popular explanation was that although their survey had a plenty large

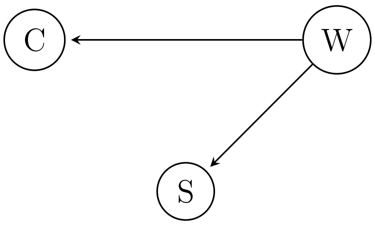


Figure 2.1: Graph corresponding to selection bias based on income

sample size, it was non-representative of the population as a whole. Particularly, both being subscribed to The Literary Digest and owning a telephone or car were more common among more affluent people, and so their survey skewed wealthier than the population as a whole, and therefore underestimated the proportion of voters who would vote for Roosevelt, a candidate with largely working class support. In fact, this explanation is unlikely to be complete. Although it is true that the survey over-sampled more affluent voters who were more likely to support Landon, a separate poll conducted by George Gallup in 1937 found that in even voters who owned both a telephone and a car were more likely to favor Roosevelt than Landon. This means that there must have been other factors at play, and in particular, that Landon voters were more likely to return the survey than Roosevelt voters (**?**), **?**. This is called (non) response bias, and we will see that it falls under the selection bias framework proposed here.

For now, though, let's assume that income was the only variable that was responsible for the non-representative sample, and that this information was collected from respondents as part of the survey. However, we will return to this case later to add some depth.

One of the important features of selection bias is that unlike sampling variation, it cannot be reduced with sufficiently large sample sizes. For this reason, in the rest of the paper we will discuss a selection biased joint distribution for this survey, and thereby ignore sample size.

2.3 Recovering $P(y|x)$ Without External Data

In the simplest case, we only have access to a distribution corresponding to a causal graph G with a set of nodes $\mathbf{V}$ equipped with a selection node S. This means that we do not have access to the full joint distribution $P(\mathbf{v})$ but rather the conditional joint distribution $P(\mathbf{v}|S = 1)$. This corresponds to the joint distribution of values that are selected to be sampled. So, returning to our election example, our graph would look like: Where C is a random variable for the candidate a person intends to vote for, W is measures income and S is a binary variable denoting inclusion in the study. Although in practice we would probably like to know $P(c)$, the population level distribution of intended votes, for now we restrict ourselves to $P(c|w)$, the conditional distribution of intended vote given SES.

In the vein, Bareinboim et al. define **s-recoverability** as the criterion which must be satisfied for a conditional distribution such as $P(c|w)$ to be obtainable under selection bias. We reproduce their definition exactly and give an explanation of why it makes sense.

Definition 2.3.1. *Given a causal graph G_s augmented with a node S encoding the selection mechanism, the distribution $Q = P(y|\mathbf{x})$ is said to be* ***s-recoverable*** *from selection biased data in G_s if the assumptions embedded in the causal model render Q expressible in terms of the distribution under selection bias $P(\mathbf{v}|S = 1)$. Formally, for every two probability distributions P_1 and P_2 compatible with G_s, $P_1(\mathbf{v}|S = 1) = P_2(\mathbf{v}|S = 1) > 0$ implies $P_1(y|\mathbf{x}) = P_2(y|\mathbf{x})$.*

If we have a selection biased distribution, we have access to a particular joint distribution, $P(\mathbf{v}|S = 1)$ where $\mathbf{V}$ is the set of all nodes other than S. So if we are to recover $P(y|\mathbf{x})$ ($\mathbf{X} \subset \mathbf{V}$, $Y \in \mathbf{V}$) we need to express $P(y|\mathbf{x})$ as values derivable from $P(\mathbf{v}|S = 1)$ and the conditional independences implied by the d-separations within the graph. This is what is meant by "if the assumptions embedded in the causal model render Q expressible in terms of the distribution under selection bias $P(\mathbf{v}|S = 1)$", and it is worth exploring why this fairly intuitive formulation is equivalent to the formal definition given in the last sentence.

Consider that the agreement of P_1 and P_2 on the joint distribution $P(\mathbf{v}|S = 1)$ means that P_1 and P_2 produce the same observational distribution under selection bias. Therefore, when $P_1(\mathbf{v}|S = 1) = P_2(\mathbf{v}|S = 1) \implies P_1(y|\mathbf{x}) = P_2(y|\mathbf{x})$ we know that no matter the differences between P_1 and P_2, they must produce the same conditional $P(y|\mathbf{x})$. So, if this is true for every P_1, P_2, then all distributions consistent with the observational $P(\mathbf{v}|S = 1)$ produce the same $P(y|\mathbf{x})$. The first important result of the paper is the following theorem:

Theorem 1. *The distribution $P(y|\mathbf{x})$ is s-recoverable from G_s if and only if $S \perp\!\!\!\perp_{G_s} Y|\mathbf{X}$.*

The proof of $\impliedby$ is quite long, but $\implies$ is easy.

Proof. Assume that $S \perp\!\!\!\perp_{G_s} Y|\mathbf{X}$. Then by the Markov property, $S \perp\!\!\!\perp Y|X$ and $P(y|\mathbf{x}, S = 1) = P(y|\mathbf{x})$. So the 'recovered' distribution is simply the biased distribution. □

Clearly this is a useful result since it is easy to test whether two nodes are d-separated by a set. In our application, we can see that $V \perp\!\!\!\perp_G S|W$, and so $P(v|w) = P(v|w, S = 1)$. So our study gives us direct access to the conditional distribution for candidate preference by income. This is potentially useful but restricted. What if we wish to recover a conditional distribution but do not want to condition on the full separating set? To do so, we will need external measurements.

