Journal of Machine Learning Research 14 (2013) 3207-3260 Submitted 9/12; Revised 3/13; Published 11/13

Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising

Léon Bottou LEON@BOTTOU.ORG
Microsoft
1 Microsoft Way
Redmond, WA 98052, USA

Jonas Peters* PETERS@STAT.MATH.ETHZ.CH
Max Planck Institute
Spemannstraße 38
72076 Tübingen, Germany

Joaquin Quiñonero-Candela† JQUINONERO@GMAIL.COM
Denis X. Charles CDX@MICROSOFT.COM
D. Max Chickering DMAX@MICROSOFT.COM
Elon Portugaly ELONP@MICROSOFT.COM
Dipankar Ray DIPANRAY@MICROSOFT.COM
Patrice Simard PATRICE@MICROSOFT.COM
Ed Snelson EDSNELSO@MICROSOFT.COM
Microsoft
1 Microsoft Way
Redmond, WA 98052, USA

Abstract

This work shows how to leverage causal inference to understand the behavior of complex learning systems interacting with their environment and predict the consequences of changes to the system. Such predictions allow both humans and algorithms to select the changes that would have improved the system performance. This work is illustrated by experiments on the ad placement system associated with the Bing search engine.

Keywords: causation, counterfactual reasoning, computational advertising

1. Introduction

Statistical machine learning technologies in the real world are never without a purpose. Using their predictions, humans or machines make decisions whose circuitous consequences often violate the modeling assumptions that justified the system design in the first place.

Such contradictions appear very clearly in the case of the learning systems that power web scale applications such as search engines, ad placement engines, or recommendation systems. For instance, the placement of advertisement on the result pages of Internet search engines depend on the bids of advertisers and on scores computed by statistical machine learning systems. Because the scores affect the contents of the result pages proposed to the users, they directly influence the occurrence of clicks and the corresponding advertiser payments. They also have important indirect effects. Ad placement decisions impact the satisfaction of the users and therefore their willingness to frequent this web site in the future. They also impact the return on investment observed by the

*. Current address: *Jonas Peters, ETH Zürich, Rämistraße 101, 8092 Zürich, Switzerland.*
†. Current address: *Joaquin Quiñonero-Candela, Facebook, 1 Hacker Way, Menlo Park, CA 94025, USA.*

advertisers and therefore their future bids. Finally they change the nature of the data collected for training the statistical models in the future.

These complicated interactions are clarified by important theoretical works. Under simplified assumptions, mechanism design (Myerson, 1981) leads to an insightful account of the advertiser feedback loop (Varian, 2007; Edelman et al., 2007). Under simplified assumptions, multiarmed bandits theory (Robbins, 1952; Auer et al., 2002; Langford and Zhang, 2008) and reinforcement learning (Sutton and Barto, 1998) describe the exploration/exploitation dilemma associated with the training feedback loop. However, none of these approaches gives a complete account of the complex interactions found in real-life systems.

This contribution proposes a novel approach: we view these *complicated interactions as manifestations of the fundamental difference that separates correlation and causation.* Using the ad placement example as a model of our problem class, we therefore argue that *the language and the methods of causal inference* provide flexible means to *describe such complex machine learning systems* and *give sound answers to the practical questions* facing the designer of such a system. Is it useful to pass a new input signal to the statistical model? Is it worthwhile to collect and label a new training set? What about changing the loss function or the learning algorithm? In order to answer such questions and improve the operational performance of the learning system, one needs to unravel how the information produced by the statistical models traverses the web of causes and effects and eventually produces measurable performance metrics.

Readers with an interest in causal inference will find in this paper (*i*) a *real world example demonstrating the value of causal inference for large-scale machine learning applications*, (*ii*) *causal inference techniques applicable to continuously valued variables with meaningful confidence intervals*, and (*iii*) *quasi-static analysis techniques for estimating how small interventions affect certain causal equilibria*. Readers with an interest in real-life applications will find (*iv*) a selection of *practical counterfactual analysis techniques applicable to many real-life machine learning systems*. Readers with an interest in computational advertising will find a principled framework that (*v*) explains *how to soundly use machine learning techniques for ad placement*, and (*vi*) *conceptually connects machine learning and auction theory* in a compelling manner.

The paper is organized as follows. Section 2 gives an overview of the advertisement placement problem which serves as our main example. In particular, we stress some of the difficulties encountered when one approaches such a problem without a principled perspective. Section 3 provides a condensed review of the essential concepts of causal modeling and inference. Section 4 centers on formulating and answering counterfactual questions such as "how would the system have performed during the data collection period if certain interventions had been carried out on the system ?" We describe importance sampling methods for counterfactual analysis, with clear conditions of validity and confidence intervals. Section 5 illustrates how the structure of the causal graph reveals opportunities to exploit prior information and vastly improve the confidence intervals. Section 6 describes how counterfactual analysis provides essential signals that can drive learning algorithms. Assume that we have identified interventions that would have caused the system to perform well during the data collection period. Which guarantee can we obtain on the performance of these same interventions in the future? Section 7 presents counterfactual differential techniques for the study of equlibria. Using data collected when the system is at equilibrium, we can estimate how a small intervention displaces the equilibrium. This provides an elegant and effective way to reason about long-term feedback effects. Various appendices complete the main text with information that we think more relevant to readers with specific backgrounds.

2. Causation Issues in Computational Advertising

After giving an overview of the advertisement placement problem, which serves as our main example, this section illustrates some of the difficulties that arise when one does not pay sufficient attention to the causal structure of the learning system.

2.1 Advertisement Placement

All Internet users are now familiar with the advertisement messages that adorn popular web pages. Advertisements are particularly effective on search engine result pages because users who are searching for something are good targets for advertisers who have something to offer. Several actors take part in this Internet advertisement game:

- Advertisers create advertisement messages, and place bids that describe how much they are willing to pay to see their ads displayed or clicked.

- Publishers provide attractive web services, such as, for instance, an Internet search engine. They display selected ads and expect to receive payments from the advertisers. The infrastructure to collect the advertiser bids and select ads is sometimes provided by an advertising network on behalf of its affiliated publishers. For the purposes of this work, we simply consider a publisher large enough to run its own infrastructure.

- Users reveal information about their current interests, for instance, by entering a query in a search engine. They are offered web pages that contain a selection of ads (Figure 1). Users sometimes click on an advertisement and are transported to a web site controlled by the advertiser where they can initiate some business.

A conventional bidding language is necessary to precisely define under which conditions an advertiser is willing to pay the bid amount. In the case of Internet search advertisement, each bid specifies (a) the advertisement message, (b) a set of keywords, (c) one of several possible matching criteria between the keywords and the user query, and (d) the maximal price the advertiser is willing to pay when a user clicks on the ad after entering a query that matches the keywords according to the specified criterion.

Whenever a user visits a publisher web page, an advertisement placement engine runs an auction in real time in order to select winning ads, determine where to display them in the page, and compute the prices charged to advertisers, should the user click on their ad. Since the placement engine is operated by the publisher, it is designed to further the interests of the publisher. Fortunately for everyone else, the publisher must balance short term interests, namely the immediate revenue brought by the ads displayed on each web page, and long term interests, namely the future revenues resulting from the continued satisfaction of both users and advertisers.

Auction theory explains how to design a mechanism that optimizes the revenue of the seller of a single object (Myerson, 1981; Milgrom, 2004) under various assumptions about the information available to the buyers regarding the intentions of the other buyers. In the case of the ad placement problem, the publisher runs multiple auctions and sells opportunities to receive a click. When nearly identical auctions occur thousand of times per second, it is tempting to consider that the advertisers have perfect information about each other. This assumption gives support to the popular generalized second price rank-score auction (Varian, 2007; Edelman et al., 2007):

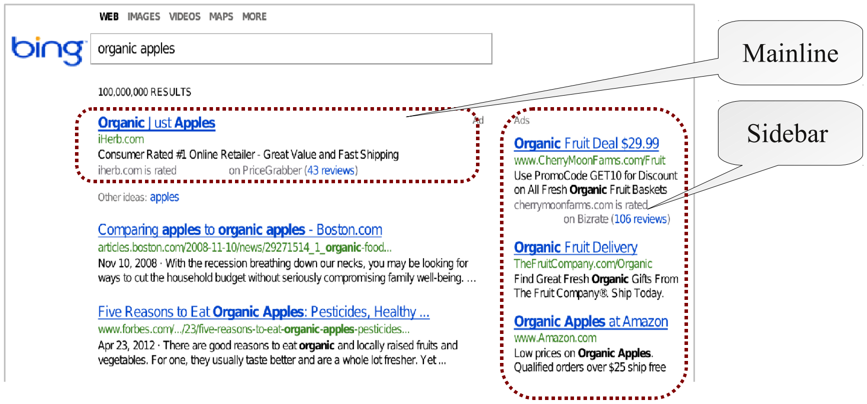

Figure 1: Mainline and sidebar ads on a search result page. Ads placed in the mainline are more likely to be noticed, increasing both the chances of a click if the ad is relevant and the risk of annoying the user if the ad is not relevant.

- Let x represent the auction context information, such as the user query, the user profile, the date, the time, etc. The ad placement engine first determines all eligible ads $a_1 \dots a_n$ and the corresponding bids $b_1 \dots b_n$ on the basis of the auction context x and of the matching criteria specified by the advertisers.

- For each selected ad a_i and each potential position p on the web page, a statistical model outputs the estimate $q_{i,p}(x)$ of the probability that ad a_i displayed in position p receives a user click. The rank-score $r_{i,p}(x) = b_i q_{i,p}(x)$ then represents the purported value associated with placing ad a_i at position p.

- Let L represent a possible ad layout, that is, a set of positions that can simultaneously be populated with ads, and let $\mathscr{L}$ be the set of possible ad layouts, including of course the empty layout. The optimal layout and the corresponding ads are obtained by maximizing the total rank-score
$$\max_{L \in \mathscr{L}} \max_{i_1, i_2, \dots} \sum_{p \in L} r_{i_p,p}(x)\,, \tag{1}$$
subject to reserve constraints
$$\forall p \in L,\ r_{i_p,p}(x) \geq R_p(x)\,,$$
and also subject to diverse policy constraints, such as, for instance, preventing the simultaneous display of multiple ads belonging to the same advertiser. Under mild assumptions, this discrete maximization problem is amenable to computationally efficient greedy algorithms (see appendix A.)

- The advertiser payment associated with a user click is computed using the generalized second price (GSP) rule: the advertiser pays the smallest bid that it could have entered without changing the solution of the discrete maximization problem, all other bids remaining equal. In other words, the advertiser could not have manipulated its bid and obtained the same treatment for a better price.

Under the perfect information assumption, the analysis suggests that the publisher simply needs to find which reserve prices $R_p(x)$ yield the best revenue *per auction*. However, the total revenue of the publisher also depends on the traffic experienced by its web site. Displaying an excessive number of irrelevant ads can train users to ignore the ads, and can also drive them to competing web sites. Advertisers can artificially raise the rank-scores of irrelevant ads by temporarily increasing the bids. Indelicate advertisers can create deceiving advertisements that elicit many clicks but direct users to spam web sites. Experience shows that the continued satisfaction of the users is more important to the publisher than it is to the advertisers.

Therefore the generalized second price rank-score auction has evolved. Rank-scores have been augmented with terms that quantify the user satisfaction or the ad relevance. Bids receive adaptive discounts in order to deal with situations where the perfect information assumption is unrealistic. These adjustments are driven by additional statistical models. The ad placement engine should therefore be viewed as a complex learning system interacting with both users and advertisers.

2.2 Controlled Experiments

The designer of such an ad placement engine faces the fundamental question of testing whether a proposed modification of the ad placement engine results in an improvement of the operational performance of the system.

The simplest way to answer such a question is to try the modification. The basic idea is to randomly split the users into treatment and control groups (Kohavi et al., 2008). Users from the control group see web pages generated using the unmodified system. Users of the treatment groups see web pages generated using alternate versions of the system. Monitoring various performance metrics for a couple months usually gives sufficient information to reliably decide which variant of the system delivers the most satisfactory performance.

Modifying an advertisement placement engine elicits reactions from both the users and the advertisers. Whereas it is easy to split users into treatment and control groups, splitting advertisers into treatment and control groups demands special attention because each auction involves multiple advertisers (Charles et al., 2012). Simultaneously controlling for both users and advertisers is probably impossible.

Controlled experiments also suffer from several drawbacks. They are expensive because they demand a complete implementation of the proposed modifications. They are slow because each experiment typically demands a couple months. Finally, although there are elegant ways to efficiently run overlapping controlled experiments on the same traffic (Tang et al., 2010), they are limited by the volume of traffic available for experimentation.

It is therefore difficult to rely on controlled experiments during the conception phase of potential improvements to the ad placement engine. It is similarly difficult to use controlled experiments to drive the training algorithms associated with click probability estimation models. Cheaper and faster statistical methods are needed to drive these essential aspects of the development of an ad placement engine. Unfortunately, interpreting cheap and fast data can be very deceiving.

2.3 Confounding Data

Assessing the consequence of an intervention using statistical data is generally challenging because it is often difficult to determine whether the observed effect is a simple consequence of the intervention or has other uncontrolled causes.

	Overall	Patients with small stones	Patients with large stones
Treatment A: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment B: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

Table 1: A classic example of Simpson's paradox. The table reports the success rates of two treatments for kidney stones (Charig et al., 1986, Tables I and II). Although the overall success rate of treatment B seems better, treatment B performs worse than treatment A on both patients with small kidney stones and patients with large kidney stones. See Section 2.3.

For instance, the empirical comparison of certain kidney stone treatments illustrates this difficulty (Charig et al., 1986). Table 2.3 reports the success rates observed on two groups of 350 patients treated with respectively open surgery (treatment A, with 78% success) and percutaneous nephrolithotomy (treatment B, with 83% success). Although treatment B seems more successful, it was more frequently prescribed to patients suffering from small kidney stones, a less serious condition. Did treatment B achieve a high success rate because of its intrinsic qualities or because it was preferentially applied to less severe cases? Further splitting the data according to the size of the kidney stones reverses the conclusion: treatment A now achieves the best success rate for both patients suffering from large kidney stones and patients suffering from small kidney stones. Such an inversion of the conclusion is called Simpson's paradox (Simpson, 1951).

The stone size in this study is an example of a *confounding variable*, that is an uncontrolled variable whose consequences pollute the effect of the intervention. Doctors knew the size of the kidney stones, chose to treat the healthier patients with the least invasive treatment B, and therefore caused treatment B to appear more effective than it actually was. If we now decide to apply treatment B to all patients irrespective of the stone size, we break the causal path connecting the stone size to the outcome, we eliminate the illusion, and we will experience disappointing results.

When we suspect the existence of a confounding variable, we can split the contingency tables and reach improved conclusions. Unfortunately we cannot fully trust these conclusions unless we are certain to have taken into account all confounding variables. The real problem therefore comes from the confounding variables we do not know.

Randomized experiments arguably provide the only correct solution to this problem (see Stigler, 1992). The idea is to randomly chose whether the patient receives treatment A or treatment B. Because this random choice is independent from all the potential confounding variables, known and unknown, they cannot pollute the observed effect of the treatments (see also Section 4.2). This is why controlled experiments in ad placement (Section 2.2) randomly distribute users between treatment and control groups, and this is also why, in the case of an ad placement engine, we should be somehow concerned by the practical impossibility to randomly distribute both users and advertisers.

	Overall	q_2 low	q_2 high
q_1 low	6.2% (124/2000)	**5.1%** (92/1823)	**18.1%** (32/176)
q_1 high	**7.5%** (149/2000)	4.8% (71/1500)	15.6% (78/500)

Table 2: Confounding data in ad placement. The table reports the click-through rates and the click counts of the second mainline ad. The overall counts suggest that the click-through rate of the second mainline ad increases when the click probability estimate q_1 of the top ad is high. However, if we further split the pages according to the click probability estimate q_2 of the second mainline ad, we reach the opposite conclusion. See Section 2.4.

2.4 Confounding Data in Ad Placement

Let us return to the question of assessing the value of passing a new input signal to the ad placement engine click prediction model. Section 2.1 outlines a placement method where the click probability estimates $q_{i,p}(x)$ depend on the ad and the position we consider, but do not depend on other ads displayed on the page. We now consider replacing this model by a new model that additionally uses the estimated click probability of the top mainline ad to estimate the click probability of the second mainline ad (Figure 1). We would like to estimate the effect of such an intervention using existing statistical data.

We have collected ad placement data for Bing search result pages served during three consecutive hours on a certain slice of traffic. Let q_1 and q_2 denote the click probability estimates computed by the existing model for respectively the top mainline ad and the second mainline ad. After excluding pages displaying fewer than two mainline ads, we form two groups of 2000 pages randomly picked among those satisfying the conditions $q_1 < 0.15$ for the first group and $q_1 \geq 0.15$ for the second group. Table 2.4 reports the click counts and frequencies observed on the second mainline ad in each group. Although the overall numbers show that users click more often on the second mainline ad when the top mainline ad has a high click probability estimate q_1, this conclusion is reversed when we further split the data according to the click probability estimate q_2 of the second mainline ad.

Despite superficial similarities, this example is considerably more difficult to interpret than the kidney stone example. The overall click counts show that the actual click-through rate of the second mainline ad is positively correlated with the click probability estimate on the top mainline ad. Does this mean that we can increase the total number of clicks by placing regular ads below frequently clicked ads?

Remember that the click probability estimates depend on the search query which itself depends on the user intention. The most likely explanation is that pages with a high q_1 are frequently associated with more commercial searches and therefore receive more ad clicks on all positions. The observed correlation occurs because the presence of a click and the magnitude of the click probability estimate q_1 have a common cause: the user intention. Meanwhile, the click probability estimate q_2 returned by the current model for the second mainline ad also depend on the query and therefore the user intention. Therefore, assuming that this dependence has comparable strength, and assuming that there are no other causal paths, splitting the counts according to the magnitude of q_2 factors out the effects of this common confounding cause. We then observe a negative correlation which now suggests that a frequently clicked top mainline ad has a negative impact on the click-through rate of the second mainline ad.

If this is correct, we would probably increase the accuracy of the click prediction model by switching to the new model. This would decrease the click probability estimates for ads placed in the second mainline position on commercial search pages. These ads are then less likely to clear the reserve and therefore more likely to be displayed in the less attractive sidebar. The net result is probably a loss of clicks and a loss of money despite the higher quality of the click probability model. Although we could tune the reserve prices to compensate this unfortunate effect, nothing in this data tells us where the performance of the ad placement engine will land. Furthermore, unknown confounding variables might completely reverse our conclusions.

Making sense out of such data is just too complex !

2.5 A Better Way

It should now be obvious that we need a more principled way to reason about the effect of potential interventions. We provide one such more principled approach using the causal inference machinery (Section 3). The next step is then the identification of a class of questions that are sufficiently expressive to guide the designer of a complex learning system, and sufficiently simple to be answered using data collected in the past using adequate procedures (Section 4).

A machine learning algorithm can then be viewed as an automated way to generate questions about the parameters of a statistical model, obtain the corresponding answers, and update the parameters accordingly (Section 6). Learning algorithms derived in this manner are very flexible: human designers and machine learning algorithms can cooperate seamlessly because they rely on similar sources of information.

3. Modeling Causal Systems

When we point out a causal relationship between two events, we describe what we expect to happen to the event we call the *effect*, should an external operator manipulate the event we call the *cause*. Manipulability theories of causation (von Wright, 1971; Woodward, 2005) raise this commonsense insight to the status of a definition of the causal relation. Difficult adjustments are then needed to interpret statements involving causes that we can only observe through their effects, *"because they love me,"* or that are not easily manipulated, *"because the earth is round."*

Modern statistical thinking makes a clear distinction between the statistical model and the world. The actual mechanisms underlying the data are considered unknown. The statistical models do not need to reproduce these mechanisms to emulate the observable data (Breiman, 2001). Better models are sometimes obtained by deliberately avoiding to reproduce the true mechanisms (Vapnik, 1982, Section 8.6). We can approach the manipulability puzzle in the same spirit by viewing causation as a reasoning model (Bottou, 2011) rather than a property of the world. Causes and effects are simply the pieces of an abstract reasoning game. Causal statements that are not empirically testable acquire validity when they are used as intermediate steps when one reasons about manipulations or interventions amenable to experimental validation.

