A. A Primer on Causal (and Proximal) Inference

Every schoolchild knows that "correlation does not imply causation," but this has not stopped people from trying to develop methods for inferring causality purely based on observational data. While trying to understand the world around us has always been a fundamental part of human civilization, ultimately, much of this knowledge has relied on our ability to perform experiments, and other interventions, in order to test the hypotheses that we have generated based on observation. The bulk of scientific progress, and other advancements in understanding our environment, relies upon the pairing of deduction with induction, rather than utilizing inductive reasoning alone.

The goal of causal inference from observational (non-experimental) data is much more ambitious: to understand cause and effect without any intervention. Clearly, such lofty aspirations must come with significant limitations and, indeed, causal inference requires strong, untestable assumptions to be satisfied in order for our causal estimates to be valid. However, miraculously, when these assumptions hold, we are able to infer causation based on observation alone.

Statisticians have long flirted with causality, with Jerzy Neyman even trying to formalize it in the form of his Potential Outcomes framework (Neyman, 1923). However, Ronald Fisher famously believed that causal inference based solely on observation had no place in statistics and that randomized experiments were necessary for drawing causal conclusions. Thus, causal inference in statistics remained out of favor until the mid 1970s, when Donald Rubin took up Neyman's framework, further developing it and applying it to the analysis of observational studies (Rubin, 1974). In the mid 1980s, James Robins generalized the Neyman-Rubin counterfactual model to studies with treatment and confounding covariates that varied over time (Robins, 1986). In the early 1990s Judea Pearl began developing graphical and structural equation based models that connected counterfactuals with Directed Acyclic Graphs (DAGs) representing the underlying causal structures (Pearl, 1993). Since then, the field has advanced significantly, with causal methods becoming widely used in statistics and machine learning in recent years.

We will present causal inference using a combination of a counterfactual framework based on Neyman's Potential Outcomes and Directed Acyclic Graphs. We consider a potential cause, X, and an outcome, Y. X and Y might be, for example, starting a medicine and the 10-year risk of death, the price of an item and annual sales, or going to college and lifetime earnings. In all these cases, X and Y also share a common cause, which we will denote by U; such common causes are referred to as confounders. We can also express these relationships graphically using DAGs, as in Figure 3.

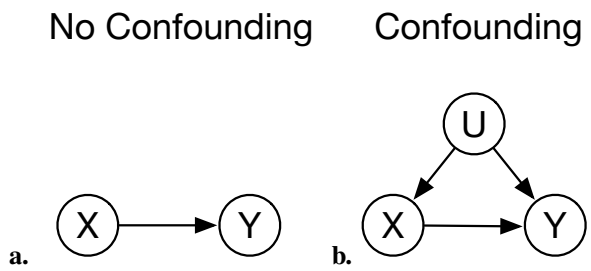

Figure 3. **a.** Unconfounded and **b.** Confounded DAGs. X is a potential cause, Y is the outcome of interest, and U is a (potentially unmeasured) confounder.

In the first example, patients in worse health are more likely to be prescribed a preventive medicine and are also more likely to die sooner, in the second example, factors like supply, geography, and consumer demand have direct effects on both price and sales, and, in the third example, whether one goes to college or not and one's lifetime earnings are both influenced by factors such as SocioEconomic Status (SES), parental education, and geographic location, among others. Thus, it is not clear how much of any observed relationship between X and Y will be due to X having a causal effect on Y (if there is even an effect at all) and how much is simply a result of the shared cause, U.

In order to formalize these notions, we introduce counterfactual outcome variables (also called Potential Outcomes), which we denote by $Y(x)$. These are to be interpreted as the value of Y that an individual would have had had they been exposed to potential cause x, rather than the exposure that they actually experienced, e.g. starting medication or not, the item being priced at x dollars, or going to college or not. One might imagine this as defining two (or more) realities, one for each of the possible exposures that an individual might have experienced, that are the same in every way, except for the exposure at time 0. For example, in the first case, $Y(1)$ (or $Y(x = 1)$) would correspond to whether the individual were still alive in 10

years if they started the medication at time 0, in the second, $Y(x)$ might be how many of an item the individual would have purchased over the course of a year if the price were x dollars, and, in the third, $Y(0)$ would correspond to an individual's lifetime earnings if they did not go to college.