2.4 Recovery With External Data

Sometimes a researcher will conduct a survey and collect variables on survey participants for which we know the population level distribution. For instance, the US Census collects and publishes data on occupation, race, income, and many other variables from the entire US population. If some of these variables are measured in our selection biased study, this information can be used to recover distributions that would otherwise be unavailable. To elaborate on this topic we will introduce a more complex model of the selection mechanism in our voter poll. As mentioned in the introduction, more recent research suggests that the assumption that income was the only variable affecting the probability of being included in the poll is not sufficient. A better explanation includes the fact that Roosevelt supporters

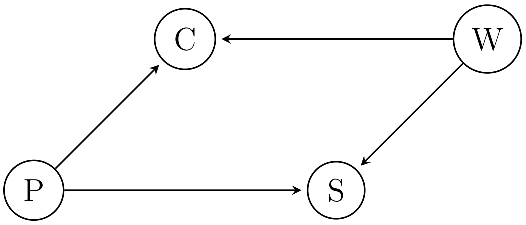


Figure 2.2: Graph corresponding to selection bias based on income and party registration

who received the survey were much less likely to respond than Landon supporters **?**. This effect is called non-response bias, and is really another form of selection bias since it arises from the differing probabilities of being included in the final sample. The exact mechanism of the non-response effect in this poll is not and likely will never be known. In fact, if candidate preference directly influences the likelihood of survey response, (i.e. $C \rightarrow S$), the two nodes cannot be d-separated and we will be out of luck. Being optimistic, however, let's say we have reason to believe that another factor is the culprit. Particularly, the political party to which a survey respondent is registered could be mediating the relationship between voter preference and response probability. This model is represented in the figure below.

Where P is a binary variable denoting the party registration status of the voter. Under this model, we have a path $S \leftarrow U \rightarrow V$ which is not blocked by W. So $S \not\perp\!\!\!\perp_G Y|X$ and by the previous theorem, we cannot recover $P(c|w)$. Of course, we could recover $P(c|w, p)$ since $S \perp\!\!\!\perp_G V|\{W, P\}$, but assuming we are still interested in $P(c|w)$, what can be done?

Well, we could consult the US census as well as state voter records and gather the joint distribution for income and party registration $P(u, w)$. This is what we would call *external data* since it is not affected by the selection mechanism. It turns out that in this case, we can do something. To address this situation, Bareinboim et al. introduce a revised definition of s-recoverability which we again reproduce and comment on.

Definition 2.4.1. *Given a causal graph G_s augmented with a node S encoding the selection mechanism, the distribution $Q = P(y|\mathbf{x})$ is said to be **s-recoverable** from selection bias in G_s with external data over $\mathbf{T} \subseteq \mathbf{V}$ and selection biased data over $\mathbf{M} \subseteq \mathbf{V}$ if the the assumptions embedded in the causal model render Q expressible in terms of the distribution under selection bias $P(\mathbf{m}|S = 1)$ and $P(\mathbf{t})$, both positive. Formally, for every two probability distributions P_1 and P_2 compatible with G_s, if they agree on the available distributions $P_1(\mathbf{v}|S = 1) = P_2(\mathbf{v}|S = 1) > 0$, $P_1(\mathbf{t}) = P_2(\mathbf{t})$ they must agree on the query distribution $P_1(y|\mathbf{x}) = P_2(y|\mathbf{x})$.*

We can see that this definition follows the same structure as the original s-recoverability definition, but is expanded to allow for the use of population level distributions. The paper's second theorem follows directly from this definition.

Theorem 2. *If there is a set $C \subseteq V$ such that $P(\mathbf{c}, \mathbf{x})$ is measured in the population and $Y \perp\!\!\!\perp_{G_s} S|\{C, X\}$ then $P(y|\mathbf{x})$ is s-recoverable as*

$$P(y|\mathbf{x}) = \sum_{\mathbf{c}} P(y|\mathbf{x}, \mathbf{c}, S = 1)P(\mathbf{c}|\mathbf{x})$$

The theorem is a straightforward application of the law of total probability.

Proof. By assumption, we have the external distribution $P(\mathbf{c}, \mathbf{x})$ and therefore $P(\mathbf{c}|\mathbf{x})$, and as usual we have $P(\mathbf{v}|S = 1)$. So can apply the law of total probability and the conditional independence $Y \perp\!\!\!\perp S|\{C, X\}$ to write:

$$P(y|\mathbf{x}) = \sum_{\mathbf{c}} P(y|\mathbf{x}, \mathbf{c})P(\mathbf{c}|\mathbf{x}) = \sum_{\mathbf{c}} P(y|\mathbf{x}, \mathbf{c}, S = 1)P(\mathbf{c}|\mathbf{x})$$

□

Therefore, if we wish to recover $P(c|w)$ party registration affects response probability, we can do so using the external data $P(p|w)$.

2.5 A Useful Extension

As we have seen, Bareinboim et al. is focused on the recovery of conditional distributions, i.e. $P(v|w)$. However, we would often like to have the unconditional distribution $P(v)$. Although it is only mentioned obliquely at the end of the second section, the results they prove give a simple condition for when $P(v)$ is recoverable using external data. In both sections we have seen that $P(v|w)$ was recoverable. Then, assuming that we have the external data for $P(w)$, we can use the law of total probability to write:

$$P(v) = \sum_{w} P(v|w)P(w).$$

More generally, we can formulate this result as a theorem, although it is not listed as such in the paper.

Theorem 3. *If there exists a set* $\mathbf{X} \subseteq V$ *such that* $P(y|\mathbf{x})$ *is s-recoverable and* $P(\mathbf{x})$ *is available externally, then* $P(y)$ *is recoverable as* $P(y) = \sum_{\mathbf{x}} P(y|\mathbf{x})P(\mathbf{x})$.