This section presents the rules of this reasoning game. We largely follow the framework proposed by Pearl (2009) because it gives a clear account of the connections between causal models and probabilistic models.

x	$=$	$f_1(u, \varepsilon_1)$	Query context x from user intent u.
a	$=$	$f_2(x, v, \varepsilon_2)$	Eligible ads (a_i) from query x and inventory v.
b	$=$	$f_3(x, v, \varepsilon_3)$	Corresponding bids (b_i).
q	$=$	$f_4(x, a, \varepsilon_4)$	Scores $(q_{i,p}, R_p)$ from query x and ads a.
s	$=$	$f_5(a, q, b, \varepsilon_5)$	Ad slate s from eligible ads a, scores q and bids b.
c	$=$	$f_6(a, q, b, \varepsilon_6)$	Corresponding click prices c.
y	$=$	$f_7(s, u, \varepsilon_7)$	User clicks y from ad slate s and user intent u.
z	$=$	$f_8(y, c, \varepsilon_8)$	Revenue z from clicks y and prices c.

Figure 2: A structural equation model for ad placement. The sequence of equations describes the flow of information. The functions f_k describe how effects depend on their direct causes. The additional noise variables ε_k represent independent sources of randomness useful to model probabilistic dependencies.

3.1 The Flow of Information

Figure 2 gives a deterministic description of the operation of the ad placement engine. Variable u represents the user and his or her intention in an unspecified manner. The query and query context x is then expressed as an unknown function of the u and of a noise variable ε_1. Noise variables in this framework are best viewed as independent sources of randomness useful for modeling a nondeterministic causal dependency. We shall only mention them when they play a specific role in the discussion. The set of eligible ads a and the corresponding bids b are then derived from the query x and the ad inventory v supplied by the advertisers. Statistical models then compute a collection of scores q such as the click probability estimates $q_{i,p}$ and the reserves R_p introduced in Section 2.1. The placement logic uses these scores to generate the "ad slate" s, that is, the set of winning ads and their assigned positions. The corresponding click prices c are computed. The set of user clicks y is expressed as an unknown function of the ad slate s and the user intent u. Finally the revenue z is expressed as another function of the clicks y and the prices c.

Such a system of equations is named *structural equation model* (Wright, 1921). Each equation asserts a functional dependency between an effect, appearing on the left hand side of the equation, and its direct causes, appearing on the right hand side as arguments of the function. Some of these causal dependencies are *unknown*. Although we postulate that the effect can be expressed as some function of its direct causes, we do not know the form of this function. For instance, the designer of the ad placement engine knows functions f_2 to f_6 and f_8 because he has designed them. However, he does not know the functions f_1 and f_7 because whoever designed the user did not leave sufficient documentation.

Figure 3 represents the directed causal graph associated with the structural equation model. Each arrow connects a direct cause to its effect. The noise variables are omitted for simplicity. The structure of this graph reveals fundamental assumptions about our model. For instance, the user clicks y do not directly depend on the scores q or the prices c because users do not have access to this information.

We hold as a principle that causation obeys the arrow of time: causes always precede their effects. Therefore the causal graph must be *acyclic*. Structural equation models then support two fundamental operations, namely simulation and intervention.

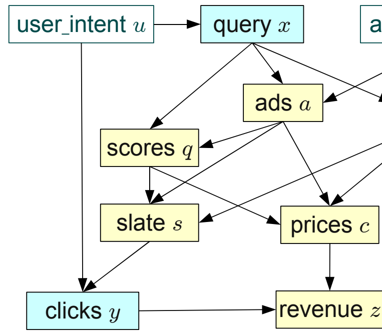

Figure 3: Causal graph associated with the structural equation model of Figure 2. The mutually independent noise variables ε_1 to ε_8 are implicit. The variables a, b, q, s, c, and z depend on their direct causes in known ways. In contrast, the variables u and v are exogenous and the variables x and y depend on their direct causes through unknown functions.

- *Simulation* – Let us assume that we know both the exact form of all functional dependencies and the value of all exogenous variables, that is, the variables that never appear in the left hand side of an equation. We can compute the values of all the remaining variables by applying the equations in their natural time sequence.
- *Intervention* – As long as the causal graph remains acyclic, we can construct derived structural equation models using arbitrary algebraic manipulations of the system of equations. For instance, we can clamp a variable to a constant value by rewriting the right-hand side of the corresponding equation as the specified constant value.

The algebraic manipulation of the structural equation models provides a powerful language to describe interventions on a causal system. This is not a coincidence. Many aspects of the mathematical notation were invented to support causal inference in classical mechanics. However, we no longer have to interpret the variable values as physical quantities: the equations simply describe the flow of information in the causal model (Wiener, 1948).

3.2 The Isolation Assumption

Let us now turn our attention to the exogenous variables, that is, variables that never appear in the left hand side of an equation of the structural model. Leibniz's *principle of sufficient reason* claims that there are no facts without causes. This suggests that the exogenous variables are the effects of a network of causes not expressed by the structural equation model. For instance, the user intent u and the ad inventory v in Figure 3 have temporal correlations because both users and advertisers worry about their budgets when the end of the month approaches. Any structural equation model should then be understood in the context of a larger structural equation model potentially describing all things in existence.

Ads served on a particular page contribute to the continued satisfaction of both users and advertisers, and therefore have an effect on their willingness to use the services of the publisher in the future. The ad placement structural equation model shown in Figure 2 only describes the causal dependencies for a single page and therefore cannot account for such effects. Consider however a very

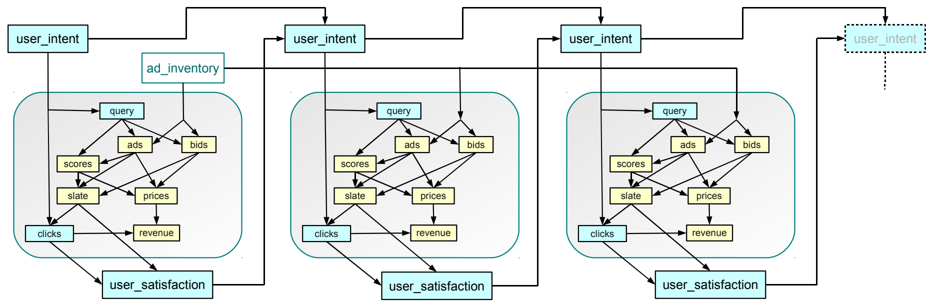

Figure 4: Conceptually unrolling the user feedback loop by threading instances of the single page causal graph (Figure 3). Both the ad slate s_t and user clicks y_t have an indirect effect on the user intent u_{t+1} associated with the next query.

large structural equation model containing a copy of the page-level model for every web page ever served by the publisher. Figure 4 shows how we can thread the page-level models corresponding to pages served to the same user. Similarly we could model how advertisers track the performance and the cost of their advertisements and model how their satisfaction affects their future bids. The resulting causal graphs can be very complex. Part of this complexity results from time-scale differences. Thousands of search pages are served in a second. Each page contributes a little to the continued satisfaction of one user and a few advertisers. The accumulation of these contributions produces measurable effects after a few weeks.

Many of the functional dependencies expressed by the structural equation model are left unspecified. Without direct knowledge of these functions, we must reason using statistical data. The most fundamental statistical data is collected from repeated trials that are assumed independent. When we consider the large structured equation model of everything, we can only have one large trial producing a single data point.[1] It is therefore desirable to identify repeated patterns of identical equations that can be viewed as repeated independent trials. Therefore, when we study a structural equation model representing such a pattern, we need to make an additional assumption to expresses the idea that the outcome of one trial does not affect the other trials. We call such an assumption an *isolation assumption* by analogy with thermodynamics.[2] This can be achieved by assuming that *the exogenous variables are independently drawn from an unknown but fixed joint probability distribution*. This assumption cuts the causation effects that could flow through the exogenous variables.

The noise variables are also exogenous variables acting as independent source of randomness. The noise variables are useful to represent the conditional distribution $\mathrm{P}(\mathsf{effect}\,|\,\mathsf{causes})$ using the equation $\mathsf{effect} = f(\mathsf{causes}, \varepsilon)$. Therefore, we also assume joint independence between all the noise variables and any of the named exogenous variable.[3] For instance, in the case of the ad placement

1. See also the discussion on reinforcement learning, Section 3.5.
2. The concept of isolation is pervasive in physics. An isolated system in thermodynamics (Reichl, 1998, Section 2.D) or a closed system in mechanics (Landau and Lifshitz, 1969, §5) evolves without exchanging mass or energy with its surroundings. Experimental trials involving systems that are assumed isolated may differ in their initial setup and therefore have different outcomes. Assuming isolation implies that the outcome of each trial cannot affect the other trials.
3. Rather than letting two noise variables display measurable statistical dependencies because they share a common cause, we prefer to name the common cause and make the dependency explicit in the graph.

$$
\mathrm{P}\left(\begin{array}{c} u,v,x,a,b \\ q,s,c,y,z \end{array}\right) \quad = \quad \left\{\begin{array}{ll}
\mathrm{P}(u,v) & \text{Exogenous vars.} \\
\times\, \mathrm{P}(x\,|\,u) & \text{Query.} \\
\times\, \mathrm{P}(a\,|\,x,v) & \text{Eligible ads.} \\
\times\, \mathrm{P}(b\,|\,x,v) & \text{Bids.} \\
\times\, \mathrm{P}(q\,|\,x,a) & \text{Scores.} \\
\times\, \mathrm{P}(s\,|\,a,q,b) & \text{Ad slate.} \\
\times\, \mathrm{P}(c\,|\,a,q,b) & \text{Prices.} \\
\times\, \mathrm{P}(y\,|\,s,u) & \text{Clicks.} \\
\times\, \mathrm{P}(z\,|\,y,c) & \text{Revenue.}
\end{array}\right.
$$

Figure 5: Markov factorization of the structural equation model of Figure 2.

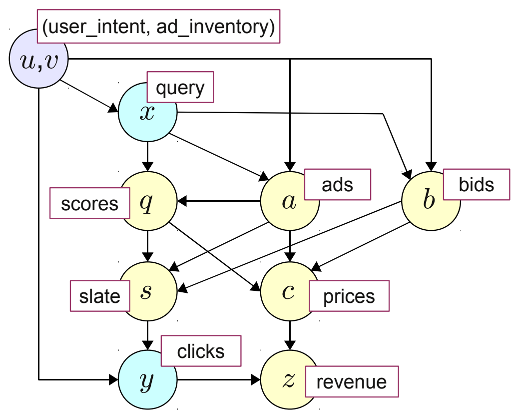

Figure 6: Bayesian network associated with the Markov factorization shown in Figure 5.

model shown in Figure 2, we assume that the joint distribution of the exogenous variables factorizes as

$$\mathrm{P}(u,v,\varepsilon_1,\ldots,\varepsilon_8) = \mathrm{P}(u,v)\,\mathrm{P}(\varepsilon_1)\ldots\mathrm{P}(\varepsilon_8)\,.$$

Since an isolation assumption is only true up to a point, it should be expressed clearly and remain under constant scrutiny. We must therefore measure additional performance metrics that reveal how the isolation assumption holds. For instance, the ad placement structural equation model and the corresponding causal graph (figures 2 and 3) do not take user feedback or advertiser feedback into account. Measuring the revenue is not enough because we could easily generate revenue at the expense of the satisfaction of the users and advertisers. When we evaluate interventions under such an isolation assumption, we also need to measure a battery of additional quantities that act as proxies for the user and advertiser satisfaction. Noteworthy examples include ad relevance estimated by human judges, and advertiser surplus estimated from the auctions (Varian, 2009).

3.3 Markov Factorization

Conceptually, we can draw a sample of the exogenous variables using the distribution specified by the isolation assumption, and we can then generate values for all the remaining variables by simulating the structural equation model.

This process defines a *generative probabilistic model* representing the joint distribution of all variables in the structural equation model. The distribution readily factorizes as the product of the joint probability of the named exogenous variables, and, for each equation in the structural equation model, the conditional probability of the effect given its direct causes (Spirtes et al., 1993; Pearl, 2000). As illustrated by figures 5 and 6, this *Markov factorization* connects the structural equation model that describes causation, and the Bayesian network that describes the joint probability distribution followed by the variables under the isolation assumption.[4]

Structural equation models and Bayesian networks appear so intimately connected that it could be easy to forget the differences. The structural equation model is an algebraic object. As long as the causal graph remains acyclic, algebraic manipulations are interpreted as interventions on the causal system. The Bayesian network is a generative statistical model representing a class of joint probability distributions, and, as such, does not support algebraic manipulations. However, the symbolic representation of its Markov factorization is an algebraic object, essentially equivalent to the structural equation model.

3.4 Identification, Transportation, and Transfer Learning

Consider a causal system represented by a structural equation model with some unknown functional dependencies. Subject to the isolation assumption, data collected during the operation of this system follows the distribution described by the corresponding Markov factorization. Let us first assume that this data is sufficient to identify the joint distribution of the subset of variables we can observe. We can intervene on the system by clamping the value of some variables. This amounts to replacing the right-hand side of the corresponding structural equations by constants. The joint distribution of the variables is then described by a new Markov factorization that shares many factors with the original Markov factorization. Which conditional probabilities associated with this new distribution can we express using only conditional probabilities identified during the observation of the original system? This is called the *identifiability* problem. More generally, we can consider arbitrarily complex manipulations of the structural equation model, and we can perform multiple experiments involving different manipulations of the causal system. Which conditional probabilities pertaining to one experiment can be expressed using only conditional probabilities identified during the observation of other experiments? This is called the *transportability* problem.

Pearl's *do*-calculus completely solves the identifiability problem and provides useful tools to address many instances of the transportability problem (see Pearl, 2012). Assuming that we *know* the conditional probability distributions involving observed variables in the original structural equation model, *do*-calculus allows us to *derive* conditional distributions pertaining to the manipulated structural equation model.

Unfortunately, we must further distinguish the conditional probabilities that we know (because we designed them) from those that we estimate from empirical data. This distinction is important because estimating the distribution of continuous or high cardinality variables is notoriously difficult. Furthermore, *do*-calculus often combines the estimated probabilities in ways that amplify estimation errors. This happens when the manipulated structural equation model exercises the variables in ways that were rarely observed in the data collected from the original structural equation model.

4. Bayesian networks are directed graphs representing the Markov factorization of a joint probability distribution: the arrows no longer have a causal interpretation.

Therefore we prefer to use much simpler causal inference techniques (see sections 4.1 and 4.2). Although these techniques do not have the completeness properties of *do*-calculus, they combine estimation and transportation in a manner that facilitates the derivation of useful confidence intervals.

3.5 Special Cases

Three special cases of causal models are particularly relevant to this work.

- In the multi-armed bandit (Robbins, 1952), a user-defined policy function π determines the distribution of action $a \in \{1 \dots K\}$, and an unknown reward function r determines the distribution of the outcome y given the action a (Figure 7). In order to maximize the accumulated rewards, the player must construct policies π that balance the exploration of the action space with the exploitation of the best action identified so far (Auer et al., 2002; Audibert et al., 2007; Seldin et al., 2012).

- The contextual bandit problem (Langford and Zhang, 2008) significantly increases the complexity of multi-armed bandits by adding one exogenous variable x to the policy function π and the reward functions r (Figure 8).

- Both multi-armed bandit and contextual bandit are special case of reinforcement learning (Sutton and Barto, 1998). In essence, a Markov decision process is a sequence of contextual bandits where the context is no longer an exogenous variable but a state variable that depends on the previous states and actions (Figure 9). Note that the policy function π, the reward function r, and the transition function s are independent of time. All the time dependencies are expressed using the states s_t.

These special cases have increasing generality. Many simple structural equation models can be reduced to a contextual bandit problem using appropriate definitions of the context x, the action a and the outcome y. For instance, assuming that the prices c are discrete, the ad placement structural equation model shown in Figure 2 reduces to a contextual bandit problem with context (u, v), actions (s, c) and reward z. Similarly, given a sufficiently intricate definition of the state variables s_t, all structural equation models with discrete variables can be reduced to a reinforcement learning problem. Such reductions lose the fine structure of the causal graph. We show in Section 5 how this fine structure can in fact be leveraged to obtain more information from the same experiments.

Modern reinforcement learning algorithms (see Sutton and Barto, 1998) leverage the assumption that the policy function, the reward function, the transition function, and the distributions of the corresponding noise variables, are independent from time. This invariance property provides great benefits when the observed sequences of actions and rewards are long in comparison with the size of the state space. Only Section 7 in this contribution presents methods that take advantage of such an invariance. The general question of leveraging arbitrary functional invariances in causal graphs is left for future work.

4. Counterfactual Analysis

We now return to the problem of formulating and answering questions about the value of proposed changes of a learning system. Assume for instance that we consider replacing the score computation

$$\begin{array}{rcll} a &=& \pi(\varepsilon) & \text{Action } a \in \{1\ldots K\} \\ y &=& r(a,\, \varepsilon') & \text{Reward } y \in \mathbb{R} \end{array}$$

Figure 7: Structural equation model for the multi-armed bandit problem. The policy π selects a discrete action a, and the reward function r determines the outcome y. The noise variables ε and ε' represent independent sources of randomness useful to model probabilistic dependencies.

$$\begin{array}{rcll} a &=& \pi(x,\, \varepsilon) & \text{Action } a \in \{1\ldots K\} \\ y &=& r(x,\, a,\, \varepsilon') & \text{Reward } y \in \mathbb{R} \end{array}$$

Figure 8: Structural equation model for contextual bandit problem. Both the action and the reward depend on an exogenous context variable x.

$$\begin{array}{rcll} a_t &=& \pi(s_{t-1},\, \varepsilon_t) & \text{Action} \\ y_t &=& r(s_{t-1},\, a_t,\, \varepsilon_t') & \text{Reward } r_t \in \mathbb{R} \\ s_t &=& s(s_{t-1},\, a_t,\, \varepsilon_t'') & \text{Next state} \end{array}$$

Figure 9: Structural equation model for reinforcement learning. The above equations are replicated for all $t \in \{0\ldots,T\}$. The context is now provided by a state variable s_{t-1} that depends on the previous states and actions.

model M of an ad placement engine by an alternate model M^*. We seek an answer to the conditional question:

> "*How will the system perform if we replace model M by model M^* ?*"

Given sufficient time and sufficient resources, we can obtain the answer using a controlled experiment (Section 2.2). However, instead of carrying out a new experiment, we would like to obtain an answer using data that we have already collected in the past.

> "*How would the system have performed if, when the data was collected, we had replaced model M by model M^*?*"

The answer of this *counterfactual question* is of course a *counterfactual statement* that describes the system performance subject to a condition that did not happen.

Counterfactual statements challenge ordinary logic because they depend on a condition that is known to be false. Although assertion $A \Rightarrow B$ is always true when assertion A is false, we certainly do not mean for all counterfactual statements to be true. Lewis (1973) navigates this paradox using a modal logic in which a counterfactual statement describes the state of affairs in an alternate world that resembles ours except for the specified differences. Counterfactuals indeed offer many subtle ways to qualify such alternate worlds. For instance, we can easily describe isolation assumptions (Section 3.2) in a counterfactual question:

> "*How would the system have performed if, when the data was collected, we had replaced model M by model M^* without incurring user or advertiser reactions?*"

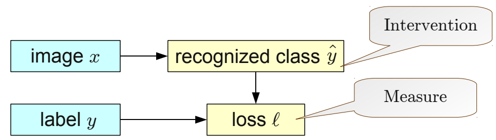

Figure 10: Causal graph for an image recognition system. We can estimate counterfactuals by replaying data collected in the past.

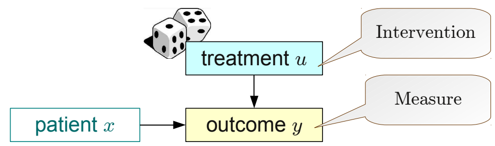

Figure 11: Causal graph for a randomized experiment. We can estimate certain counterfactuals by reweighting data collected in the past.