In order for causal inference to be possible, we need a condition referred to as Consistency. Consistency states that if an individual experiences exposure x (in the real world), their outcome Y will be the same as their counterfactual outcome $Y(x)$. (We can also write this as $Y = Y(X)$.) This is necessary to allow us to connect observed outcomes to counterfactual outcomes. With these preliminaries, we are now ready to explore when the causal effect of X on Y (which may be 0) can be identified, meaning that it can be unambiguously determined from the distribution of the observed data. Unfortunately, it is impossible to do this at the individual level (without being able to peer into alternate realities), but, under certain conditions, it will be possible to estimate the expected causal effect at the population level.

In our example, this corresponds to determining $\mathrm{E}Y(x)$ for all possible values of x. If U is unknown or unmeasured, this will be impossible, however, if we know the value of U, we can determine the expected value as follows.

We first note that $Y(x) \perp\!\!\!\perp X|U$. This is because since $Y(x)$ is a counterfactual variable that does not depend on the value of X that the individual actually experienced, and, thus, the only association between X and $Y(x)$ must be via the shared influence of U. One can see this by examining Figure 3b, but it is perhaps even clearer from Figure 4. Figure 4 is called a Single-World Intervention Graph (SWIG); such graphs were introduced by Richardson and Robins in order to make the task of identifying independences easier (Richardson and Robins, 2013). For any intervention, we can produce a SWIG from the corresponding DAG by splitting the node corresponding to the relevant exposure into two parts. The first part, on the left, is attached to all the incoming arrows to the original node, while the second part, on the right, is attached to all the outgoing arrows. We label the left node with the original random variable, in this case X, and the right node with the corresponding small letter, here x, to indicate that it is being set to a specific value by the intervention. Finally, we label all downstream nodes with their counterfactual equivalents, in this case $Y(x)$.

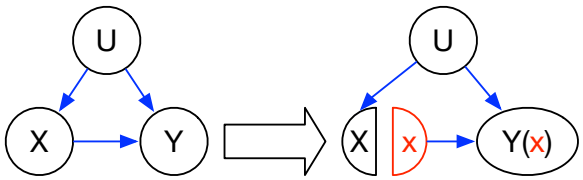

Figure 4. For any intervention, a SWIG can be made from a DAG by splitting the relevant node in half, directing incoming arrows into the left (X) half and outgoing arrows out of the right (x) half, and relabeling downstream nodes with their counterfactual equivalents ($Y(x)$).

From the SWIG it is easy to see that $Y(x) \perp\!\!\!\perp X|U$, since there is no path from X to $Y(x)$ that does not pass through U. We say that conditioning on U blocks the path.

The fact that $Y(x) \perp\!\!\!\perp X|U$ suggests using the law of iterated expectations, to introduce conditioning on U:

$$\mathrm{E}Y(x) = \mathrm{E}\left[\mathrm{E}\left[Y(x)|U\right]\right]$$

Then, since $Y(x) \perp\!\!\!\perp X|U$,

$$\mathrm{E}\left[Y(x)|U\right] = \mathrm{E}\left[Y(x)|U, X = x\right] = \mathrm{E}\left[Y|U, X = x\right]$$

where we have used consistency in the last equality. Then,

$$\mathrm{E}Y(x) = \mathrm{E}\left[\mathrm{E}\left[Y|U, X = x\right]\right]$$

This expression is now purely in terms of variables that we can observe rather than counterfactuals, which we cannot. Thus, if we knew U, we could identify the expected causal effect of X on Y for each value of x. However, we don't actually need to know all of U, or even U itself. We merely need to know enough to block the path from X to U to Y. Figure 5 shows two examples where another variable, V, is on the path either from U to X or from U to Y.

Conditioning on V, in either case, is enough to block the path from X to U to Y, meaning that $Y(x) \perp\!\!\!\perp X|V$, and, thereby, allow us to identify the causal effect. Then, we can identify $\mathrm{E}Y(x)$, as above, instead conditioning on V in place of U.

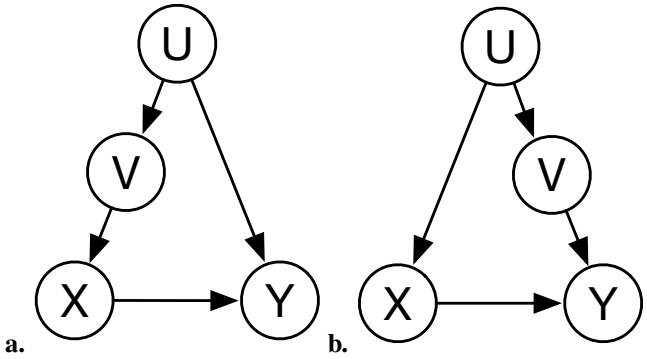

Figure 5. In both DAGs, the path from X to U to Y can be blocked by conditioning on V by blocking either (**a.**) the path from U to X or (**b.**) the path from U to Y.

Then,

$$\mathrm{E}Y(x) = \mathrm{E}\left[\mathrm{E}\left[Y|V, X = x\right]\right]$$

Thus, it is good enough to measure V, even if we can't measure U. This is the identification strategy in traditional causal inference: measure enough variables to block the confounding path from X to U to Y and then identify the causal effect, as above. We can interpret $\mathrm{E}Y(x)$ as the expected outcome if everyone in the population experienced exposure x, e.g. started treatment, could buy the item for x dollars, or didn't go to college. We can also consider causal contrasts, such as $\mathrm{E}Y(1) - \mathrm{E}Y(0)$, which correspond to questions such as, what would the expected 10-year mortality be if everyone started taking a medication vs. if no one did or what would a society's average lifetime earnings be if everyone went to college vs. if no one did. These contrasts are typically what arise when trying to answer policy questions.

Though we have only considered the simplest case, this strategy is quite general, although, in more complicated graphs, when the causal effect can be identified, and what paths need to be blocked (and variables measured), can rapidly become quite complicated. However, the primary difficulty in causal inference is that one can never know with certainty the causal structure of a given system, particularly, what unmeasured confounders may be common causes of X and Y. Thus, a large part of whether or not we can successfully identify a causal effect depends on whether we can correctly determine the relevant causal structure and (possibly unmeasured) confounders, as well as whether we can find sufficient additional variables that we can measure to control for confounding. For more on this see such references as (Pearl, 2000).

A.1. Instrumental Variables

Sometimes we are unable to determine the unmeasured confounders or find enough other things to measure to block the confounding. In such cases, other techniques for causal inference can be useful. One of the oldest is Instrumental Variables (IV), which is widely used in econometrics. An Instrument is a variable, Z, that is associated with the potential cause, X, and not associated with the outcome, Y, except for its association with X, the so-called exclusion restriction. The causal structure of most IV problems is shown in Figure 6.

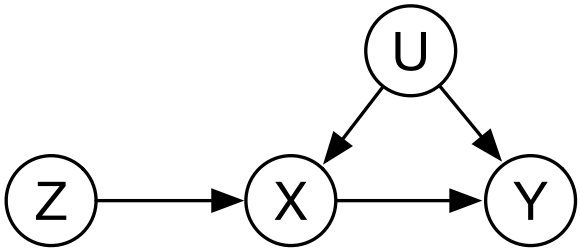

Figure 6. DAG for a standard IV model. X is a potential cause, Y is the outcome of interest, Z is an instrument, and U is an unmeasured confounder.

There are different ways to interpret the role of the instrument. One traditional way of viewing it is as a source that causes

changes in X that are not confounded by U. The stronger the instrument, which, in the classical, linear case, corresponds to a higher correlation between X and Z, the more that changes in Z induce changes in X and the more that changes in Y indirectly induced by Z through X, reflect the effect of the unconfounded part of X. Thus, strong instruments act as good surrogates for the part of X that is not caused by U.

Another way to view instruments, which is particularly relevant when many instruments are used (or in nonparametric IV), is simply as a tool to remove the effects of U from X. IV regression can be viewed as first projecting onto Z and then using this "unconfounded" part of X to perform the actual regression. The stronger the instrument, the less information about X is lost during the projection, so the better the final regression performs.