The fact that we could not have changed the model without incurring the user and advertiser reactions does not matter any more than the fact that we did not replace model M by model M^* in the first place. This does not prevent us from using counterfactual statements to reason about causes and effects. Counterfactual questions and statements provide a natural framework to express and share our conclusions.

The remaining text in this section explains how we can answer certain counterfactual questions using data collected in the past. More precisely, we seek to estimate performance metrics that can be expressed as expectations with respect to the distribution that would have been observed if the counterfactual conditions had been in force.[5]

4.1 Replaying Empirical Data

Figure 10 shows the causal graph associated with a simple image recognition system. The classifier takes an image x and produces a prospective class label $\hat{y}$. The loss measures the penalty associated with recognizing class $\hat{y}$ while the true class is y.

To estimate the expected error of such a classifier, we collect a representative data set composed of labelled images, run the classifier on each image, and average the resulting losses. In other words, we *replay* the data set to estimate what (counterfactual) performance would have been observed if we had used a different classifier. We can then select in retrospect the classifier that would have worked the best and hope that it will keep working well. This is the counterfactual viewpoint on empirical risk minimization (Vapnik, 1982).

Replaying the data set works because both the alternate classifier and the loss function are known. More generally, to estimate a counterfactual by replaying a data set, we need to know all the functional dependencies associated with all causal paths connecting the intervention point to the measurement point. This is obviously not always the case.

5. Although counterfactual expectations can be viewed as expectations of unit-level counterfactuals (Pearl, 2009, Definition 4), they elude the semantic subtleties of unit-level counterfactuals and can be measured with randomized experiments (see Section 4.2.)

4.2 Reweighting Randomized Trials

Figure 11 illustrates the randomized experiment suggested in Section 2.3. The patients are randomly split into two equally sized groups receiving respectively treatments A and B. The overall success rate for this experiment is therefore $Y = (Y_A + Y_B)/2$ where Y_A and Y_B are the success rates observed for each group. We would like to estimate which (counterfactual) overall success rate Y^* would have been observed if we had selected treatment A with probability p and treatment B with probability $1-p$.

Since we do not know how the outcome depends on the treatment and the patient condition, we cannot compute which outcome y^* would have been obtained if we had treated patient x with a different treatment u^*. Therefore we cannot answer this question by replaying the data as we did in Section 4.1.

However, observing different success rates Y_A and Y_B for the treatment groups reveals an empirical correlation between the treatment u and the outcome y. Since the only cause of the treatment u is an independent roll of the dices, this correlation cannot result from any known or unknown confounding common cause.[6] Having eliminated this possibility, we can *reweight* the observed outcomes and compute the estimate $Y^* \approx p\,Y_A + (1-p)\,Y_B$.

4.3 Markov Factor Replacement

The reweighting approach can in fact be applied under much less stringent conditions. Let us return to the ad placement problem to illustrate this point.

The average number of ad clicks per page is often called *click yield*. Increasing the click yield usually benefits both the advertiser and the publisher, whereas increasing the revenue per page often benefits the publisher at the expense of the advertiser. Click yield is therefore a very useful metric when we reason with an isolation assumption that ignores the advertiser reactions to pricing changes.

Let ω be a shorthand for all variables appearing in the Markov factorization of the ad placement structural equation model,

$$\begin{aligned} \mathrm{P}(\omega) \quad = \quad & \mathrm{P}(u,v)\,\mathrm{P}(x\,|\,u)\,\mathrm{P}(a\,|\,x,v)\,\mathrm{P}(b\,|\,x,v)\,\mathrm{P}(q\,|\,x,a) \\ & \times\, \mathrm{P}(s\,|\,a,q,b)\,\mathrm{P}(c\,|\,a,q,b)\,\mathrm{P}(y\,|\,s,u)\,\mathrm{P}(z\,|\,y,c)\;. \end{aligned} \tag{2}$$

Variable y was defined in Section 3.1 as the set of user clicks. In the rest of the document, we slightly abuse this notation by using the same letter y to represent the number of clicks. We also write the expectation $Y = \mathbb{E}_{\omega\sim\mathrm{P}(\omega)}[y]$ using the integral notation

$$Y = \int_\omega y\,\mathrm{P}(\omega)\;.$$

We would like to estimate what the expected click yield Y^* would have been if we had used a different scoring function (Figure 12). This intervention amounts to replacing the actual factor $\mathrm{P}(\,q\,|\,x,a)$ by a counterfactual factor $\mathrm{P}^*(\,q\,|\,x,a)$ in the Markov factorization.

$$\begin{aligned} \mathrm{P}^*(\omega) \quad = \quad & \mathrm{P}(u,v)\,\mathrm{P}(x\,|\,u)\,\mathrm{P}(a\,|\,x,v)\,\mathrm{P}(b\,|\,x,v)\,\mathrm{P}^*(q\,|\,x,a) \\ & \times\, \mathrm{P}(s\,|\,a,q,b)\,\mathrm{P}(c\,|\,a,q,b)\,\mathrm{P}(y\,|\,s,u)\,\mathrm{P}(z\,|\,x,c)\;. \end{aligned} \tag{3}$$

6. See also the discussion of Reichenbach's common cause principle and of its limitations in Spirtes et al. (1993) and Spirtes and Scheines (2004).

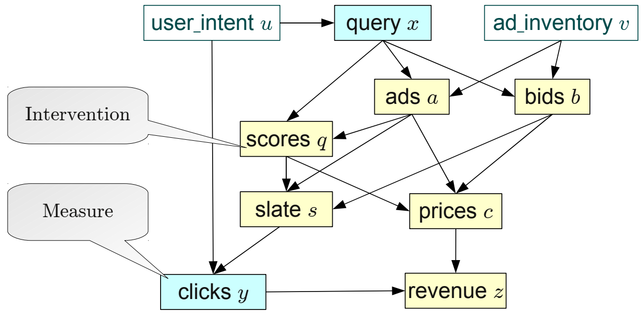

Figure 12: Estimating which average number of clicks per page would have been observed if we had used a different scoring model.

Let us assume, for simplicity, that the actual factor $\mathrm{P}(\,q\,|\,x,a)$ is nonzero everywhere. We can then estimate the counterfactual expected click yield Y^* using the transformation

$$Y^* \;=\; \int_\omega y\; \mathrm{P}^*(\omega) \;=\; \int_\omega y\; \frac{\mathrm{P}^*(\,q\,|\,x,a)}{\mathrm{P}(\,q\,|\,x,a)}\; \mathrm{P}(\omega) \;\approx\; \frac{1}{n}\sum_{i=1}^{n} y_i\; \frac{\mathrm{P}^*(\,q_i\,|\,x_i,a_i)}{\mathrm{P}(\,q_i\,|\,x_i,a_i)}\;, \tag{4}$$

where the data set of tuples (a_i,x_i,q_i,y_i) is distributed according to the actual Markov factorization instead of the counterfactual Markov factorization. This data could therefore have been collected during the normal operation of the ad placement system. Each sample is reweighted to reflect its probability of occurrence under the counterfactual conditions.

In general, we can use *importance sampling* to estimate the counterfactual expectation of any quantity $\ell(\omega)$:

$$Y^* \;=\; \int_\omega \ell(\omega)\; \mathrm{P}^*(\omega) \;=\; \int_\omega \ell(\omega)\; \frac{\mathrm{P}^*(\omega)}{\mathrm{P}(\omega)}\; \mathrm{P}(\omega) \;\approx\; \frac{1}{n}\sum_{i=1}^{n} \ell(\omega_i)\, w_i \tag{5}$$

with weights

$$w_i \;=\; w(\omega_i) \;=\; \frac{\mathrm{P}^*(\omega_i)}{\mathrm{P}(\omega_i)} \;=\; \frac{\text{factors appearing in } \mathrm{P}^*(\omega_i) \text{ but not in } \mathrm{P}(\omega_i)}{\text{factors appearing in } \mathrm{P}(\omega_i) \text{ but not in } \mathrm{P}^*(\omega_i)}\;. \tag{6}$$

Equation (6) emphasizes the simplifications resulting from the algebraic similarities of the actual and counterfactual Markov factorizations. Because of these simplifications, the evaluation of the weights only requires the knowledge of the few factors that differ between $\mathrm{P}(\omega)$ and $\mathrm{P}^*(\omega)$. Each data sample needs to provide the value of $\ell(\omega_i)$ and the values of all variables needed to evaluate the factors that do not cancel in the ratio (6).

In contrast, the replaying approach (Section 4.1) demands the knowledge of all factors of $\mathrm{P}^*(\omega)$ connecting the point of intervention to the point of measurement $\ell(\omega)$. On the other hand, it does not require the knowledge of factors appearing only in $\mathrm{P}(\omega)$.

Importance sampling relies on the assumption that all the factors appearing in the denominator of the reweighting ratio (6) are nonzero whenever the factors appearing in the numerator are nonzero. Since these factors represents conditional probabilities resulting from the effect of an independent noise variable in the structural equation model, this assumption means that the data must be collected with an experiment involving active randomization. We must therefore design cost-effective randomized experiments that yield enough information to estimate many interesting counterfactual expectations with sufficient accuracy. This problem cannot be solved without answering the confidence interval question: given data collected with a certain level of randomization, with which accuracy can we estimate a given counterfactual expectation?

4.4 Confidence Intervals

At first sight, we can invoke the law of large numbers and write

$$Y^* = \int_\omega \ell(\omega)\, w(\omega)\, \mathrm{P}(\omega) \;\approx\; \frac{1}{n}\sum_{i=1}^{n} \ell(\omega_i)\, w_i\,. \tag{7}$$

For sufficiently large n, the central limit theorem provides confidence intervals whose width grows with the standard deviation of the product $\ell(\omega)\, w(\omega)$.

Unfortunately, when $\mathrm{P}(\omega)$ is small, the reweighting ratio $w(\omega)$ takes large values with low probability. This heavy tailed distribution has annoying consequences because the variance of the integrand could be very high or infinite. When the variance is infinite, the central limit theorem does not hold. When the variance is merely very large, the central limit convergence might occur too slowly to justify such confidence intervals. Importance sampling works best when the actual distribution and the counterfactual distribution overlap.

When the counterfactual distribution has significant mass in domains where the actual distribution is small, the few samples available in these domains receive very high weights. Their noisy contribution dominates the reweighted estimate (7). We can obtain better confidence intervals by eliminating these few samples drawn in poorly explored domains. The resulting bias can be bounded using prior knowledge, for instance with an assumption about the range of values taken by $\ell(\omega)$,

$$\forall \omega \quad \ell(\omega) \,\in\, [0, M]\,. \tag{8}$$

Let us choose the maximum weight value R deemed acceptable for the weights. We have obtained very consistent results in practice with R equal to the fifth largest reweighting ratio observed on the empirical data.[7] We can then rely on *clipped weights* to eliminate the contribution of the poorly explored domains,

$$\bar{w}(\omega) \;=\; \begin{cases} w(\omega) & \text{if } \mathrm{P}^*(\omega) < R\,\mathrm{P}(\omega) \\ 0 & \text{otherwise.} \end{cases}$$

The condition $\mathrm{P}^*(\omega) < R\,\mathrm{P}(\omega)$ ensures that the ratio has a nonzero denominator $\mathrm{P}(\omega)$ and is smaller than R. Let Ω_R be the set of all values of ω associated with acceptable ratios:

$$\Omega_R \;=\; \{\, \omega :\; \mathrm{P}^*(\omega) < R\,\mathrm{P}(\omega) \,\}\,.$$

We can decompose Y^* in two terms:

$$Y^* \;=\; \int_{\omega\in\Omega_R} \ell(\omega)\,\mathrm{P}^*(\omega) \;+\; \int_{\omega\in\Omega\setminus\Omega_R} \ell(\omega)\,\mathrm{P}^*(\omega) \;=\; \bar{Y}^* + (Y^* - \bar{Y}^*)\,. \tag{9}$$

7. This is in fact a slight abuse because the theory calls for choosing R before seeing the data.

The first term of this decomposition is the *clipped expectation* $\bar{Y}^*$. Estimating the clipped expectation $\bar{Y}^*$ is much easier than estimating Y^* from (7) because the clipped weights $\bar{w}(\omega)$ are bounded by R.

$$\bar{Y}^* \;=\; \int_{\omega\in\Omega_R} \ell(\omega)\,\mathrm{P}^*(\omega) \;=\; \int_{\omega} \ell(\omega)\,\bar{w}(\omega)\,\mathrm{P}(\omega) \;\approx\; \widehat{Y}^* \;=\; \frac{1}{n}\sum_{i=1}^{n} \ell(\omega_i)\,\bar{w}(\omega_i)\,. \tag{10}$$

The second term of Equation (9) can be bounded by leveraging assumption (8). The resulting bound can then be conveniently estimated using only the clipped weights.

$$Y^* - \bar{Y}^* \;=\; \int_{\omega\in\Omega\setminus\Omega_R} \ell(\omega)\,\mathrm{P}^*(\omega) \;\in\; \Big[0,\; M\,\mathrm{P}^*(\Omega\setminus\Omega_R)\Big] \;=\; \Big[0,\; M\,(1-\bar{W}^*)\Big] \quad \text{with}$$

$$\bar{W}^* \;=\; \mathrm{P}^*(\Omega_R) \;=\; \int_{\omega\in\Omega_R} \mathrm{P}^*(\omega) \;=\; \int_{\omega} \bar{w}(\omega)\,\mathrm{P}(\omega) \;\approx\; \widehat{W}^* \;=\; \frac{1}{n}\sum_{i=1}^{n} \bar{w}(\omega_i)\,. \tag{11}$$

Since the clipped weights are bounded, the estimation errors associated with (10) and (11) are well characterized using either the central limit theorem or using empirical Bernstein bounds (see appendix B for details). Therefore we can derive an *outer confidence interval* of the form

$$\mathbb{P}\Big\{\; \widehat{Y}^* - \varepsilon_R \;\le\; \bar{Y}^* \;\le\; \widehat{Y}^* + \varepsilon_R \;\Big\} \;\ge\; 1-\delta \tag{12}$$

and an *inner confidence interval* of the form

$$\mathbb{P}\Big\{\; \bar{Y}^* \;\le\; Y^* \;\le\; \bar{Y}^* + M(1-\widehat{W}^* + \xi_R) \;\Big\} \;\ge\; 1-\delta\,. \tag{13}$$

The names *inner* and *outer* are in fact related to our preferred way to visualize these intervals (e.g., Figure 13). Since the bounds on $Y^* - \bar{Y}^*$ can be written as

$$\bar{Y}^* \;\le\; Y^* \;\le\; \bar{Y}^* + M\,(1-\bar{W}^*)\,, \tag{14}$$

we can derive our final confidence interval,

$$\mathbb{P}\Big\{\; \widehat{Y}^* - \varepsilon_R \;\le\; Y^* \;\le\; \widehat{Y}^* + M(1-\widehat{W}^* + \xi_R) + \varepsilon_R \;\Big\} \ge 1-2\delta\,. \tag{15}$$

In conclusion, replacing the unbiased importance sampling estimator (7) by the clipped importance sampling estimator (10) with a suitable choice of R leads to improved confidence intervals. Furthermore, since the derivation of these confidence intervals does not rely on the assumption that $\mathrm{P}(\omega)$ is nonzero everywhere, the clipped importance sampling estimator remains valid when the distribution $\mathrm{P}(\omega)$ has a limited support. This relaxes the main restriction associated with importance sampling.

4.5 Interpreting the Confidence Intervals

The estimation of the counterfactual expectation Y^* can be inaccurate because the sample size is insufficient or because the sampling distribution $\mathrm{P}(\omega)$ does not sufficiently explore the counterfactual conditions of interest.

By construction, the clipped expectation $\bar{Y}^*$ ignores the domains poorly explored by the sampling distribution $\mathrm{P}(\omega)$. The difference $Y^* - \bar{Y}^*$ then reflects the inaccuracy resulting from a lack of exploration. Therefore, assuming that the bound R has been chosen competently, the relative sizes of the outer and inner confidence intervals provide precious cues to determine whether we can continue collecting data using the same experimental setup or should adjust the data collection experiment in order to obtain a better coverage.

- The *inner confidence interval* (13) witnesses the uncertainty associated with the domain G_R insufficiently explored by the actual distribution. A large inner confidence interval suggests that the most practical way to improve the estimate is to adjust the data collection experiment in order to obtain a better coverage of the counterfactual conditions of interest.

- The *outer confidence interval* (12) represents the uncertainty that results from the limited sample size. A large outer confidence interval indicates that the sample is too small. To improve the result, we simply need to continue collecting data using the same experimental setup.

4.6 Experimenting with Mainline Reserves

We return to the ad placement problem to illustrate the reweighting approach and the interpretation of the confidence intervals. Manipulating the reserves $R_p(x)$ associated with the mainline positions (Figure 1) controls which ads are prominently displayed in the mainline or displaced into the sidebar.

We seek in this section to answer counterfactual questions of the form:

> "*How would the ad placement system have performed if we had scaled the mainline reserves by a constant factor* ρ*, without incurring user or advertiser reactions?*"

Randomization was introduced using a modified version of the ad placement engine. Before determining the ad layout (see Section 2.1), a random number ε is drawn according to the standard normal distribution $\mathcal{N}(0,1)$, and all the mainline reserves are multiplied by $m = \rho\, e^{-\sigma^2/2+\sigma\varepsilon}$. Such multipliers follow a log-normal distribution[8] whose mean is ρ and whose width is controlled by σ. This effectively provides a parametrization of the conditional score distribution $\mathrm{P}(\,q\,|\,x,a)$ (see Figure 5.)

The Bing search platform offers many ways to select traffic for controlled experiments (Section 2.2). In order to match our isolation assumption, individual page views were randomly assigned to traffic buckets without regard to the user identity. The main treatment bucket was processed with mainline reserves randomized by a multiplier drawn as explained above with $\rho = 1$ and $\sigma = 0.3$. With these parameters, the mean multiplier is exactly 1, and 95% of the multipliers are in range $[0.52, 1.74]$. Samples describing 22 million search result pages were collected during five consecutive weeks.

We then use this data to estimate what would have been measured if the mainline reserve multipliers had been drawn according to a distribution determined by parameters ρ^* and σ^*. This is achieved by reweighting each sample ω_i with

$$w_i = \frac{\mathrm{P}^*(\,q_i\,|\,x_i,a_i)}{\mathrm{P}(\,q_i\,|\,x_i,a_i)} = \frac{p(m_i\,;\,\rho^*,\,\sigma^*)}{p(m_i\,;\,\rho,\,\sigma)}\,,$$

where m_i is the multiplier drawn for this sample during the data collection experiment, and $p(t\,;\,\rho,\sigma)$ is the density of the log-normal multiplier distribution.