Nonparametric IV formalizes the notion of information loss using the concept of completeness, which, informally, ensures that Z, and functions of Z, provide a rich enough space to capture all the relevant (unconfounded) information in X. Formally, the completeness condition says that for all $f \in L^2_{\mathcal{P}_X}$, $\mathrm{E}\left[f(X)|Z\right] = 0$ if and only if $f = 0$ $a.s.$

In the classical case, as long as the assumptions hold, one can estimate the causal effect via a variety of methods. In the nonparametric case, the causal effect is estimated via the solution to an integral equation, although, a multitude of solution techniques have been proposed.

A.2. Proximal Inference

An interesting recent advance in causal inference has been the development of proximal inference by Tchetgen Tchetgen and colleagues (Miao et al., 2018; Ghassami et al., 2022; Cui et al., 2022), with some notable recent improvements by Kallus and colleagues (Kallus et al., 2022). In situations where traditional methods for causal inference may not apply, say because we cannot measure enough additional covariates to block the confounding path from X to Y, and no good instruments exist, one can still use proximal inference, since it is based on a very different approach than either classical causal inference or IV.

The idea behind proximal inference is that, if the system obeys certain independence relations, and one can obtain two good surrogates W and Z for the unmeasured confounders U that satisfy these restrictions, then we can estimate the causal effect by solving an integral equation. Proximal inference is compatible with a number of graphs; one of the most common is shown in Figure 7.

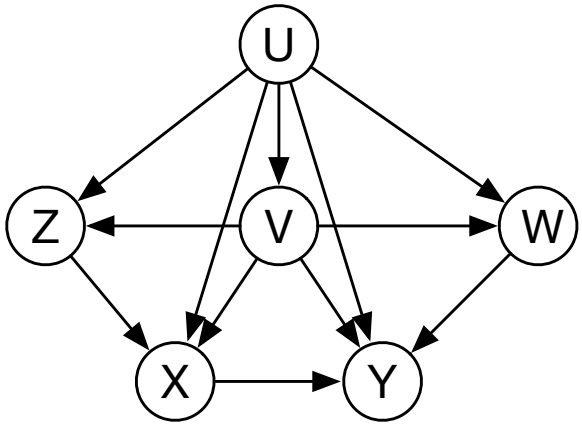

Figure 7. DAG compatible with proximal inference. X is a potential cause, Y is the outcome of interest, W is the outcome proxy, Z is the exposure proxy, U is an unmeasured confounder, and V is additional covariates.

Formally, we must have the following independences: $T \perp\!\!\!\perp Z|X, U, V$ and $W \perp\!\!\!\perp (X, Z)|U, V$.

In order to identify the causal effect, we also need some additional conditions. Kallus and colleagues, explore a variety of these conditions (Kallus et al., 2022). One reasonable set of assumptions, which is closely related to the assumptions made in nonparametric IV, is a pair of completeness assumptions, which can be interpreted to mean that the two proxies, W and Z, each contain all of the information from U. Formally, we have:

$$\text{For all } f \in L^2_{\mathrm{P}_U}, \text{ and all } x \in \mathcal{X}, v \in \mathcal{V},$$

$$\mathrm{E}\left[f(U)|X = x, V = v, Z = z\right] = 0 \text{ for all } z \in \mathcal{Z} \text{ if and only if } f(U) = 0 \text{ a.s.}$$

and

$$\mathrm{E}\left[f(U)|X=x,V=v,W=w\right]=0 \text{ for all } w\in\mathcal{W} \text{ if and only if } f(U)=0 \text{ a.s.}$$

In practice, the conditions for all $z \in \mathcal{Z}$ and all $w \in \mathcal{W}$ can be weakened to almost surely.

Given any valid set of additional assumptions, we can estimate the causal effect in two steps. First we must solve an integral equation for an auxiliary function, h, called the bridge function:

$$\mathrm{E}\left[Y-h(x,w,v)|X=x,Z=z,V=v\right]=0$$

We can then estimate the causal effect as,

$$\mathrm{E}Y(x)=\mathrm{E}[h(x,W,V)]$$

Thus, the bridge function takes on the role played by the conditional expectation, $\mathrm{E}[Y|V,X=x]$, in classical causal inference.

While proximal inference seems miraculous in its ability to identify the causal effect without worrying about blocking all the confounding paths between X and Y, it is important to note that the conditional independence assumptions, as well as the additional completeness assumptions (or their equivalents), are quite restrictive, making it very difficult to find valid proxies in most observational settings.