Figure 13 reports results obtained by varying ρ^* while keeping $\sigma^* = \sigma$. This amounts to estimating what would have been measured if all mainline reserves had been multiplied by ρ^* while keeping the same randomization. The curves bound 95% confidence intervals on the variations of the average number of mainline ads displayed per page, the average number of ad clicks per page,

8. More precisely, $ln\mathcal{N}(\mu,\sigma^2)$ with $\mu = \sigma^2/2 + \log\rho$.

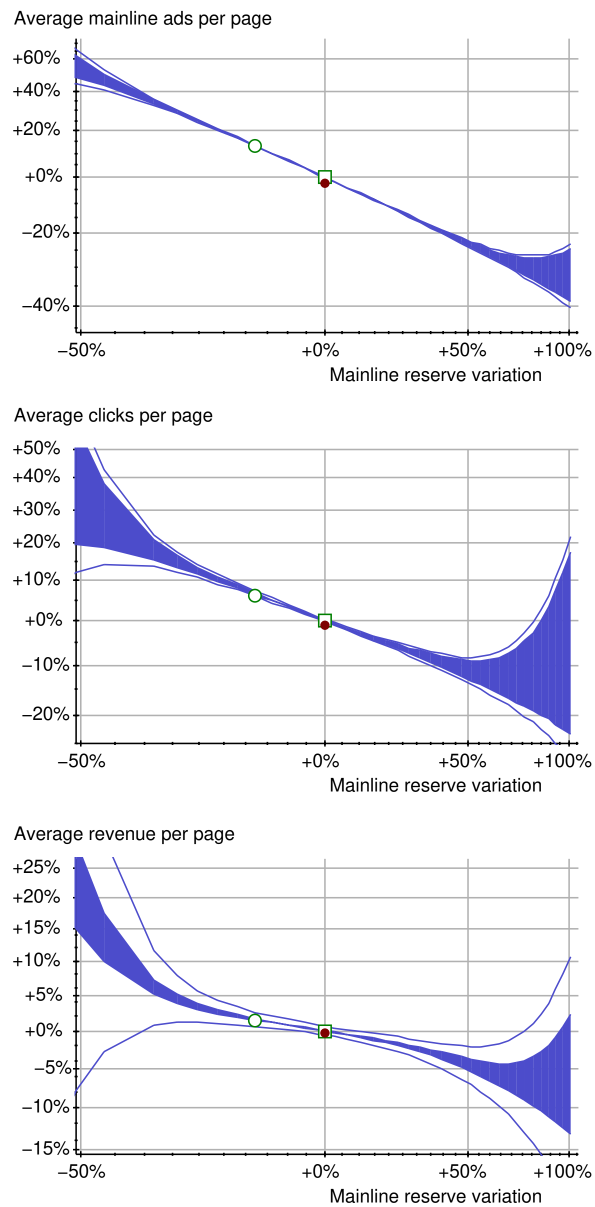

Figure 13: Estimated variations of three performance metrics in response to mainline reserve changes. The curves delimit 95% confidence intervals for the metrics we would have observed if we had increased the mainline reserves by the percentage shown on the horizontal axis. The filled areas represent the inner confidence intervals. The hollow squares represent the metrics measured on the experimental data. The hollow circles represent metrics measured on a second experimental bucket with mainline reserves reduced by 18%. The filled circles represent the metrics effectively measured on a control bucket running without randomization.

and the average revenue per page, as functions of ρ^*. The inner confidence intervals, represented by the filled areas, grow sharply when ρ^* leaves the range explored during the data collection experiment. The average revenue per page has more variance because a few very competitive queries command high prices.

In order to validate the accuracy of these counterfactual estimates, a second traffic bucket of equal size was configured with mainline reserves reduced by about 18%. The hollow circles in Figure 13 represent the metrics effectively measured on this bucket during the same time period. The effective measurements and the counterfactual estimates match with high accuracy.

Finally, in order to measure the cost of the randomization, we also ran the unmodified ad placement system on a control bucket. The brown filled circles in Figure 13 represent the metrics effectively measured on the control bucket during the same time period. The randomization caused a small but statistically significant increase of the number of mainline ads per page. The click yield and average revenue differences are not significant.

This experiment shows that we can obtain accurate counterfactual estimates with affordable randomization strategies. However, this nice conclusion does not capture the true practical value of the counterfactual estimation approach.

4.7 More on Mainline Reserves

The main benefit of the counterfactual estimation approach is the ability to *use the same data* to answer a *broad range of counterfactual questions*. Here are a few examples of counterfactual questions that can be answered using data collected using the simple mainline reserve randomization scheme described in the previous section:

- *Different variances* – Instead of estimating what would have been measured if we had increased the mainline reserves without changing the randomization variance, that is, letting $\sigma^* = \sigma$, we can use the same data to estimate what would have been measured if we had also changed σ. This provides the means to determine which level of randomization we can afford in future experiments.

- *Pointwise estimates* – We often want to estimate what would have been measured if we had set the mainline reserves to a specific value without randomization. Although computing estimates for small values of σ often works well enough, very small values lead to large confidence intervals.

  Let $Y_\nu(\rho)$ represent the expectation we would have observed if the multipliers m had mean ρ and variance ν. We have then $Y_\nu(\rho) = \mathbb{E}_m[\mathbb{E}[y|m]] = \mathbb{E}_m[Y_0(m)]$. Assuming that the pointwise value Y_0 is smooth enough for a second order development,

  $$Y_\nu(\rho) \;\approx\; \mathbb{E}_m\big[\,Y_0(\rho) + (m-\rho)Y_0'(\rho) + (m-\rho)^2 Y_0''(\rho)/2\,\big] \;=\; Y_0(\rho) + \nu Y_0''(\rho)/2\;.$$

  Although the reweighting method cannot estimate the point-wise value $Y_0(\rho)$ directly, we can use the reweighting method to estimate both $Y_\nu(\rho)$ and $Y_{2\nu}(\rho)$ with acceptable confidence intervals and write $Y_0(\rho) \approx 2Y_\nu(\rho) - Y_{2\nu}(\rho)$ (Goodwin, 2011).

- *Query-dependent reserves* – Compare for instance the queries "car insurance" and "common cause principle" in a web search engine. Since the advertising potential of a search varies considerably with the query, it makes sense to investigate various ways to define query-dependent reserves (Charles and Chickering, 2012).

The data collected using the simple mainline reserve randomization can also be used to estimate what would have been measured if we had increased all the mainline reserves by a query-dependent multiplier $\rho^*(x)$. This is simply achieved by reweighting each sample ω_i with

$$w_i = \frac{\mathrm{P}^*(\,q_i\,|\,x_i, a_i)}{\mathrm{P}(\,q_i\,|\,x_i, a_i)} = \frac{p(m_i\,;\,\rho^*(x_i)\,,\,\sigma)}{p(m_i\,;\,\mu,\,\sigma)}\,.$$

Considerably broader ranges of counterfactual questions can be answered when data is collected using randomization schemes that explore more dimensions. For instance, in the case of the ad placement problem, we could apply an independent random multiplier for each score instead of applying a single random multiplier to the mainline reserves only. However, the more dimensions we randomize, the more data needs to be collected to effectively explore all these dimensions. Fortunately, as discussed in section 5, the structure of the causal graph reveals many ways to leverage a priori information and improve the confidence intervals.

4.8 Related Work

Importance sampling is widely used to deal with covariate shifts (Shimodaira, 2000; Sugiyama et al., 2007). Since manipulating the causal graph changes the data distribution, such an intervention can be viewed as a covariate shift amenable to importance sampling. Importance sampling techniques have also been proposed without causal interpretation for many of the problems that we view as causal inference problems. In particular, the work presented in this section is closely related to the Monte-Carlo approach of reinforcement learning (Sutton and Barto, 1998, Chapter 5) and to the offline evaluation of contextual bandit policies (Li et al., 2010, 2011).

Reinforcement learning research traditionally focuses on control problems with relatively small discrete state spaces and long sequences of observations. This focus reduces the need for characterizing exploration with tight confidence intervals. For instance, Sutton and Barto suggest to normalize the importance sampling estimator by $1/\sum_i w(\omega_i)$ instead of $1/n$. This would give erroneous results when the data collection distribution leaves parts of the state space poorly explored. Contextual bandits are traditionally formulated with a finite set of discrete actions. For instance, Li's (2011) unbiased policy evaluation assumes that the data collection policy always selects an arbitrary policy with probability greater than some small constant. This is not possible when the action space is infinite.

Such assumptions on the data collection distribution are often impractical. For instance, certain ad placement policies are not worth exploring because they cannot be implemented efficiently or are known to elicit fraudulent behaviors. There are many practical situations in which one is only interested in limited aspects of the ad placement policy involving continuous parameters such as click prices or reserves. Discretizing such parameters eliminates useful a priori knowledge: for instance, if we slightly increase a reserve, we can reasonable believe that we are going to show slightly less ads.

Instead of making assumptions on the data collection distribution, we construct a biased estimator (10) and bound its bias. We then interpret the inner and outer confidence intervals as resulting from a lack of exploration or an insufficient sample size.

Finally, the causal framework allows us to easily formulate counterfactual questions that pertain to the practical ad placement problem and yet differ considerably in complexity and exploration requirements. We can address specific problems identified by the engineers without incurring the risks associated with a complete redesign of the system. Each of these incremental steps helps demonstrating the soundness of the approach.

5. Structure

This section shows how the structure of the causal graph reveals many ways to leverage a priori knowledge and improve the accuracy of our counterfactual estimates. Displacing the reweighting point (Section 5.1) improves the inner confidence interval and therefore reduce the need for exploration. Using a prediction function (Section 5.2) essentially improve the outer confidence interval and therefore reduce the sample size requirements.

5.1 Better Reweighting Variables

Many search result pages come without eligible ads. We then know with certainty that such pages will have zero mainline ads, receive zero clicks, and generate zero revenue. This is true for the randomly selected value of the reserve, and this would have been true for any other value of the reserve. We can exploit this knowledge by pretending that the reserve was drawn from the counterfactual distribution $\mathrm{P}^*(q\,|\,x_i,a_i)$ instead of the actual distribution $\mathrm{P}(q\,|\,x_i,a_i)$. The ratio $w(\omega_i)$ is therefore forced to the unity. This does not change the estimate but reduces the size of the inner confidence interval. The results of Figure 13 were in fact helped by this little optimization.

There are in fact many circumstances in which the observed outcome would have been the same for other values of the randomized variables. This prior knowledge is in fact encoded in the structure of the causal graph and can be exploited in a more systematic manner. For instance, we know that users make click decisions without knowing which scores were computed by the ad placement engine, and without knowing the prices charged to advertisers. The ad placement causal graph encodes this knowledge by showing the clicks y as direct effects of the user intent u and the ad slate s. This implies that the exact value of the scores q does not matter to the clicks y as long as the ad slate s remains the same.

Because the causal graph has this special structure, we can simplify both the actual and counterfactual Markov factorizations (2) (3) without eliminating the variable y whose expectation is sought. Successively eliminating variables z, c, and q gives:

$$\begin{aligned}
\mathrm{P}(u,v,x,a,b,s,y) &= \mathrm{P}(u,v)\,\mathrm{P}(x\,|\,u)\,\mathrm{P}(a\,|\,x,v)\,\mathrm{P}(b\,|\,x,v)\,\mathrm{P}(s\,|\,x,a,b)\,\mathrm{P}(y\,|\,s,u)\ , \\
\mathrm{P}^*(u,v,x,a,b,s,y) &= \mathrm{P}(u,v)\,\mathrm{P}(x\,|\,u)\,\mathrm{P}(a\,|\,x,v)\,\mathrm{P}(b\,|\,x,v)\,\mathrm{P}^*(s\,|\,x,a,b)\,\mathrm{P}(y\,|\,s,u)\ .
\end{aligned}$$

The conditional distributions $\mathrm{P}(s\,|\,x,a,b)$ and $\mathrm{P}^*(s\,|\,x,a,b)$ did not originally appear in the Markov factorization. They are defined by marginalization as a consequence of the elimination of the variable q representing the scores.

$$\mathrm{P}(s\,|\,x,a,b) = \int_q \mathrm{P}(s\,|\,a,q,b)\,\mathrm{P}(q\,|\,x,a)\ , \quad \mathrm{P}^*(s\,|\,x,a,b) = \int_q \mathrm{P}(s\,|\,a,q,b)\,\mathrm{P}^*(q\,|\,x,a)\,.$$

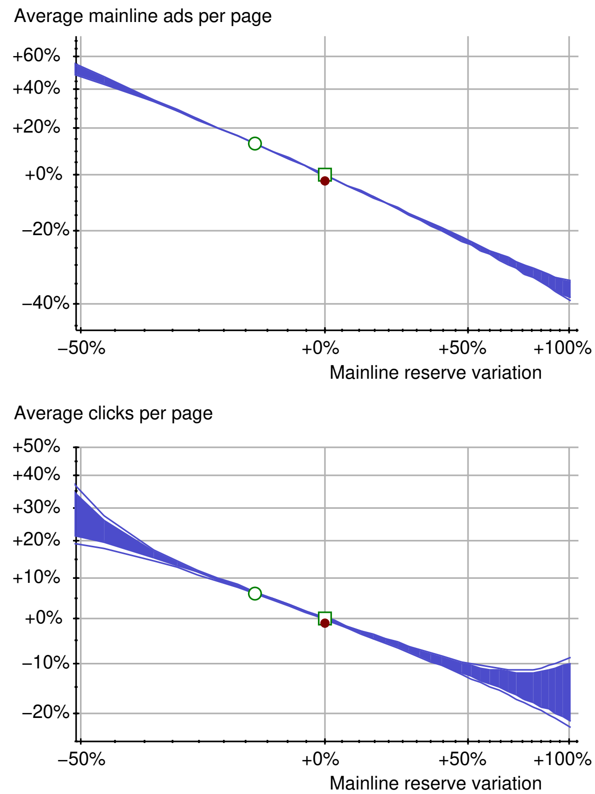

Figure 14: Estimated variations of two performance metrics in response to mainline reserve changes. These estimates were obtained using the ad slates s as reweighting variable. Compare the inner confidence intervals with those shown in Figure 13.

We can estimate the counterfactual click yield Y^* using these simplified factorizations:

$$
\begin{aligned}
Y^* &= \int y\, \mathrm{P}^*(u,v,x,a,b,s,y) = \int y\, \frac{\mathrm{P}^*(s\,|\,x,a,b)}{\mathrm{P}(s\,|\,x,a,b)}\, \mathrm{P}(u,v,x,a,b,s,y) \\
&\approx \frac{1}{n}\sum_{i=1}^{n} y_i\, \frac{\mathrm{P}^*(s_i\,|\,x_i,a_i,b_i)}{\mathrm{P}(s_i\,|\,x_i,a_i,b_i)}\,. \qquad (16)
\end{aligned}
$$

We have reproduced the experiments described in Section 4.6 with the counterfactual estimate (16) instead of (4). For each example ω_i, we determine which range $[m_i^{\max}, m_i^{\min}]$ of mainline reserve multipliers could have produced the observed ad slate s_i, and then compute the reweighting ratio using the formula:

$$
w_i = \frac{\mathrm{P}^*(s_i\,|\,x_i,a_i,b_i)}{\mathrm{P}(s_i\,|\,x_i,a_i,b_i)} = \frac{\Psi(m_i^{\max};\rho^*,\sigma^*) - \Psi(m_i^{\min};\rho^*,\sigma^*)}{\Psi(m_i^{\max};\rho,\sigma) - \Psi(m_i^{\min};\rho,\sigma)}\ ,
$$

where $\Psi(m;\rho,\sigma)$ is the cumulative of the log-normal multiplier distribution. Figure 14 shows counterfactual estimates obtained using the same data as Figure 13. The obvious improvement of

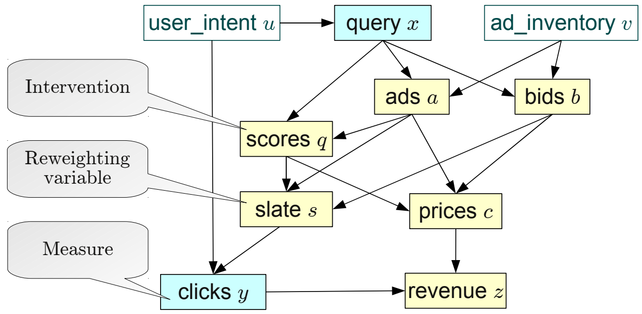

Figure 15: The reweighting variable(s) must intercept all causal paths from the point of intervention to the point of measurement.

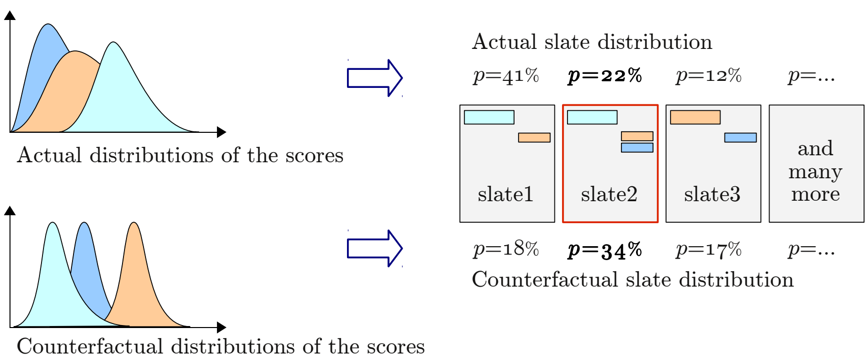

Figure 16: A distribution on the scores q induce a distribution on the possible ad slates s. If the observed slate is `slate2`, the reweighting ratio is $34/22$.

the inner confidence intervals significantly extends the range of mainline reserve multipliers for which we can compute accurate counterfactual expectations using this same data.

Comparing (4) and (16) makes the difference very clear: instead of computing the ratio of the probabilities of the observed scores under the counterfactual and actual distributions, we compute the ratio of the probabilities of the observed ad slates under the counterfactual and actual distributions. As illustrated by Figure 15, we now distinguish the reweighting variable (or variables) from the intervention. In general, the corresponding manipulation of the Markov factorization consists of marginalizing out all the variables that appear on the causal paths connecting the point of intervention to the reweighting variables and factoring all the independent terms out of the integral. This simplification works whenever the reweighting variables intercept all the causal paths connecting the point of intervention to the measurement variable. In order to compute the new reweighting ratios, all the factors remaining inside the integral, that is, all the factors appearing on the causal paths connecting the point of intervention to the reweighting variables, have to be known.

Figure 14 does not report the average revenue per page because the revenue z also depends on the scores q through the click prices c. This causal path is not intercepted by the ad slate variable s alone. However, we can introduce a new variable $\tilde{c} = f(c, y)$ that filters out the click prices computed for ads that did not receive a click. Markedly improved revenue estimates are then obtained by reweighting according to the joint variable $(s, \tilde{c})$.

Figure 16 illustrates the same approach applied to the simultaneous randomization of all the scores q using independent log-normal multipliers. The weight $w(\omega_i)$ is the ratio of the probabilities of the observed ad slate s_i under the counterfactual and actual multiplier distributions. Computing these probabilities amounts to integrating a multivariate Gaussian distribution (Genz, 1992). Details will be provided in a forthcoming publication.

5.2 Variance Reduction with Predictors

Although we do not know exactly how the variable of interest $\ell(\omega)$ depends on the measurable variables and are affected by interventions on the causal graph, we may have strong a priori knowledge about this dependency. For instance, if we augment the slate s with an ad that usually receives a lot of clicks, we can expect an increase of the number of clicks.

Let the *invariant variables* υ be all observed variables that are not direct or indirect effects of variables affected by the intervention under consideration. This definition implies that the distribution of the invariant variables is not affected by the intervention. Therefore the values υ_i of the invariant variables sampled during the actual experiment are also representative of the distribution of the invariant variables under the counterfactual conditions.

We can leverage a priori knowledge to construct a predictor $\zeta(\omega)$ of the quantity $\ell(\omega)$ whose counterfactual expectation Y^* is sought. We assume that the predictor $\zeta(\omega)$ depends only on the invariant variables or on variables that depend on the invariant variables through known functional dependencies. Given sampled values υ_i of the invariant variables, we can replay both the original and manipulated structural equation model as explained in Section 4.1 and obtain samples ζ_i and ζ_i^* that respectively follow the actual and counterfactual distributions

Then, regardless of the quality of the predictor,

$$
\begin{aligned}
Y^* = \int_\omega \ell(\omega)\,\mathrm{P}^*(\omega) \;&=\; \int_\omega \zeta(\omega)\,\mathrm{P}^*(\omega) + \int_\omega (\ell(\omega) - \zeta(\omega))\,\mathrm{P}^*(\omega) \\
&\approx\; \frac{1}{n}\sum_{i=1}^{n} \zeta_i^* + \frac{1}{n}\sum_{i=1}^{n} (\ell(\omega_i) - \zeta_i)\,w(\omega_i)\,.
\end{aligned} \tag{17}
$$

The first term in this sum represents the counterfactual expectation of the predictor and can be accurately estimated by averaging the simulated counterfactual samples ζ_i^* without resorting to potentially large importance weights. The second term in this sum represents the counterfactual expectation of the residuals $\ell(\omega) - \zeta(\omega)$ and must be estimated using importance sampling. Since the magnitude of the residuals is hopefully smaller than that of $\ell(\omega)$, the variance of $(\ell(\omega) - \zeta(\omega))\,w(\omega)$ is reduced and the importance sampling estimator of the second term has improved confidence intervals. The more accurate the predictor $\zeta(\omega)$, the more effective this variance reduction strategy.

This variance reduction technique is in fact identical to the doubly robust contextual bandit evaluation technique of Dudík et al. (2012). Doubly robust variance reduction has also been extensively used for causal inference applied to biostatistics (see Robins et al., 2000; Bang and Robins, 2005). We subjectively find that viewing the predictor as a component of the causal graph (Figure 17) clarifies how a well designed predictor can leverage prior knowledge. For instance, in order to estimate the counterfactual performance of the ad placement system, we can easily use a predictor that runs the ad auction and simulate the user clicks using a click probability model trained offline.

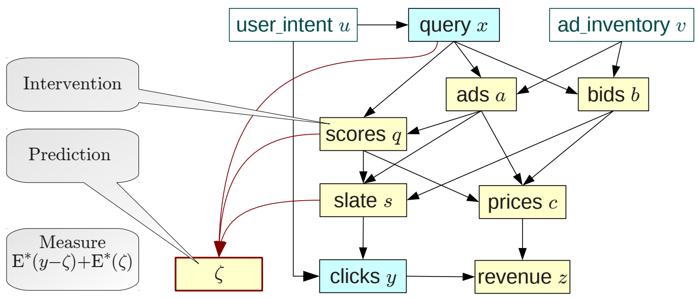

Figure 17: Leveraging a predictor. Yellow nodes represent known functional relations in the structural equation model. We can estimate the counterfactual expectation Y^* of the number of clicks per page as the sum of the counterfactual expectations of a predictor ζ, which is easy to estimate by replaying empirical data, and $y - \zeta$, which has to be estimated by importance sampling but has reduced variance.

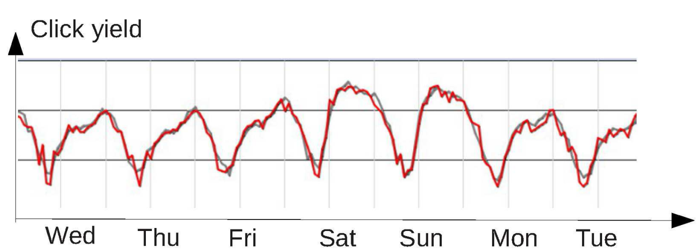

Figure 18: The two plots show the hourly click yield for two variants of the ad placement engine. The daily variations dwarf the differences between the two treatments.

5.3 Invariant Predictors

In order to evaluate which of two interventions is most likely to improve the system, the designer of a learning system often seeks to estimate a *counterfactual difference*, that is, the difference $Y^+ - Y^*$ of the expectations of a same quantity $\ell(\omega)$ under two different counterfactual distributions $\mathrm{P}^+(\omega)$ and $\mathrm{P}^*(\omega)$. These expectations are often affected by variables whose value is left unchanged by the interventions under consideration. For instance, seasonal effects can have very large effects on the number of ad clicks (Figure 18) but affect Y^+ and Y^* in similar ways.

Substantially better confidence intervals on the difference $Y^+ - Y^*$ can be obtained using an *invariant predictor*, that is, a predictor function that depends only on invariant variables υ such as the time of the day. Since the invariant predictor $\zeta(\upsilon)$ is not affected by the interventions under consideration,

$$\int_\omega \zeta(\upsilon)\,\mathrm{P}^*(\omega) = \int_\omega \zeta(\upsilon)\,\mathrm{P}^+(\omega)\,. \tag{18}$$

Therefore

$$\begin{aligned} Y^+ - Y^* \;&=\; \int_\omega \zeta(\upsilon)\,\mathrm{P}^+(\omega) + \int_\omega (\ell(\omega) - \zeta(\upsilon))\,\mathrm{P}^+(\omega) \\ &\qquad - \int_\omega \zeta(\upsilon)\,\mathrm{P}^*(\omega) - \int_\omega (\ell(\omega) - \zeta(\upsilon))\,\mathrm{P}^*(\omega) \\ &\approx\; \frac{1}{n}\sum_{i=1}^{n} \big(\ell(\omega_i) - \zeta(\upsilon_i)\big)\,\frac{P^+(\omega_i) - P^*(\omega_i)}{P(\omega_i)}\,. \end{aligned}$$

This direct estimate of the counterfactual difference $Y^+ - Y^*$ benefits from the same variance reduction effect as (17) without need to estimate the expectations (18). Appendix C provide details on the computation of confidence intervals for estimators of the counterfactual differences. Appendix D shows how the same approach can be used to compute *counterfactual derivatives* that describe the response of the system to very small interventions.

6. Learning

The previous sections deal with the identification and the measurement of interpretable signals that can justify the actions of human decision makers. These same signals can also justify the actions of machine learning algorithms. This section explains why optimizing a counterfactual estimate is a sound learning procedure.

6.1 A Learning Principle

We consider in this section interventions that depend on a parameter θ. For instance, we might want to know what the performance of the ad placement engine would have been if we had used different values for the parameter θ of the click scoring model. Let $\mathrm{P}^\theta(\omega)$ denote the counterfactual Markov factorization associated with this intervention. Let Y^θ be the counterfactual expectation of $\ell(\omega)$ under distribution P^θ. Figure 19 illustrates our simple learning setup. Training data is collected from a single experiment associated with an initial parameter value θ^0 chosen using prior knowledge acquired in an unspecified manner. A preferred parameter value θ^* is then determined using the training data and loaded into the system. The goal is of course to observe a good performance on data collected during a test period that takes place after the switching point.

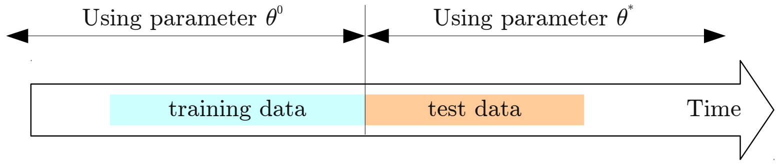

Figure 19: Single design – A preferred parameter value θ^* is determined using randomized data collected in the past. Test data is collected after loading θ^* into the system.

The isolation assumption introduced in Section 3.2 states that the exogenous variables are drawn from an unknown but fixed joint probability distribution. This distribution induces a joint distribution $\mathrm{P}(\omega)$ on all the variables ω appearing in the structural equation model associated with the parameter θ. Therefore, if the *isolation assumption remains valid during the test period*, the test data follows the same distribution $\mathrm{P}^{\theta^*}(\omega)$ that would have been observed during the training data collection period if the system had been using parameter θ^* all along.

We can therefore formulate this problem as the optimization of the expectation Y^θ of the reward $\ell(\omega)$ with respect to the distribution $\mathrm{P}^\theta(\omega)$

$$\max_\theta \; Y^\theta \;=\; \int_\omega \ell(\omega)\,\mathrm{P}^\theta(\omega)$$

on the basis of a finite set of training examples $\omega_1,\dots,\omega_n$ sampled from $\mathrm{P}(\omega)$.

However, it would be unwise to maximize the estimates obtained using approximation (5) because they could reach a maximum for a value of θ that is poorly explored by the actual distribution. As explained in Section 4.5, the gap between the upper and lower bound of inequality (14) reveals the uncertainty associated with insufficient exploration. Maximizing an empirical estimate $\widehat{Y}^\theta$ of the lower bound $\bar{Y}^\theta$ ensures that the optimization algorithm finds a trustworthy answer

$$\theta^* \;=\; \arg\max_\theta \widehat{Y}^\theta \;. \tag{19}$$

We shall now discuss the statistical basis of this learning principle.[9]

6.2 Uniform Confidence Intervals

As discussed in Section 4.4, inequality (14),

$$\bar{Y}^\theta \;\le\; Y^\theta \;\le\; \bar{Y}^\theta + M(1-\bar{W}^\theta)\;,$$

where

$$\bar{Y}^\theta \;=\; \int_\omega \ell(\omega)\,\bar{w}(\omega)\,\mathrm{P}(\omega) \;\;\approx\;\; \widehat{Y}^\theta \;=\; \frac{1}{n}\sum_{i=1}^{n}\ell(\omega_i)\,\bar{w}(\omega_i)\;,$$

$$\bar{W}^\theta \;=\; \int_\omega \bar{w}(\omega)\,\mathrm{P}(\omega) \;\;\approx\;\; \widehat{W}^\theta \;=\; \frac{1}{n}\sum_{i=1}^{n}\bar{w}(\omega_i)\;,$$

9. The idea of maximizing the lower bound may surprise readers familiar with the UCB algorithm for multi-armed bandits (Auer et al., 2002). UCB performs exploration by maximizing the upper confidence interval bound and updating the confidence intervals online. Exploration in our setup results from the active system randomization during the offline data collection. See also Section 6.4.

leads to confidence intervals (15) of the form

$$\forall\delta>0,\ \forall\theta \quad \mathbb{P}\Big\{\ \widehat{Y}^{\theta}-\varepsilon_R\ \leq\ Y^{\theta}\ \leq\ \widehat{Y}^{\theta}+M(1-\widehat{W}^{\theta}+\xi_R)+\varepsilon_R\ \Big\}\geq 1-\delta\,. \tag{20}$$

Both ε_R and ξ_R converge to zero in inverse proportion to the square root of the sample size n. They also increase at most linearly in $\log\delta$ and depend on both the capping bound R and the parameter θ through the empirical variances (see appendix B.)

Such confidence intervals are insufficient to provide guarantees for a parameter value θ^* that depends on the sample. In fact, the optimization (19) procedure is likely to select values of θ for which the inequality is violated. We therefore seek uniform confidence intervals (Vapnik and Chervonenkis, 1968), simultaneously valid for all values of θ.

- When the parameter θ is chosen from a finite set $\mathscr{F}$, applying the union bound to the ordinary intervals (20) immediately gives the uniform confidence interval :

$$\mathbb{P}\Big\{\ \forall\theta\in\mathscr{F},\ \widehat{Y}^{\theta}-\varepsilon_R\leq Y^{\theta}\leq\widehat{Y}^{\theta}+M(1-\widehat{W}^{\theta}+\xi_R)+\varepsilon_R\ \Big\}\geq 1-|\mathscr{F}|\,\delta\,.$$

- Following the pioneering work of Vapnik and Chervonenkis, a broad choice of mathematical tools have been developed to construct uniform confidence intervals when the set $\mathscr{F}$ is infinite. For instance, appendix E leverages uniform empirical Bernstein bounds (Maurer and Pontil, 2009) and obtains the uniform confidence interval

$$\mathbb{P}\Big\{\ \forall\theta\in\mathscr{F},\ \widehat{Y}^{\theta}-\varepsilon_R\leq Y^{\theta}\leq\widehat{Y}^{\theta}+M(1-\widehat{W}^{\theta}+\xi_R)+\varepsilon_R\ \Big\}\geq 1-\mathscr{M}(n)\,\delta\,, \tag{21}$$

  where the growth function $\mathscr{M}(n)$ measures the capacity of the family of functions

$$\{\ f_\theta:\omega\mapsto\ell(\omega)\bar{w}(\omega)\ ,\ \ g_\theta:\omega\mapsto\bar{w}(\omega)\ ,\ \ \forall\theta\in\mathscr{F}\ \}\ .$$

  Many practical choices of $\mathrm{P}^*(\omega)$ lead to functions $\mathscr{M}(n)$ that grow polynomially with the sample size. Because both ε_R and ξ_R are $\mathscr{O}(n^{-1/2}\log\delta)$, they converge to zero with the sample size when one maintains the confidence level $1-\mathscr{M}(n)\,\delta$ equal to a predefined constant.

The interpretation of the inner and outer confidence intervals (Section 4.5) also applies to the uniform confidence interval (21). When the sample size is sufficiently large and the capping bound R chosen appropriately, the inner confidence interval reflects the upper and lower bound of inequality (14).

The uniform confidence interval therefore ensures that Y^{θ^*} is close to the maximum of the lower bound of inequality (14) which essentially represents the best performance that can be guaranteed using training data sampled from $\mathrm{P}(\omega)$. Meanwhile, the upper bound of this same inequality reveals which values of θ could potentially offer better performance but have been insufficiently probed by the sampling distribution (Figure 20.)

6.3 Tuning Ad Placement Auctions

We now present an application of this learning principle to the optimization of auction tuning parameters in the ad placement engine. Despite increasingly challenging engineering difficulties, comparable optimization procedures can obviously be applied to larger numbers of tunable parameters.

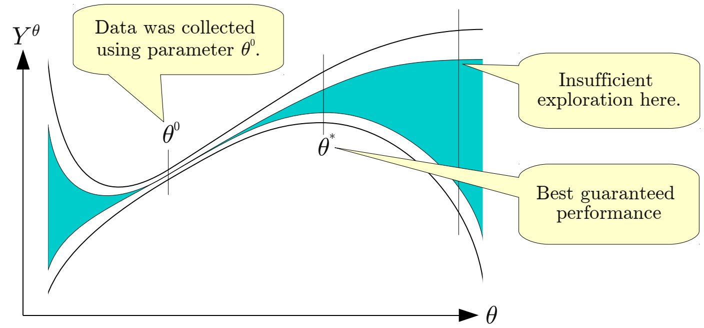

Figure 20: The uniform inner confidence interval reveals where the best guaranteed Y^θ is reached and where additional exploration is needed.

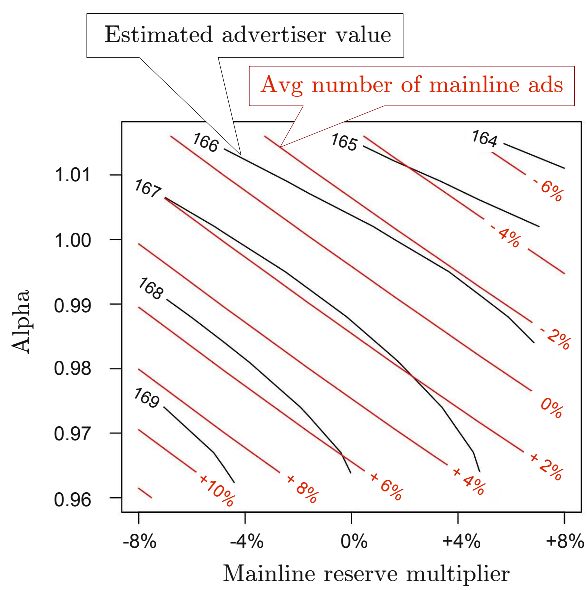

Figure 21: Level curves associated with the average number of mainline ads per page (red curves labelled from -6% to $+10\%$) and the average estimated advertisement value generated per page (black curves, labelled with arbitrary units ranging from 164 to 169) that would have been observed for a certain query cluster if we had changed the mainline reserves by the multiplicative factor shown on the horizontal axis, and if we had applied a squashing exponent α shown on the vertical axis to the estimated click probabilities $q_{i,p}(x)$.

Lahaie and McAfee (2011) propose to account for the uncertainty of the click probability estimation by introducing a squashing exponent α to control the impact of the estimated probabilities on the rank scores. Using the notations introduced in Section 2.1, and assuming that the estimated probability of a click on ad i placed at position p after query x has the form $q_{ip}(x) = \gamma_p\, \beta_i(x)$ (see appendix A), they redefine the rank-score $r_{ip}(x)$ as:

$$r_{ip}(x) = \gamma_p \, b_i \, \beta_i(x)^{\alpha} \; .$$

Using a squashing exponent $\alpha < 1$ reduces the contribution of the estimated probabilities and increases the reliance on the bids b_i placed by the advertisers.

Because the squashing exponent changes the rank-score scale, it is necessary to simultaneously adjust the reserves in order to display comparable number of ads. In order to estimate the counterfactual performance of the system under interventions affecting both the squashing exponent and the mainline reserves, we have collected data using a random squashing exponent following a normal distribution, and a mainline reserve multiplier following a log-normal distribution as described in Section 4.6. Samples describing 12 million search result pages were collected during four consecutive weeks.

Following Charles and Chickering (2012), we consider separate squashing coefficients α_k and mainline reserve multipliers ρ_k per query cluster $k \in \{1..K\}$, and, in order to avoid negative user or advertiser reactions, we seek the auction tuning parameters α_k and ρ_k that maximize an estimate of the advertisement value[10] subject to a global constraint on the average number of ads displayed in the mainline. Because maximizing the advertisement value instead of the publisher revenue amounts to maximizing the size of the advertisement pie instead of the publisher slice of the pie, this criterion is less likely to simply raise the prices without improving the ads. Meanwhile the constraint ensures that users are not exposed to excessive numbers of mainline ads.

We then use the collected data to estimate bounds on the counterfactual expectations of the advertiser value and the counterfactual expectation of the number of mainline ads per page. Figure 21 shows the corresponding level curves for a particular query cluster. We can then run a simple optimization algorithm and determine the optimal auction tuning parameters for each cluster subject to the global mainline footprint constraint. Appendix D describes how to estimate off-policy counterfactual derivatives that greatly help the numerical optimization.

The obvious alternative (see Charles and Chickering, 2012) consists of replaying the auctions with different parameters and simulating the user using a click probability model. However, it may be unwise to rely on a click probability model to estimate the best value of a squashing coefficient that is expected to compensate for the uncertainty of the click prediction model itself. The counterfactual approach described here avoids the problem because it does not rely on a click prediction model to simulate users. Instead it estimates the counterfactual performance of the system using the actual behavior of the users collected under moderate randomization.

6.4 Sequential Design

Confidence intervals computed after a first randomized data collection experiment might not offer sufficient accuracy to choose a final value of the parameter θ. It is generally unwise to simply collect additional samples using the same experimental setup because the current data already reveals information (Figure 20) that can be used to design a better data collection experiment. Therefore, it seems natural to extend the learning principle discussed in Section 6.1 to a sequence of data collection experiments. The parameter θ_t characterizing the t-th experiment is then determined using samples collected during the previous experiments (Figure 22).

10. The value of an ad click from the point of view of the advertiser. The advertiser payment then splits the advertisement value between the publisher and the advertiser.

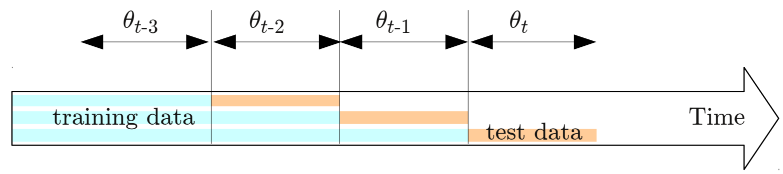

Figure 22: Sequential design – The parameter θ_t of each data collection experiment is determined using data collected during the previous experiments.

Although it is relatively easy to construct convergent sequential design algorithms, reaching the *optimal* learning performance is notoriously difficult (Wald, 1945) because the selection of parameter θ_t involves a trade-off between exploitation, that is, the maximization of the immediate reward Y^{θ_t}, and exploration, that is, the collection of samples potentially leading to better Y^{θ} in the more distant future.

The optimal exploration exploitation trade-off for multi-armed bandits is well understood (Gittins, 1989; Auer et al., 2002; Audibert et al., 2007) because an essential property of multi-armed bandits makes the analysis much simpler: the outcome observed after performing a particular action brings no information about the value of other actions. Such an assumption is both unrealistic and pessimistic. For instance, the outcome observed after displaying a certain ad in response to a certain query brings very useful information about the value of displaying similar ads on similar queries.

Refined contextual bandit approaches (Slivkins, 2011) account for similarities in the context and action spaces but do not take advantage of all the additional opportunities expressed by structural equation models. For instance, in the contextual bandit formulation of the ad placement problem outlined in Section 3.5, actions are pairs (s,c) describing the ad slate s and the corresponding click prices c, policies select actions by combining individual ad scores in very specific ways, and actions determine the rewards through very specific mechanisms.

Meanwhile, despite their suboptimal asymptotic properties, heuristic exploration strategies perform surprisingly well during the time span in which the problem can be considered stationary. Even in the simple case of multi-armed bandits, excellent empirical results have been obtained using Thompson sampling (Chapelle and Li, 2011) or fixed strategies (Vermorel and Mohri, 2005; Kuleshov and Precup, 2010). Leveraging the problem structure seems more important in practice than perfecting an otherwise sound exploration strategy.