Additionally, using proximal inference requires solving an integral equation to estimate the bridge function, which presents its own challenges. A variety of such methods have been proposed. We describe one particular method, NMMR, below.

A.2.1. Neural Maximum Moment Restriction (NMMR)

NMMR is a recent approach to estimating the bridge function that relies on a combination of kernels and neural nets (Kompa et al., 2022) Like earlier work from Zhang, et al. and Mastouri, et al., NMMR transforms the problem of solving the integral equation into a convex minimization problem (Zhang et al., 2021; Mastouri et al., 2021) Like its predecessors, it does this by noting that, since the integral equation can be written in the form of a conditional expectation, which is equal to zero for all values of x, v, z,

$$\mathrm{E}\left[Y-h(x,w,v)|X=x,Z=z,V=v\right]=0$$

it must also be the case that multiplying this quantity by any function $g:\mathcal{X}\times\mathcal{Z}\times\mathcal{V}\rightarrow\mathbb{R}$, $g\in\mathcal{G}$ must also be zero, so

$$\mathrm{E}\left[Y-h(x,w,v)|X=x,Z=z,V=v\right]g(x,z,v)=0$$

Thus, the supremum of this expression over $\mathcal{G}$ must also be equal to zero.

$$\sup_{g\in\mathcal{G}}\mathrm{E}\left[Y-h(x,w,v)|X=x,Z=z,V=v\right]g(x,z,v)=0$$

Finally, due to linearity, it is sufficient to consider only those g of norm ≤ 1, yielding

$$\sup_{g:\|g\|\leq 1}\mathrm{E}\left[Y-h(x,w,v)|X=x,Z=z,V=v\right]g(x,z,v)=0$$

Mastouri, et al. show that when $g\in\mathcal{G}$ and $\mathcal{G}$ is a Reproducing Kernel Hilbert Space (RKHS), the left hand side, $R_k(h)$ has a closed form (Mastouri et al., 2021),

$$R_k(h)=\mathrm{E}\left[\left(Y-h\left(X,W,V\right)\right)\left(Y'-h\left(X',W',V'\right)\right)k\left(\left(X,Z,V\right)\left(X',Z',V'\right)\right)\right]$$

and, thus, we can estimate h using ERM with the following loss function,

$$R_k(h) = \sum_{i \neq j} \left(y_i - h\left(x_i, w_i, v_i\right)\right)\left(y_j - h\left(x_j, w_j, v_j\right)\right) k\left(\left(x_i, z_i, v_i\right)\left(x_j, z_j, v_j\right)\right)$$

which is a U-statistic. NMMR implements h using a neural network, since neural nets can function as very flexible function approximators due to universality. Thus, the final estimator combines a kernel-based U-statistic, with neural net based bridge functions. This makes it straightforward to analyze using Theorem 3.15, which results in Theorem 4.1, which represents a significant improvement over the original generalization bound given in Kompa, et al (Kompa et al., 2022).

B. An Introduction to U-Statistics

In this section, we provide a brief discussion of the history and use of U-statistics in statistics and machine learning. U-statistics were first proposed by Wassily Hoeffding in what is considered one of statistics' seminal papers, not only for its originality and thoroughness, but for the impact it had on how future statisticians approached the field (Hoeffding, 1948). The U in U-statistics stands for unbiased, and, indeed, given any statistic of the form $\mathrm{E}f\left(X_1, \ldots X_r\right)$, we can produce an efficient empirical estimator in the form of a U-statistic.

For any function $f : \mathcal{X}^m \to \mathbb{R}$, which is symmetric in its arguments, we can define a U-statistic of order m with kernel f, as follows. Given a sample $(X_1, \ldots, X_n) \in \mathcal{X}^n$ we define $\mathbb{U}_n(f) = \frac{n!}{(n-m)!} \sum_{i_1 \neq \ldots \neq i_m}^{n} f(X_{i_1}, \ldots, X_{i_m})$. This will be an unbiased estimator of $\mathbb{E}(f(X_1, \ldots, X_m))$ provided the sample $(X_1, \ldots, X_n) \in \mathcal{X}^n$ is jointly exchangeable (which holds for i.i.d. sampling). To see this in the IID case, note that, since for each term in the sum, none of the arguments can be the same (since they cannot share the same indices), then the expectation of every term must be $\mathrm{E}f\left(X_1, \ldots, X_m\right)$ and, thus, $\mathrm{E}\mathbb{U}_n(f) = \mathrm{E}f\left(X_1, \ldots, X_m\right)$, as well. Also note that the requirement that f be symmetric imposes no burden, since we can always replace any nonsymmetric f by a symmetrized version of itself.