Therefore, in the absence of sufficient theoretical guidance, it is both expedient and practical to maximizing $\widehat{Y}^{\theta}$ at each round, as described in Section 6.1, subject to additional ad-hoc constraints ensuring a minimum level of exploration.

7. Equilibrium Analysis

All the methods discussed in this contribution rely on the isolation assumption presented in Section 3.2. This assumption lets us interpret the samples as repeated independent trials that follow the pattern defined by the structural equation model and are amenable to statistical analysis.

The isolation assumption is in fact a component of the counterfactual conditions under investigation. For instance, in Section 4.6, we model single auctions (Figure 3) in order to empirically determine how the ad placement system would have performed if we had changed the mainline reserves *without incurring a reaction from the users or the advertisers*.

Since the future publisher revenues depend on the continued satisfaction of users and advertisers, lifting this restriction is highly desirable.

- We can in principle work with larger structural equation models. For instance, Figure 4 suggests to thread single auction models with additional causal links representing the impact of the displayed ads on the future user goodwill. However, there are practical limits on the number of trials we can consider at once. For instance, it is relatively easy to simultaneously model all the auctions associated with the web pages served to the same user during a thirty minute web session. On the other hand, it is practically impossible to consider several weeks worth of auctions in order to model their accumulated effect on the continued satisfaction of users and advertisers.

- We can sometimes use problem-specific knowledge to construct alternate performance metrics that anticipate the future effects of the feedback loops. For instance, in Section 6.3, we optimize the advertisement value instead of the publisher revenue. Since this alternative criterion takes the advertiser interests into account, it can be viewed as a heuristic proxy for the future revenues of the publisher.

This section proposes an alternative way to account for such feedback loops using the *quasistatic equilibrium* method familiar to physicists: we assume that the publisher changes the parameter θ so slowly that the system remains at equilibrium at all times. Using data collected while the system was at equilibrium, we describe empirical methods to determine how an infinitesimal intervention $\mathrm{d}\theta$ on the model parameters would have displaced the equilibrium:

> "*How would the system have performed during the data collection period if a small change* $\mathrm{d}\theta$ *had been applied to the model parameter* θ *and the equilibrium had been reached before the data collection period.*"

A learning algorithm can then update θ to improve selected performance metrics.

7.1 Rational Advertisers

The ad placement system is an example of game where each actor furthers his or her interests by controlling some aspects of the system: the publisher controls the placement engine parameters, the advertisers control the bids, and the users control the clicks.

As an example of the general quasi-static approach, this section focuses on the reaction of *rational advertisers* to small changes of the scoring functions driving the ad placement system. Rational advertisers always select bids that maximize their economic interests. Although there are more realistic ways to model advertisers, this exercise is interesting because the auction theory approaches also rely on the rational advertiser assumption (see Section 2.1). This analysis seamlessly integrates the auction theory and machine learning perspectives.

As illustrated in Figure 23, we treat the bid vector $b_\star = (b_1 \dots b_A) \in [0, b_{\max}]^A$ as the parameter of the conditional distribution $\mathrm{P}^{b_\star}(b|x,v)$ of the bids associated with the eligible ads.[11] The vari-

11. Quantities measured when a feedback causal system reaches equilibrium often display conditional independence patterns that cannot be represented with directed acyclic graphs (Lauritzen and Richardson, 2002; Dash, 2003). Treating the feedback loop as parameters instead of variables works around this difficulty in a manner that appears sufficient to perform the quasi-static analysis.

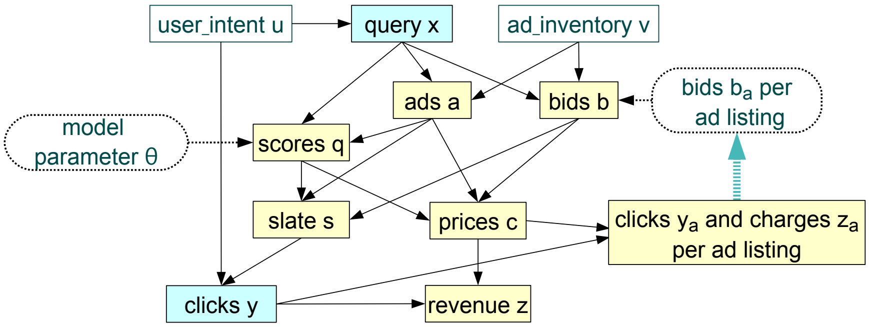

Figure 23: Advertisers select the bid amounts b_a on the basis of the past number of clicks y_a and the past prices z_a observed for the corresponding ads.

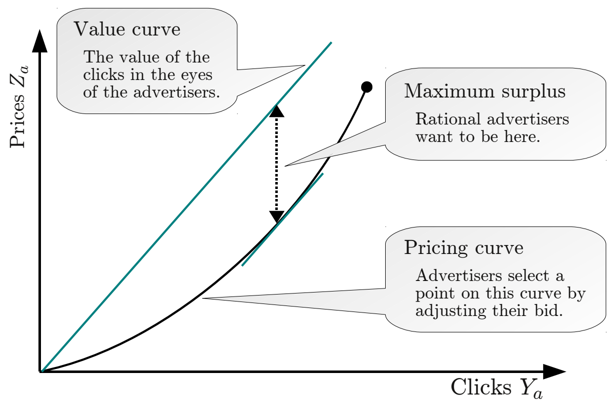

Figure 24: Advertisers control the expected number of clicks Y_a and expected prices Z_a by adjusting their bids b_a. Rational advertisers select bids that maximize the difference between the value they see in the clicks and the price they pay.

ables y_a in the structural equation model represents the number of clicks received by ads associated with bid b_a. The variables z_a represents the amount charged for these clicks to the corresponding advertiser. The advertisers select their bids b_a according to their anticipated impact on the number of resulting clicks y_a and on their cost z_a.

Following the pattern of the perfect information assumption (see Section 2.1), we assume that the advertisers eventually acquire full knowledge of the expectations

$$Y_a(\theta,b_\star) = \int_\omega y_a\, \mathrm{P}^{\theta,b_\star}(\omega) \quad \text{and} \quad Z_a(\theta,b_\star) = \int_\omega z_a\, \mathrm{P}^{\theta,b_\star}(\omega)\,.$$

Let V_a denote the value of a click for the corresponding advertiser. Rational advertiser seek to maximize the difference between the value they see in the clicks and the price they pay to the publisher, as illustrated in Figure 24. This is expressed by the utility functions

$$U_a^\theta(b_\star) \;=\; V_a\, Y_a(\theta,b_\star) - Z_a(\theta,b_\star)\,.$$

Following Athey and Nekipelov (2010), we argue that the injection of smooth random noise into the auction mechanism changes the discrete problem into a continuous problem amenable to standard differential methods. Mild regularity assumption on the densities probability $\mathrm{P}^{b_\star}(b|x,v)$ and $\mathrm{P}^{\theta}(q|x,a)$ are in fact sufficient to ensure that the expectations $Y_a(\theta,b_\star)$ and $Z_a(\theta,b_\star)$ are continuously differentiable functions of the distribution parameters $b_\star$ and θ. Further assuming that utility functions $U_a^\theta(b_\star)$ are diagonally quasiconcave, Athey and Nekipelov establish the existence of a unique Nash equilibrium

$$\forall a \quad b_a \in \operatorname*{ArgMax}_b \; U_a^\theta(b_1,\dots,b_{a-1},b,b_{a+1},\dots,b_A)$$

characterized by its first order Karush-Kuhn-Tucker conditions

$$\forall a \qquad V_a \frac{\partial Y_a}{\partial b_a} - \frac{\partial Z_a}{\partial b_a} \begin{cases} \leq 0 & \text{if } b_a = 0, \\ \geq 0 & \text{if } b_a = b_{\max}, \\ = 0 & \text{if } 0 < b_a < b_{\max}. \end{cases} \tag{22}$$

We use the first order equilibrium conditions (22) for two related purposes. Section 7.2 explains how to complete the advertiser model by estimating the values V_a. Section 7.3 estimates how the equilibrium bids and the system performance metrics respond to a small change $\mathrm{d}\theta$ of the model parameters.

Interestingly, this approach remains sensible when key assumptions of the equilibrium model are violated. The perfect information assumption is unlikely to hold in practice. The quasi-concavity of the utility functions is merely plausible. However, after observing the operation of the stationary ad placement system for a sufficiently long time, it is reasonable to assume that the most active advertisers have tried small bid variations and have chosen locally optimal ones. Less active advertisers may leave their bids unchanged for longer time periods, but can also update them brutally if they experience a significant change in return on investment. Therefore it makes sense to use data collected when the system is stationary to estimate advertiser values V_a that are consistent with the first order equilibrium conditions. We then hope to maintain the conditions that each advertisers had found sufficiently attractive, by first estimating how a small change $\mathrm{d}\theta$ displaces this posited local equilibrium, then by using performance metrics that take this displacement into account.

7.2 Estimating Advertiser Values

We first need to estimate the partial derivatives appearing in the equilibrium condition (22). These derivatives measure how the expectations Y_a and Z_a would have been changed if each advertiser had placed a slightly different bid b_a. Such quantities can be estimated by randomizing the bids and computing on-policy counterfactual derivatives as explained in appendix D. Confidence intervals can be derived with the usual tools.

Unfortunately, the publisher is not allowed to directly randomize the bids because the advertisers expect to pay prices computed using the bid they have specified and not the potentially higher bids resulting from the randomization. However, the publisher has full control on the estimated click probabilities $q_{i,p}(x)$. Since the rank-scores $r_{i,p}(x)$ are the products of the bids and the estimated click probabilities (see Section 2.1), a random multiplier applied to the bids can also be interpreted as a random multiplier applied to the estimated click probabilities. Under these two interpretations, the same ads are shown to the users, but different click prices are charged to the advertisers. Therefore, the publisher can simultaneously charge prices computed as if the multiplier had been applied to the estimated click probabilities, and collect data as if the multiplier had been applied to the bid. This data can then be used to estimate the derivatives.

Solving the first order equilibrium equations then yields estimated advertiser values V_a that are consistent with the observed data.[12]

$$V_a \approx \frac{\partial Z_a}{\partial b_a} \Big/ \frac{\partial Y_a}{\partial b_a} \qquad \text{[Corrected 9/9/2014]}$$

There are however a couple caveats:

- The advertiser bid b_a may be too small to cause ads to be displayed. In the absence of data, we have no means to estimate a click value for these advertisers.

- Many ads are not displayed often enough to obtain accurate estimates of the partial derivatives $\frac{\partial Y_a}{\partial b_a}$ and $\frac{\partial Z_a}{\partial b_a}$. This can be partially remediated by smartly aggregating the data of advertisers deemed similar.

- Some advertisers attempt to capture all the available ad opportunities by placing extremely high bids and hoping to pay reasonable prices thanks to the generalized second price rule. Both partial derivatives $\frac{\partial Y_a}{\partial b_a}$ and $\frac{\partial Z_a}{\partial b_a}$ are equal to zero in such cases. Therefore we cannot recover V_a by solving the equilibrium Equation (22). It is however possible to collect useful data by selecting for these advertisers a maximum bid $b_{\max}$ that prevents them from monopolizing the eligible ad opportunities. Since the equilibrium condition is an inequality when $b_a = b_{\max}$, we can only determine a lower bound of the values V_a for these advertisers.

These caveats in fact underline the limitations of the advertiser modelling assumptions. When their ads are not displayed often enough, advertisers have no more chance to acquire a full knowledge of the expectations Y_a and Z_a than the publisher has a chance to determine their value. Similarly, advertisers that place extremely high bids are probably underestimating the risk to occasionally experience a very high click price. A more realistic model of the advertiser information acquisition is required to adequately handle these cases.

7.3 Estimating the Equilibrium Response

Let $\mathscr{A}$ be the set of the *active advertisers*, that is, the advertisers whose value can be estimated (or lower bounded) with sufficient accuracy. Assuming that the other advertisers leave their bids unchanged, we can estimate how the active advertisers adjust their bids in response to an infinitesimal change $\mathrm{d}\theta$ of the scoring model parameters. This is achieved by differentiating the equilibrium equations (22):

$$\forall a' \in \mathscr{A}, \quad 0 \;=\; \left(V_{a'} \frac{\partial^2 Y_{a'}}{\partial b_{a'} \partial \theta} - \frac{\partial^2 Z_{a'}}{\partial b_{a'} \partial \theta} \right) \mathrm{d}\theta + \sum_{a \in \mathscr{A}} \left(V_{a'} \frac{\partial^2 Y_{a'}}{\partial b_{a'} \partial b_a} - \frac{\partial^2 Z_{a'}}{\partial b_{a'} \partial b_a} \right) \mathrm{d}b_a \,. \tag{23}$$

The partial second derivatives must be estimated as described in appendix D. Solving this linear system of equations then yields an expression of the form

$$\mathrm{d}b_a \;=\; \Xi_a \, \mathrm{d}\theta \,.$$

12. This approach is of course related to the value estimation method proposed by Athey and Nekipelov (2010) but strictly relies on the explicit randomization of the scores. In contrast, practical considerations force Athey and Nekipelov to rely on the apparent noise and hope that the noise model accounts for all potential confounding factors.

This expression can then be used to estimate how any counterfactual expectation Y of interest changes when the publisher applies an infinitesimal change $\mathrm{d}\theta$ to the scoring parameter θ and the active advertisers $\mathscr{A}$ rationally adjust their bids b_a in response:

$$\mathrm{d}Y = \left(\frac{\partial Y}{\partial \theta} + \sum_a \Xi_a \frac{\partial Y}{\partial b_a} \right) \mathrm{d}\theta \ . \tag{24}$$

Although this expression provides useful information, one should remain aware of its limitations. Because we only can estimate the reaction of active advertisers, expression (24) does not includes the potentially positive reactions of advertisers who did not bid but could have. Because we only can estimate a lower bound of their values, this expression does not model the potential reactions of advertisers placing unrealistically high bids. Furthermore, one needs to be very cautious when the system (23) approaches singularities. Singularities indicate that the rational advertiser assumption is no longer sufficient to determine the reactions of certain advertisers. This happens for instance when advertisers cannot find bids that deliver a satisfactory return. The eventual behavior of such advertisers then depends on factors not taken in consideration by our model.

To alleviate these issues, we could alter the auction mechanism in a manner that forces advertisers to reveal more information, and we could enforce policies ensuring that the system (23) remains safely nonsingular. We could also design experiments revealing the impact of the fixed costs incurred by advertisers participating into new auctions. Although additional work is needed to design such refinements, the quasistatic equilibrium approach provides a generic framework to take such aspects into account.

7.4 Discussion

The rational advertiser assumption is the cornerstone of seminal works describing simplified variants of the ad placement problem using auction theory (Varian, 2007; Edelman et al., 2007). More sophisticated works account for more aspects of the ad placement problem, such as the impact of click prediction learning algorithms (Lahaie and McAfee, 2011), the repeated nature of the ad auctions (Bergemann and Said, 2010), or for the fact that advertisers place bids valid for multiple auctions (Athey and Nekipelov, 2010). Despite these advances, it seems technically very challenging to use these methods and account for all the effects that can be observed in practical ad placement systems.

We believe that our counterfactual reasoning framework is best viewed as a modular toolkit that lets us apply insights from auction theory and machine learning to problems that are far more complex than those studied in any single paper. For instance, the quasi-static equilibrium analysis technique illustrated in this section extends naturally to the analysis of multiple simultaneous causal feedback loops involving additional players:

- The first step consists in designing ad-hoc experiments to identify the parameters that determine the equilibrium equation of each player. In the case of the advertisers, we have shown how to use randomized scores to reveal the advertiser values. In the case of the user feedback, we must carefully design experiments that reveal how users respond to changes in the quality of the displayed ads.

- Differentiating all the equilibrium equations yields a linear system of equations linking the variations of the parameter under our control, such as $d\theta$, and all the parameters under the control of the other players, such as the advertiser bids, or the user willingness to visit the site and click on ads. Solving this system and writing the total derivative of the performance measure gives the answer to our question.

Although this programme has not yet been fully realized, the existence of a principled framework to handle such complex interactions is remarkable. Furthermore, thanks to the flexibility of the causal inference frameworks, these techniques can be infinitely adapted to various modeling assumptions and various system complexities.

8. Conclusion

Using the ad placement example, this work demonstrates the central role of causal inference (Pearl, 2000; Spirtes et al., 1993) for the design of learning systems interacting with their environment. Thanks to importance sampling techniques, data collected during randomized experiments gives precious cues to assist the designer of such learning systems and useful signals to drive learning algorithms.

Two recurrent themes structure this work. First, we maintain a sharp distinction between the learning algorithms and the extraction of the signals that drive them. Since real world learning systems often involve a mixture of human decision and automated processes, it makes sense to separate the discussion of the learning signals from the discussion of the learning algorithms that leverage them. Second, we claim that the mathematical and philosophical tools developed for the analysis of physical systems appear very effective for the analysis of causal information systems and of their equilibria. These two themes are in fact a vindication of cybernetics (Wiener, 1948).

Acknowledgments

We would like to acknowledge extensive discussions with Susan Athey, Miroslav Dudík, Patrick Jordan, John Langford, Lihong Li, Sebastien Lahaie, Shie Mannor, Chris Meek, Alex Slivkins, and Paul Viola. We also thank the Microsoft adCenter RnR team for giving us the invaluable opportunity to deploy these ideas at scale and prove their worth. Finally we gratefully acknowledge the precious comments of our JMLR editor and reviewers.

Appendix A. Greedy Ad Placement Algorithms

Section 2.1 describes how to select and place ads on a web page by maximizing the total rank-score (1). Following (Varian, 2007; Edelman et al., 2007) , we assume that the click probability estimates are expressed as the product of a positive position term γ_p and a positive ad term $\beta_i(x)$. The rank-scores can therefore be written as $r_{i,p}(x) = \gamma_p b_i \beta_i(x)$. We also assume that the policy constraints simply state that a web page should not display more than one ad belonging to any given advertiser. The discrete maximization problem is then amenable to computationally efficient greedy algorithms.

Let us fix a layout L and focus on the inner maximization problem. Without loss of generality, we can renumber the positions such that

$$L = \{1, 2, \dots N\} \quad \text{and} \quad \gamma_1 \geq \gamma_2 \geq \cdots \geq 0\,.$$

and write the inner maximization problem as

$$\max_{i_1,\ldots,i_N} \mathscr{R}_L(i_1,\ldots,i_N) \;=\; \sum_{p\in L} r_{i_p,p}(x)$$

subject to the policy constraints and reserve constraints $r_{i,p}(x) \geq R_p(x)$.

Let S_i denote the advertiser owning ad i. The set of ads is then partitioned into subsets $\mathscr{I}_s = \{i : S_i = s\}$ gathering the ads belonging to the same advertiser s. The ads that maximize the product $b_i\beta_i(x)$ within set $\mathscr{I}_s$ are called the best ads for advertiser s. If the solution of the discrete maximization problem contains one ad belonging to advertiser s, then it is easy to see that this ad must be one of the best ads for advertiser s: were it not the case, replacing the offending ad by one of the best ads would yield a higher $\mathscr{R}_L$ without violating any of the constraints. It is also easy to see that one could select any of the best ads for advertiser s without changing $\mathscr{R}_L$.

Let the set $\mathscr{I}^*$ contain exactly one ad per advertiser, arbitrarily chosen among the best ads for this advertiser. The inner maximization problem can then be simplified as:

$$\max_{i_1,\ldots,i_N\in\mathscr{I}^*} \mathscr{R}_L(i_1,\ldots,i_N) \;=\; \sum_{p\in L} \gamma_p\, b_{i_p}\, \beta_{i_p}(x)$$

where all the indices $i_1,\ldots,i_N$ are distinct, and subject to the reserve constraints.

Assume that this maximization problem has a solution $i_1,\ldots,i_N$, meaning that there is a feasible ad placement solution for the layout L. For $k = 1\ldots N$, let us define $I_k^* \subset \mathscr{I}^*$ as

$$I_k^* = \mathop{\mathrm{Arg\,Max}}_{i\in\mathscr{I}^*\setminus\{i_1,\ldots,i_{k-1}\}} \quad b_i\beta_i(x)\,.$$

It is easy to see that I_k^* intersects $\{i_k,\ldots,i_N\}$ because, were it not the case, replacing i_k by any element of I_k^* would increase $\mathscr{R}_L$ without violating any of the constraints. Furthermore it is easy to see that $i_k \in I_k^*$ because, were it not the case, there would be $h > k$ such that $i_h \in I_k^*$, and swapping i_k and i_h would increase $\mathscr{R}_L$ without violating any of the constraints.

Therefore, if the inner maximization problem admits a solution, we can compute a solution by recursively picking $i_1,\ldots,i_N$ from $I_1^*, I_2^*,\ldots,I_N^*$. This can be done efficiently by first sorting the $b_i\beta_i(x)$ in decreasing order, and then greedily assigning ads to the best positions subject to the reserve constraints. This operation has to be repeated for all possible layouts, including of course the empty layout.

The same analysis can be carried out for click prediction estimates expressed as arbitrary monotone combination of a position term $\gamma_p(x)$ and an ad term $\beta_i(x)$, as shown, for instance, by Graepel et al. (2010).

Appendix B. Confidence Intervals

Section 4.4 explains how to obtain improved confidence intervals by replacing the unbiased importance sampling estimator (7) by the clipped importance sampling estimator (10). This appendix provides details that could have obscured the main message.

B.1 Outer Confidence Interval

We first address the computation of the outer confidence interval (12) which describes how the estimator $\widehat{Y}^*$ approaches the clipped expectation $\bar{Y}^*$.

$$\bar{Y}^* = \int_\omega \ell(\omega)\,\bar{w}(\omega)\;\mathrm{P}(\omega) \;\;\approx\;\; \widehat{Y}^* = \frac{1}{n}\sum_{i=1}^{n}\ell(\omega_i)\,\bar{w}(\omega_i)\,.$$

Since the samples $\ell(\omega_i)\,\bar{w}(\omega_i)$ are independent and identically distributed, the central limit theorem (e.g., Cramér, 1946, Section 17.4) states that the empirical average $\widehat{Y}^*$ converges in law to a normal distribution of mean $\bar{Y}^* = \mathbb{E}[\ell(\omega)\,\bar{w}(\omega)]$ and variance $\bar{V} = \mathrm{var}[\ell(\omega)\,\bar{w}(\omega)]$. Since this convergence usually occurs quickly, it is widely accepted to write

$$\mathbb{P}\Big\{\;\widehat{Y}^* - \varepsilon_R \le \bar{Y}^* \le \widehat{Y}^* + \varepsilon_R\;\Big\} \ge 1-\delta\,,$$

with

$$\varepsilon_R \;=\; \mathrm{erf}^{-1}(1-\delta)\;\sqrt{2\bar{V}}\;. \tag{25}$$

and to estimate the variance $\bar{V}$ using the sample variance $\widehat{V}$

$$\bar{V} \;\approx\; \widehat{V} = \frac{1}{n-1}\sum_{i=1}^{n}\Big(\ell(\omega_i)\,\bar{w}(\omega_i) - \widehat{Y}^*\Big)^2\,.$$

This approach works well when the ratio ceiling R is relatively small. However the presence of a few very large ratios makes the variance estimation noisy and might slow down the central limit convergence.

The first remedy is to bound the variance more rigorously. For instance, the following bound results from (Maurer and Pontil, 2009, Theorem 10).

$$\mathbb{P}\left\{\sqrt{\bar{V}} \;>\; \sqrt{\widehat{V}} \;+\; (M-m)R\sqrt{\frac{2\log(2/\delta)}{n-1}}\right\} \le \delta$$

Combining this bound with (25) gives a confidence interval valid with probability greater than $1-2\delta$. Although this approach eliminates the potential problems related to the variance estimation, it does not address the potentially slow convergence of the central limit theorem.

The next remedy is to rely on *empirical Bernstein bounds* to derive rigorous confidence intervals that leverage both the sample mean and the sample variance (Audibert et al., 2007; Maurer and Pontil, 2009).

Theorem 1 (Empirical Bernstein bound) (Maurer and Pontil, 2009, thm 4)
Let $X, X_1, X_2, \ldots, X_n$ be i.i.d. random variable with values in $[a,b]$ and let $\delta > 0$. Then, with probability at least $1-\delta$,

$$\mathbb{E}[X] - M_n \;\le\; \sqrt{\frac{2V_n\log(2/\delta)}{n}} + (b-a)\,\frac{7\log(2/\delta)}{3(n-1)}\;,$$

where M_n and V_n respectively are the sample mean and variance

$$M_n = \frac{1}{n}\sum_{i=1}^{n}X_i\;, \qquad V_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2\;.$$

Applying this theorem to both $\ell(\omega_i)\,\bar{w}(\omega_i)$ and $-\ell(\omega_i)\,\bar{w}(\omega_i)$ provides confidence intervals that hold for for the worst possible distribution of the variables $\ell(\omega)$ and $\bar{w}(\omega)$.

$$\mathbb{P}\left\{\ \widehat{Y}^* - \varepsilon_R \leq \bar{Y}^* \leq \widehat{Y}^* + \varepsilon_R\ \right\} \geq 1 - 2\delta$$

where

$$\varepsilon_R \ = \ \sqrt{\frac{2\widehat{V}\log(2/\delta)}{n}} + M\,R\,\frac{7\log(2/\delta)}{3(n-1)}. \tag{26}$$

Because they hold for the worst possible distribution, confidence intervals obtained in this way are less tight than confidence intervals based on the central limit theorem. On the other hand, thanks to the Bernstein bound, they remains reasonably competitive, and they provide a much stronger guarantee.

B.2 Inner Confidence Interval

Inner confidence intervals are derived from inequality (14) which bounds the difference between the counterfactual expectation Y^* and the clipped expectation $\bar{Y}^*$:

$$0 \ \leq\ Y^* - \bar{Y}^* \ \leq\ M\,(1 - \bar{W}^*)\,.$$

The constant M is defined by assumption (8). The first step of the derivation consists in obtaining a lower bound of $\bar{W}^* - \widehat{W}^*$ using either the central limit theorem or an empirical Bernstein bound.

For instance, applying theorem 1 to $-\bar{w}(\omega_i)$ yields

$$\mathbb{P}\left\{\ \bar{W}^* \ \geq \widehat{W}^* - \sqrt{\frac{2\widehat{V}_w\log(2/\delta)}{n}} - R\,\frac{7\log(2/\delta)}{3(n-1)}\right\} \ \geq\ 1-\delta$$

where $\widehat{V}_w$ is the sample variance of the clipped weights

$$\widehat{V}_w \ = \ \frac{1}{n-1}\sum_{i=1}^{n}\Big(\bar{w}(\omega_i) - \widehat{W}^*\Big)^2.$$

Replacing in inequality (14) gives the outer confidence interval

$$\mathbb{P}\left\{\ \bar{Y}^* \ \leq\ Y^* \ \leq\ \bar{Y}^* + M(1 - \widehat{W}^* + \xi_R)\ \right\} \geq 1-\delta\,.$$

with

$$\xi_R \ = \ \sqrt{\frac{2\widehat{V}_w\log(2/\delta)}{n}} + R\,\frac{7\log(2/\delta)}{3(n-1)}\ . \tag{27}$$

Note that $1 - \widehat{W} + \xi_R$ can occasionally be negative. This occurs in the unlucky cases where the confidence interval is violated, with probability smaller than δ.

Putting together the inner and outer confidence intervals,

$$\mathbb{P}\left\{\ \widehat{Y}^* - \varepsilon_R \leq Y^* \leq \widehat{Y}^* + M(1-\widehat{W}^* + \xi_R) + \varepsilon_R\ \right\} \geq 1 - 3\delta\,,$$

with ε_R and ξ_R computed as described in expressions (26) and (27).

Appendix C. Counterfactual Differences

We now seek to estimate the difference $Y^+ - Y^*$ of the expectations of a same quantity $\ell(\omega)$ under two different counterfactual distributions $\mathrm{P}^+(\omega)$ and $\mathrm{P}^*(\omega)$. These expectations are often affected by variables whose value is left unchanged by the interventions under consideration. For instance, seasonal effects can have very large effects on the number of ad clicks. When these variables affect both Y^+ and Y^* in similar ways, we can obtain substantially better confidence intervals for the difference $Y^+ - Y^*$.

In addition to the notation ω representing all the variables in the structural equation model, we use notation υ to represent all the variables that are not direct or indirect effects of variables affected by the interventions under consideration.

Let $\zeta(\upsilon)$ be a known function believed to be a good predictor of the quantity $\ell(\omega)$ whose counterfactual expectation is sought. Since $\mathrm{P}^*(\upsilon) = \mathrm{P}(\upsilon)$, the following equality holds regardless of the quality of this prediction:

$$\begin{aligned} Y^* \;=\; \int_\omega \ell(\omega)\, \mathrm{P}^*(\omega) \;&=\; \int_\upsilon \zeta(\upsilon)\, \mathrm{P}^*(\upsilon) \;+\; \int_\omega \big[\ell(\omega) - \zeta(\upsilon)\big]\, \mathrm{P}^*(\omega) \\ &=\; \int_\upsilon \zeta(\upsilon)\, \mathrm{P}(\upsilon) \;+\; \int_\omega \big[\ell(\omega) - \zeta(\upsilon)\big]\, w(\omega)\, \mathrm{P}(\omega)\,. \end{aligned} \tag{28}$$

Decomposing both Y^+ and Y^* in this way and computing the difference,

$$Y^+ - Y^* \;=\; \int_\omega \,[\ell(\omega) - \zeta(\upsilon)]\, \Delta w(\omega)\, \mathrm{P}(\omega) \;\approx\; \frac{1}{n}\sum_{i=1}^{n} \big[\ell(\omega_i) - \zeta(\upsilon_i)\big]\, \Delta w(\omega_i)\,,$$

$$\text{with} \qquad \Delta w(\omega) \;=\; \frac{\mathrm{P}^+(\omega)}{\mathrm{P}(\omega)} - \frac{\mathrm{P}^*(\omega)}{\mathrm{P}(\omega)} \;=\; \frac{\mathrm{P}^+(\omega) - \mathrm{P}^*(\omega)}{\mathrm{P}(\omega)}\,.$$

The outer confidence interval size is reduced if the variance of the residual $\ell(\omega) - \zeta(\upsilon)$ is smaller than the variance of the original variable $\ell(\omega)$. For instance, a suitable predictor function $\zeta(\upsilon)$ can significantly capture the seasonal click yield variations regardless of the interventions under consideration. Even a constant predictor function can considerably change the variance of the outer confidence interval. Therefore, in the absence of better predictor, we still can (and always should) center the integrand using a constant predictor.

The rest of this appendix describes how to construct confidence intervals for the estimation of counterfactual differences. Additional bookkeeping is required because both the weights $\Delta w(\omega_i)$ and the integrand $\ell(\omega) - \zeta(\upsilon)$ can be positive or negative. We use the notation υ to represent the variables of the structural equation model that are left unchanged by the intervention under considerations. Such variables satisfy the relations $\mathrm{P}^*(\upsilon) = \mathrm{P}(\upsilon)$ and $\mathrm{P}^*(\omega) = \mathrm{P}^*(\,\omega\backslash\upsilon\,|\,\upsilon)\,\mathrm{P}(\upsilon)$, where we use notation $\omega\backslash\upsilon$ to denote all remaining variables in the structural equation model. An invariant predictor is then a function $\zeta(\upsilon)$ that is believed to be a good predictor of $\ell(\omega)$. In particular, it is expected that $\mathrm{var}[\ell(\omega) - \zeta(\upsilon)]$ is smaller than $\mathrm{var}[\ell(\omega)]$.

C.1 Inner Confidence Interval with Dependent Bounds

We first describe how to construct finer inner confidence intervals by using more refined bounds on $\ell(\omega)$. In particular, instead of the simple bound (8), we can use bounds that depend on invariant variables:

$$\forall \omega \qquad m \le m(\upsilon) \le \ell(\omega) \le M(\upsilon) \le M\,.$$

The key observation is the equality

$$\mathbb{E}[w^*(\omega)|\upsilon] \;=\; \int_{\omega\backslash\upsilon} w^*(\omega)\,\mathrm{P}(\,\omega\backslash\upsilon\,|\,\upsilon) \;=\; \int_{\omega\backslash\upsilon} \frac{P^*(\,\omega\backslash\upsilon\,|\,\upsilon)\,P(\upsilon)}{P(\,\omega\backslash\upsilon\,|\,\upsilon)\,P(\upsilon)}\,P(\,\omega\backslash\upsilon\,|\,\upsilon) \;=\; 1\,.$$

We can then write

$$\begin{aligned} Y^* - \bar{Y}^* \;&=\; \int_\omega \left[w^*(\omega) - \bar{w}^*(\omega)\right]\,\ell(\omega)\,\mathrm{P}(\omega) \;\leq\; \int_\upsilon \mathbb{E}[\,w^*(\omega) - \bar{w}^*(\omega)\,|\,\upsilon\,]\,M(\upsilon)\,\mathrm{P}(\upsilon) \\ &=\; \int_\upsilon (\,1 - \mathbb{E}[\bar{w}^*(\omega)|\upsilon]\,)\,M(\upsilon)\,\mathrm{P}(\upsilon) \;=\; \int_\omega (\,1 - \bar{w}^*(\omega)\,)\,M(\upsilon)\,\mathrm{P}(\omega) \;=\; \mathscr{B}_{\mathrm{hi}}\,. \end{aligned}$$

Using a similar derivation for the lower bound $\mathscr{B}_{\mathrm{lo}}$, we obtain the inequality

$$\mathscr{B}_{\mathrm{lo}} \;\leq\; Y^* - \bar{Y}^* \;\leq\; \mathscr{B}_{\mathrm{hi}}$$

With the notations

$$\widehat{\mathscr{B}}_{\mathrm{lo}} = \frac{1}{n}\sum_{i=1}^{n}(1 - \bar{w}^*(\omega_i))\,m(\upsilon_i)\,, \qquad \widehat{\mathscr{B}}_{\mathrm{hi}} = \frac{1}{n}\sum_{i=1}^{n}(1 - \bar{w}^*(\omega_i))\,M(\upsilon_i)\,,$$

$$\widehat{V}_{\mathrm{lo}} = \frac{1}{n-1}\sum_{i=1}^{n}\Big[\,(1-\bar{w}^*(\omega_i))\,m(\upsilon_i) - \widehat{\mathscr{B}}_{\mathrm{lo}}\Big]^2\,, \quad \widehat{V}_{\mathrm{hi}} = \frac{1}{n-1}\sum_{i=1}^{n}\Big[\,(1-\bar{w}^*(\omega_i))\,M(\upsilon_i) - \widehat{\mathscr{B}}_{\mathrm{hi}}\Big]^2\,,$$

$$\xi_{\mathrm{lo}} \;=\; \sqrt{\frac{2\widehat{V}_{\mathrm{lo}}\log(2/\delta)}{n}} + |m|R\,\frac{7\log(2/\delta)}{3(n-1)}\,, \quad \xi_{\mathrm{hi}} \;=\; \sqrt{\frac{2\widehat{V}_{\mathrm{hi}}\log(2/\delta)}{n}} + |M|R\,\frac{7\log(2/\delta)}{3(n-1)}\,,$$

two applications of theorem 1 give the inner confidence interval:

$$\mathbb{P}\Big\{\;\bar{Y}^* + \widehat{\mathscr{B}}_{\mathrm{lo}} - \xi_{\mathrm{lo}} \;\leq\; Y^* \;\leq\; \bar{Y}^* + \widehat{\mathscr{B}}_{\mathrm{hi}} + \xi_{\mathrm{hi}}\;\Big\} \;\geq\; 1 - 2\delta\,.$$

C.2 Confidence Intervals for Counterfactual Differences

We now describe how to leverage invariant predictors in order to construct tighter confidence intervals for the difference of two counterfactual expectations.

$$Y^+ - Y^* \;\approx\; \frac{1}{n}\sum_{i=1}^{n}\,\left[\,\ell(\omega_i) - \zeta(\upsilon_i)\,\right]\Delta w(\omega_i) \qquad \text{with}\;\; \Delta w(\omega) = \frac{\mathrm{P}^+(\omega) - \mathrm{P}^*(\omega)}{\mathrm{P}(\omega)}\,.$$

Let us define the reweigthing ratios $w^+(\omega) = \mathrm{P}^+(\omega)/\mathrm{P}(\omega)$ and $w^*(\omega) = \mathrm{P}^*(\omega)/\mathrm{P}(\omega)$, their clipped variants $\bar{w}^+(\omega)$ and $\bar{w}^*(\omega)$, and the clipped centered expectations

$$\bar{Y}^+_c = \int_\omega [\ell(\omega) - \zeta(\upsilon)]\,\bar{w}^+(\omega)\,\mathrm{P}(\omega) \quad \text{and} \quad \bar{Y}^*_c = \int_\omega [\ell(\omega) - \zeta(\upsilon)]\,\bar{w}^*(\omega)\,\mathrm{P}(\omega)\,.$$

The outer confidence interval is obtained by applying the techniques of Section B.1 to

$$\bar{Y}^+_c - \bar{Y}^*_c \;=\; \int_\omega\,[\,\ell(\omega) - \zeta(\upsilon)\,]\,[\,\bar{w}^+(\omega) - \bar{w}^*(\omega)\,]\,\mathrm{P}(\omega)\,.$$

Since the weights $\bar{w}^+ - \bar{w}^*$ can be positive or negative, adding or removing a constant to $\ell(\omega)$ can considerably change the variance of the outer confidence interval. This means that one should *always* use a predictor. Even a *constant predictor* can vastly improve the outer confidence interval difference.

The inner confidence interval is then obtained by writing the difference

$$\begin{aligned}\left(Y^{+}-Y^{*}\right)-\left(\bar{Y}_c^{+}-\bar{Y}_c^{*}\right) &= \int_\omega \left[\ell(\omega)-\zeta(\upsilon)\right]\,\left[w^{+}(\omega)-\bar{w}^{+}(\omega)\right]\mathrm{P}(\omega)\\ &- \int_\omega \left[\ell(\omega)-\zeta(\upsilon)\right]\,\left[w^{*}(\omega)-\bar{w}^{*}(\omega)\right]\mathrm{P}(\omega)\end{aligned}$$

and bounding both terms by leveraging υ-dependent bounds on the integrand:

$$\forall\omega \qquad -M \;\leq\; -\zeta(\upsilon) \;\leq\; \ell(\omega)-\zeta(\upsilon) \;\leq\; M-\zeta(\upsilon) \;\leq\; M\,.$$

This can be achieved as shown in Section C.1.

Appendix D. Counterfactual Derivatives

We now consider interventions that depend on a continuous parameter θ. For instance, we might want to know what the performance of the ad placement engine would have been if we had used a parametrized scoring model. Let $\mathrm{P}^\theta(\omega)$ represent the counterfactual Markov factorization associated with this intervention. Let Y^θ be the counterfactual expectation of $\ell(\omega)$ under distribution P^θ.

Computing the derivative of (28) immediately gives

$$\frac{\partial Y^\theta}{\partial\theta} \;=\; \int_w \left[\ell(\omega)-\zeta(\upsilon)\right] w'_\theta(\omega)\,\mathrm{P}(\omega) \;\approx\; \frac{1}{n}\sum_{i=1}^{n}\,\left[\ell(\omega_i)-\zeta(\upsilon_i)\right] w'_\theta(\omega_i)$$

$$\text{with}\quad w_\theta(\omega)=\frac{\mathrm{P}^\theta(\omega)}{\mathrm{P}(\omega)} \qquad \text{and}\quad w'_\theta(\omega)=\frac{\partial w_\theta(\omega)}{\partial\theta}=w_\theta(\omega)\,\frac{\partial\log\mathrm{P}^\theta(\omega)}{\partial\theta}\,. \tag{29}$$

Replacing the expressions $\mathrm{P}(\omega)$ and $\mathrm{P}^\theta(\omega)$ by the corresponding Markov factorizations gives many opportunities to simplify the reweighting ratio $w'_\theta(\omega)$. The term $w_\theta(\omega)$ simplifies as shown in (6). The derivative of $\log\mathrm{P}^\theta(\omega)$ depends only on the factors parametrized by θ. Therefore, in order to evaluate $w'_\theta(\omega)$, we only need to know the few factors affected by the intervention.