Many common statistics including the sample mean and variance, as well as a variety of test statistics, can naturally be written as U-statistics, which makes their study important for understanding the behavior of many widely used sample estimators. However, in addition, the unbiasedness of U-statistics has led to them finding many other creative uses in both statistics and machine learning. We discuss three such applications here.

The first class of applications pertains to hypothesis tests with high dimensional data that are used to examine marginal or low-dimensional features of a high-dimensional joint distributions – such as testing of mean vectors, covariance matrices and regression coefficients. In this regard, (He et al., 2021) considers a large collection of examples to demonstrate the usefulness of a class of U-statistics based statistical methods. Specifically, (He et al., 2021) and references therein consider quantities like $\mathcal{E}_{l,a} = \sum_{l \in \mathcal{L}} \mathbf{e}_l^a$ where $\mathbf{e}_l$'s are generic parameters with $a \in \mathbb{N}$. For any such quantity one can find an unbiased estimator utilizing a U-statistic of the form $\sum_{l \in \mathcal{L}} \frac{n|}{(n-\gamma_{l,a})!} \sum_{i_1 \neq \ldots \neq \gamma_{l,a}} K(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_{\gamma_{l,a}}})$ where the choice of the kernel K and order $\gamma_{l,a}$ changes according to the specific problem. One can then consider hypothesis testing using $H_0 : \mathcal{E}_{l,a} = 0$. The collection of examples that falls into this class includes high dimensional mean testing, covariance testing, and many other useful applications. Similarly, U-statistics based methods have also become popular for the independence testing of high dimensional vectors – see e.g. (Albert et al., 2022; Han et al., 2017) and references therein.

A second class of applications that benefits from the rich theory of U-statistics is the problem of ranking. In ranking, one considers i.i.d. data $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $X \in \mathcal{X}$ are feature vectors and $Y \in \{-1, +1\}$ are their corresponding labels. The goal is to design a ranking rule that predicts the relative magnitude of the Y's based on the X's. In order to do so, one considers risk functions of the form $L(r) = \mathbb{P}\left(\left\{\frac{Y-Y'}{2}\right\} r(X, X') < 0\right)$ where $(X, Y), (X', Y') \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ are any two independent draws and $r(X, X') \in \{-1, +1\}$ is any ranking rule. Since the population level ranking loss involves two independent draws from the population, the natural sample version of the problem involves minimizing a second order U-statistic $\hat{L}_r = \frac{n!}{(n-2)!} \sum_{i_1 \neq i_2} \mathbf{1}\left(\frac{Y_{i_1} - Y_{i_2}}{2} \cdot r(X_{i_1}, X_{i_2}) < 0\right)$ over a class of ranking rules r. This problem is naturally covered by our results on U-statistics from the main text. For further details, the interested reader can find a detailed exposition in (Clemençon et al., 2005; Clémençon et al., 2006; Clémençon et al., 2008; 2016) and references therein.

Last but not the least, U-statistics can be used to convert the problem of finding the zero of a linear equation into a convex minimization problem, which can often be easier to solve. This is frequently useful when considering estimators that arise as the solution of integral equations. Such estimators are common in areas such as causal inference using observational data. In both nonparametric IV and proximal inference one must solve an integral equation in order to obtain the estimator of interest. These expressions can typically be rewritten as a family of linear moment equations. Through the use of RKHS methods, the problem of finding the zeros of these linear functionals can instead be transformed into the problem of minimizing a bilinear form, which is naturally a second order U-statistic. This strategy has been widely applied, beginning with work in the nonparametric IV literature and continuing into proximal inference (Zhang et al., 2021; Mastouri et al., 2021; Kompa et al., 2022). We discuss this in more detail in the previous Section A.2.1.