Higher order derivatives can be estimated using the same approach. For instance,

$$\frac{\partial^2 Y^\theta}{\partial\theta_i\,\partial\theta_j} \;=\; \int_w \left[\ell(\omega)-\zeta(\upsilon)\right] w''_{ij}(\omega)\,\mathrm{P}(\omega) \;\approx\; \frac{1}{n}\sum_{i=1}^{n}\,\left[\ell(\omega_i)-\zeta(\upsilon_i)\right] w''_{ij}(\omega_i)$$

$$\text{with}\quad w''_{ij}(\omega)=\frac{\partial^2 w_\theta(\omega)}{\partial\theta_i\,\partial\theta_j}=w_\theta(\omega)\,\frac{\partial\log\mathrm{P}^\theta(\omega)}{\partial\theta_i}\,\frac{\partial\log\mathrm{P}^\theta(\omega)}{\partial\theta_j}+w_\theta(\omega)\,\frac{\partial^2\log\mathrm{P}^\theta(\omega)}{\partial\theta_i\,\partial\theta_j}\,.$$

The second term in $w''_{ij}(\omega)$ vanishes when θ_i and θ_j parametrize distinct factors in $\mathrm{P}^\theta(\omega)$.

D.1 Infinitesimal Interventions and Policy Gradient

Expression (29) becomes particularly attractive when $\mathrm{P}(\omega)=\mathrm{P}^\theta(\omega)$, that is, when one seeks derivatives that describe the effect of an infinitesimal intervention on the system from which the data was collected. The resulting expression is then identical to the celebrated *policy gradient* (Aleksandrov et al., 1968; Glynn, 1987; Williams, 1992) which expresses how the accumulated rewards in a reinforcement learning problem are affected by small changes of the parameters of the policy function.

$$\frac{\partial Y^\theta}{\partial \theta} \;=\; \int_\omega \big[\,\ell(\omega) - \zeta(\upsilon)\,\big]\, w'_\theta(\omega)\, \mathrm{P}^\theta(\omega) \;\approx\; \frac{1}{n}\sum_{i=1}^{n} \big[\,\ell(\omega_i) - \zeta(\upsilon_i)\,\big]\, w'_\theta(\omega_i)$$

$$\text{where } \omega_i \text{ are sampled i.i.d. from } \mathrm{P}^\theta \text{ and } w'_\theta(\omega) \;=\; \frac{\partial \log \mathrm{P}^\theta(\omega)}{\partial \theta}\,.$$

Sampling from $\mathrm{P}^\theta(\omega)$ eliminates the potentially large ratio $w_\theta(\omega)$ that usually plagues importance sampling approaches. Choosing a parametrized distribution that depends smoothly on θ is then sufficient to contain the size of the weights $w'_\theta(\omega)$. Since the weights can be positive or negative, centering the integrand with a prediction function $\zeta(\upsilon)$ remains very important. Even a constant predictor ζ can substantially reduce the variance

$$\begin{aligned}
\mathrm{var}[\,(\ell(\omega) - \zeta)\, w'_\theta(\omega)\,] \;&=\; \mathrm{var}[\,\ell(\omega)\, w'_\theta(\omega) - \zeta\, w'_\theta(\omega)\,] \\
&=\; \mathrm{var}[\ell(\omega)\, w'_\theta(\omega)] - 2\,\zeta\, \mathrm{cov}[\,\ell(\omega)\, w'_\theta(\omega),\, w'_\theta(\omega)\,] + \zeta^2\, \mathrm{var}[w'_\theta(\omega)]
\end{aligned}$$

whose minimum is reached for $\zeta = \dfrac{\mathrm{cov}[\ell(\omega) w'_\theta(\omega), w'_\theta(\omega)]}{\mathrm{var}[w'_\theta(\omega)]} = \dfrac{\mathbb{E}[\ell(\omega) w'_\theta(\omega)^2]}{\mathbb{E}[w'_\theta(\omega)^2]}$.

We sometimes want to evaluate expectations under a counterfactual distribution that is too far from the actual distribution to obtain reasonable confidence intervals. Suppose, for instance, that we are unable to reliably estimate which click yield would have been observed if we had used a certain parameter θ^* for the scoring models. We still can estimate how quickly and in which direction the click yield would have changed if we had slightly moved the current scoring model parameters θ in the direction of the target θ^*. Although such an answer is not as good as a reliable estimate of Y^{θ^*}, it is certainly better than no answer.

D.2 Off-Policy Gradient

We assume in this subsection that the parametrized probability distribution $\mathrm{P}^\theta(\omega)$ is regular enough to ensure that all the derivatives of interest are defined and that the event $\{w_\theta(\omega) = R\}$ has probability zero. Furthermore, in order to simplify the exposition, the following derivation does not leverage an invariant predictor function.

Estimating derivatives using data sampled from a distribution $\mathrm{P}(\omega)$ different from $\mathrm{P}^\theta(\omega)$ is more challenging because the ratios $w_\theta(\omega_i)$ in Equation (29) can take very large values. However it is comparatively easy to estimate the derivatives of lower and upper bounds using a slightly different way to clip the weights. Using notation $\mathbb{1}(x)$ represent the indicator function, equal to one if condition x is true and zero otherwise, let us define respectively the clipped weights $\bar{w}^{\mathrm{Z}}_\theta$ and the capped weights $\bar{w}^{\mathrm{M}}_\theta$:

$$\bar{w}^{\mathrm{Z}}_\theta(\omega) = w_\theta(\omega)\, \mathbb{1}\{\mathrm{P}^*(\omega) < R\,\mathrm{P}(\omega)\} \quad \text{and} \quad \bar{w}^{\mathrm{M}}_\theta(\omega) = \min\{w_\theta(\omega), R\}\,.$$

Although Section 4.4 illustrates the use of clipped weights, the confidence interval derivation can be easily extended to the capped weights. Defining the capped quantities

$$\bar{Y}^\theta = \int_\omega \ell(\omega)\, \bar{w}^{\mathrm{M}}_\theta(\omega)\, \mathrm{P}(\omega) \quad \text{and} \quad \bar{W}^\theta = \int_\omega \bar{w}^{\mathrm{M}}_\theta(\omega)\, \mathrm{P}(\omega)$$

and writing

$$\begin{aligned} 0 \;\le\; Y^\theta - \bar{Y}^\theta \;&=\; \int_{\omega\in\Omega\setminus\Omega_R} \ell(\omega)\,\big(\,\mathrm{P}^*(\omega) - R\,\mathrm{P}(\omega)\,\big) \\ &\le\; M\left(1 - \mathrm{P}^*(\Omega_R) - R\,\mathrm{P}(\Omega\setminus\Omega_R)\right) \;=\; M\left(1 - \int_\omega \bar{w}^{\mathrm{M}}_\theta(\omega)\,\mathrm{P}(\omega)\right) \end{aligned}$$

yields the inequality

$$\bar{Y}^\theta \;\le\; Y^\theta \;\le\; \bar{Y}^\theta + M(1-\bar{W}^\theta)\;. \tag{30}$$

In order to obtain reliable estimates of the derivatives of these upper and lower bounds, it is of course sufficient to obtain reliable estimates of the derivatives of $\bar{Y}^\theta$ and $\bar{W}^\theta$. By separately considering the cases $w_\theta(\omega) < R$ and $w_\theta(\omega) > R$, we easily obtain the relation

$$\bar{w}^{\mathrm{M}'}_\theta(\omega) \;=\; \frac{\partial \bar{w}^{\mathrm{M}}_\theta(\omega)}{\partial\theta} \;=\; \bar{w}^{\mathrm{Z}}_\theta(\omega)\,\frac{\partial \log P^\theta(\omega)}{\partial\theta} \quad \text{when } \; w_\theta(\omega) \ne R$$

and, thanks to the regularity assumptions, we can write

$$\begin{aligned} \frac{\partial \bar{Y}^\theta}{\partial\theta} \;&=\; \int_\omega \ell(\omega)\,\bar{w}^{\mathrm{M}'}_\theta(\omega)\,\mathrm{P}(\omega) \;\approx\; \frac{1}{n}\sum_{i=1}^n \ell(\omega_i)\,\bar{w}^{\mathrm{M}'}_\theta(\omega_i)\;, \\ \frac{\partial \bar{W}^\theta}{\partial\theta} \;&=\; \int_\omega \bar{w}^{\mathrm{M}'}_\theta(\omega)\,\mathrm{P}(\omega) \;\approx\; \frac{1}{n}\sum_{i=1}^n \bar{w}^{\mathrm{M}'}_\theta(\omega_i)\,, \end{aligned}$$

Estimating these derivatives is considerably easier than using approximation (29) because they involve the bounded quantity $\bar{w}^{\mathrm{Z}}_\theta(\omega)$ instead of the potentially large ratio $w_\theta(\omega)$. It is still necessary to choose a sufficiently smooth sampling distribution $\mathrm{P}(\omega)$ to limit the magnitude of $\partial \log \mathrm{P}^\theta/\partial\theta$.

Such derivatives are very useful to drive optimization algorithms. Assume for instance that we want to find the parameter θ that maximizes the counterfactual expectation Y^θ as illustrated in Section 6.3. Maximizing the estimate obtained using approximation (5) could reach its maximum for a value of θ that is poorly explored by the actual distribution. Maximizing an estimate of the lower bound (30) ensures that the optimization algorithm finds a trustworthy answer.

Appendix E. Uniform Empirical Bernstein Bounds

This appendix reviews the uniform empirical Bernstein bound given by Maurer and Pontil (2009) and describes how it can be used to construct the uniform confidence interval (21). The first step consists of characterizing the size of a family $\mathscr{F}$ of functions mapping a space $\mathscr{X}$ into the interval $[a,b] \subset \mathbb{R}$. Given n points $\mathbf{x} = (x_1 \ldots x_n) \in \mathscr{X}^n$, the trace $\mathscr{F}(\mathbf{x}) \in \mathbb{R}^n$ is the set of vectors $\big(f(x_1),\ldots,f(x_n)\big)$ for all functions $f \in \mathscr{F}$.

Definition 2 (Covering numbers, etc.) *Given $\varepsilon > 0$, the covering number $\mathscr{N}(\mathbf{x},\varepsilon,\mathscr{F})$ is the smallest possible cardinality of a subset $C \subset \mathscr{F}(\mathbf{x})$ satisfying the condition*

$$\forall v \in \mathscr{F}(\mathbf{x}) \quad \exists c \in C \quad \max_{i=1\ldots n} |v_i - c_i| \;\le\; \varepsilon\;,$$

and the growth function $\mathscr{N}(n,\varepsilon,\mathscr{F})$ is

$$\mathscr{N}(n,\varepsilon,\mathscr{F}) \;=\; \sup_{\mathbf{x}\in\mathscr{X}^n} \mathscr{N}(\mathbf{x},\varepsilon,\mathscr{F})\;.$$

Thanks to a famous combinatorial lemma (Vapnik and Chervonenkis, 1968, 1971; Sauer, 1972), for many usual parametric families $\mathscr{F}$, the growth function $\mathscr{N}(n,\varepsilon,\mathscr{F})$ increases at most polynomially[13] with both n and $1/\varepsilon$.

Theorem 3 (Uniform empirical Bernstein bound) (Maurer and Pontil, 2009, thm 6)
Let $\delta \in (0,1)$, $n >= 16$. Let $X, X_1, \ldots, X_n$ be i.i.d. random variables with values in $\mathscr{X}$. Let $\mathscr{F}$ be a set of functions mapping $\mathscr{X}$ into $[a,b] \subset \mathbb{R}$ and let $\mathscr{M}(n) = 10\,\mathscr{N}(2n,\mathscr{F},1/n)$. Then we probability at least $1-\delta$,

$$\forall f \in \mathscr{F},\ \mathbb{E}[f(X)] - M_n \ \leq\ \sqrt{\frac{18\,V_n \log(\mathscr{M}(n)/\delta)}{n}} + (b-a)\frac{15\log(\mathscr{M}(n)/\delta)}{n-1}\ ,$$

where M_n and V_n respectively are the sample mean and variance

$$M_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i)\ , \qquad V_n = \frac{1}{n-1}\sum_{i=1}^{n}(f(X_i)-M_n)^2\ .$$

The statement of this theorem emphasizes its similarity with the non-uniform empirical Bernstein bound (theorem 1). Although the constants are less attractive, the uniform bound still converges to zero when n increases, provided of course that $\mathscr{M}(n) = 10\,\mathscr{N}(2n,\mathscr{F},1/n)$ grows polynomially with n.

Let us then define the family of functions

$$\mathscr{F} = \{\ f_\theta : \omega \mapsto \ell(\omega)\bar{w}^{\mathrm{M}}_\theta(\omega)\ ,\ \ g_\theta : \omega \mapsto \bar{w}^{\mathrm{M}}_\theta(\omega)\ ,\ \ \forall\theta \in \mathscr{F}\ \}\ ,$$

and use the uniform empirical Bernstein bound to derive an outer inequality similar to (26) and an inner inequality similar to (27). The theorem implies that, with probability $1-\delta$, both inequalities are simultaneously true for all values of the parameter θ. The uniform confidence interval (21) then follows directly.

References

V M. Aleksandrov, V. I. Sysoyev, and V. V. Shemeneva. Stochastic optimization. *Engineering Cybernetics*, 5:11–16, 1968.

Susan Athey and Denis Nekipelov. A structural model of sponsored search advertising. Working paper, 2010. URL `http://kuznets.harvard.edu/~athey/papers/Structural_Sponsored_Search.pdf`.

Jean-Yves Audibert, Remi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In *Proc. 18th International Conference on Algorithmic Learning Theory (ALT 2007)*, pages 150–165, 2007.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fisher. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002.

13. For a simple proof of this fact, slice $[a,b]$ into intervals S_k of maximal width ε and apply the lemma to the family of indicator functions $(x_i, S_k) \mapsto \mathbb{1}\{f(x_i) \in S_k\}$.

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:692–972, 2005.

Dirk Bergemann and Maher Said. Dynamic auctions: a survey. Discussion Paper 1757R, Cowles Foundation for Research in Economics, Yale University, 2010.

Léon Bottou. From machine learning to machine reasoning. `http://arxiv.org/abs/1102.1808v3`, Feb 2011.

Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems 24*, pages 2249–2257. NIPS Foundation, 2011.

C. R. Charig, D. R. Webb, S. R. Payne, and J. E. A. Wickham. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *British Medical Journal (Clin Res Ed)*, 292(6254):879–882, 1986.

Denis X. Charles and D. Max Chickering. Optimization for paid search auctions. Manuscript in preparation, 2012.

Denis X. Charles, D. Max Chickering, and Patrice Simard. Micro-market experimentation for paid search. Manuscript in preparation, 2012.

Harald Cramér. *Mathematical Methods of Statistics*. Princeton University Press, 1946.

Denver Dash. *Caveats for Causal Reasoning with Equilibrium Models*. PhD thesis, University of Pittsburgh, 2003.

Miroslav Dudík, Dimitru Erhan, John Langford, and Lihong Li. Sample-efficient nonstationary-policy evaluation for contextual bandits. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, pages 247–254, 2012.

Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. Internet advertising and the generalized second price auction: Selling billions of dollars worth of keywords. *American Economic Review*, 97(1):242–259, 2007.

Alan Genz. Numerical computation of multivariate normal probabilities. *Journal Computation of Multivariate Normal Probabilities*, 1:141–149, 1992.

John C. Gittins. *Bandit Processes and Dynamic Allocation Indices*. Wiley, 1989.

Peter W. Glynn. Likelihood ratio gradient estimation: an overview. In *Proceedings of the 1987 Winter Simulation Conference*, pages 366–375, 1987.

John Goodwin. Microsoft adCenter. Personal communication, 2011.

Thore Graepel, Joaquin Quinonero Candela, Thomas Borchert, and Ralf Herbrich. Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine. In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010), Invited Applications Track*. Omnipress, 2010.

Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, July 2008.

Volodymyr Kuleshov and Doina Precup. Algorithms for multi-armed bandit problems, October 2010. `http://www.cs.mcgill.ca/~vkules/bandits.pdf`.

Sébastien Lahaie and R. Preston McAfee. Efficient ranking in sponsored search. In *Proc. 7th International Workshop on Internet and Network Economics (WINE 2011)*, pages 254–265. LNCS 7090, Springer, 2011.

Lev Landau and Evgeny Lifshitz. *Course in Theoretical Physics, Volume 1: Mechanics*. Pergamon Press, 1969. 2nd edition.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems 20*, pages 817–824. MIT Press, Cambridge, MA, 2008.

Steffen L. Lauritzen and Thomas S. Richardson. Chain graph models and their causal interpretation. *Journal of the Royal Statistical Society, Series B*, 64:321–361, 2002.

David K. Lewis. *Counterfactuals*. Harvard University Press, 1973. 2nd edition: Wiley-Blackwell, 2001.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on the World Wide Web (WWW 2010)*, pages 661–670. ACM, 2010.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proc. 4th ACM International Conference on Web Search and Data Mining (WSDM 2011)*, pages 297–306, 2011.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample-variance penalization. In *Proc. The 22nd Conference on Learning Theory (COLT 2009)*, 2009.

Paul Milgrom. *Putting Auction Theory to Work*. Cambridge University Press, 2004.

Roger B. Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000. 2nd edition: 2009.

Judea Pearl. Causal inference in statistics: an overview. *Statistics Surveys*, 3:96–146, 2009.

Judea Pearl. The do-calculus revisited. In *Proc. Twenty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI-2012)*, pages 3–11, 2012.

Linda E. Reichl. *A Modern Course in Statistical Physics, 2nd Edition*. Wiley, 1998.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

James M. Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, Sep 2000.

Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory*, 13:145–147, 1972.

Yevgeny Seldin, cois Laviolette Fran Nicolò Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12): 7086–7093, 2012.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Edward H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Ser. B*, 13:238–241, 1951.

Alexsandrs Slivkins. Contextual bandits with similarity information. *JMLR Conference and Workshop Proceedings*, 19:679–702, 2011.

Peter Spirtes and Richard Scheines. Causal inference of ambiguous manipulations. *Philosophy of Science*, 71(5):833–845, Dec 2004.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction and Search*. Springer Verlag, New York, 1993. 2nd edition: MIT Press, Cambridge (Mass.), 2011.

Stephen M. Stigler. A historical view of statistical concepts in psychology and educational research. *American Journal of Education*, 101(1):60–70, Nov 1992.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

Rich S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings 16th Conference on Knowledge Discovery and Data Mining (KDD 2010)*, pages 17–26, 2010.

Vladimir N. Vapnik. *Estimation of Dependences based on Empirical Data*. Springer Series in Statistics. Springer Verlag, Berlin, New York, 1982.

Vladimir N. Vapnik and Alexey Ya. Chervonenkis. Uniform convegence of the frequencies of occurence of events to their probabilities. *Proc. Academy of Sciences of the USSR*, 181(4), 1968. English translation: *Soviet Mathematics - Doklady*, 9:915-918, 1968.

Vladimir N. Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

Hal R. Varian. Position auctions. *International Journal of Industrial Organization*, 25:1163–1178, 2007.

Hal R. Varian. Online ad auctions. *American Economic Review*, 99(2):430–434, 2009.

Joannes Vermorel and Mehryar Mohri. Multi-armed bandit algorithms and empirical evaluation. In *Proc. European Conference on Machine Learning*, pages 437–448, 2005.

Georg H. von Wright. *Explanation and Understanding*. Cornell University Press, 1971.

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

Norbert Wiener. *Cybernetics, or Control and Communication in the Animal and the Machine*. Hermann et Cie (Paris), MIT Press (Cambridge, Mass.), Wiley and Sons (New York), 1948. 2nd Edition (expanded): MIT Press, Wiley and Sons, 1961.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(229–256), 1992.

James Woodward. *Making Things Happen*. Oxford University Press, 2005.

Sewall S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.