September 12, 2012

Couterfactual Reasoning and Learning Systems

Léon Bottou
LEON@BOTTOU.ORG
Microsoft Research, Redmond, WA.

Jonas Peters†
JONAS.PETERS@TUEBINGEN.MPG.DE
Max Planck Institute, Tübingen.

Joaquin Quiñonero-Candela,[a‡] **Denis X. Charles,**[b] **D. Max Chickering,**[b] **Elon Portugualy,**[a] **Dipankar Ray,**[c] **Patrice Simard,**[b] **Ed Snelson**[a]

[a] *Microsoft Cambridge, UK.*
[b] *Microsoft Research, Redmond, WA.*
[c] *Microsoft Online Services Division, Bellevue, WA.*

Abstract

Using the search engine ad placement problem as an example, we explain the central role of causal inference for the design of learning systems interacting with their environments. 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. Thanks to a sharp distinction between the learning algorithms and the extraction of the signals that drive them, these methods can be tailored to causal models with different structures. Thanks to mathematical foundations shared with physics, these signals can describe the response of the system when equilibrium conditions are reached.

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 assumptions that justified the statistical approach in the first place.

- Consider for instance the detection of fraudulent credit card transactions (e.g., Hand and Weston, 2008). Informed by the outputs of a statistical system, the card operator makes decisions such as declining a transaction. These decisions have an immediate impact on the card operator earnings. They also affect the satisfaction of both the merchant and the customer and therefore impact the future earnings of the card operator. The card operator decisions can also change the behavior of the fraudsters. Finally the card operator decisions change the data that are collected and used to train or update the fraud detection system itself.

- Consider the placement of advertisements on the result pages of Internet search engines (e.g., Edelman et al., 2007). The placement decisions depend on the bids of

†. Jonas Peters has moved to ETH Zürich.
‡. Joaquin Quiñonero-Candela has joined Facebook.

the advertisers and on scores computed by statistical machine learning systems. Because these decisions define the contents of the result page proposed to the user, they directly influence both the occurrence of clicks and the corresponding advertiser payments. The placement decisions also impact the user satisfaction with this search engine. Meanwhile the future bids of advertisers depend on how much value they see in the Internet traffic they receive because of their advertisements. Finally the placement decisions affect the collection of potential training data.

Meanwhile the designer of the learning system faces a different set of questions: 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 this web of causes and consequences and produces measurable losses and rewards.

We can phrase similar questions about the parameters of the model. Is it worthwhile to move the parameters along this specific direction? Is it worthwhile to replace them by this new value? An automated way to generate and answer such questions immediately leads to a learning algorithm. Therefore, in order to design a learning algorithm, one also needs to understand how the graph of causes and consequences maps the model outputs into measurable losses and rewards.

This work assumes that we have some knowledge of the structure of the causal graph and describes how to address these questions using the principles of causal inference. The tools described in the following sections are closely related to methods of reinforcement learning (Sutton and Barto, 1998) and methods proposed for various special cases such as multi-armed bandits (Robbins, 1952) and contextual bandits (Langford and Zhang, 2008) problems (see also the bibliographical notes, appendix A.5). The fundamental contribution of this work is to demonstrate the qualitative and quantitative benefits afforded by paying a closer attention to the detailed structure of the graph of causes and consequences.

This paper is structured 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. A special attention is paid to the isolation assumption which allows us to interpret the data as repeated independent trials amenable to statistical analysis.

- Section 4 centers on formulating and answering questions of the form: how would the system have performed during the data collection period if certain interventions had been carried out on the system? This process is called counterfactual analysis because such questions pertain to events that did not happen but could have happened. We describe importance sampling methods for counterfactual analysis, with clear conditions of validity and confidence intervals.

- Section 5 describes useful importance sampling techniques, including techniques to reduce the variance and improve confidence intervals, and techniques to estimate derivatives.

- Section 6 describes how counterfactual analysis provides the essential signal for the design of learning algorithms. Assume that we have identified specific 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 small interventions change the point equilibrium. This provides an elegant and effective way to reason about long-term feedback effects.

This work does not discuss learning algorithms but deals with the identification and the measurement of interpretable signals that justify the actions of humans and machines alike. Whether these signals are exploited by human decision makers or by machine learning algorithms is marginally relevant to our approach. 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. This is not a new idea. Wiener (1948) argues that the study of the propagation of learning signals constitutes the discipline that he calls cybernetics.

2. Difficulties

After giving an overview of the advertisement placement problem, which serves as our main example in this work, 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 containing 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,

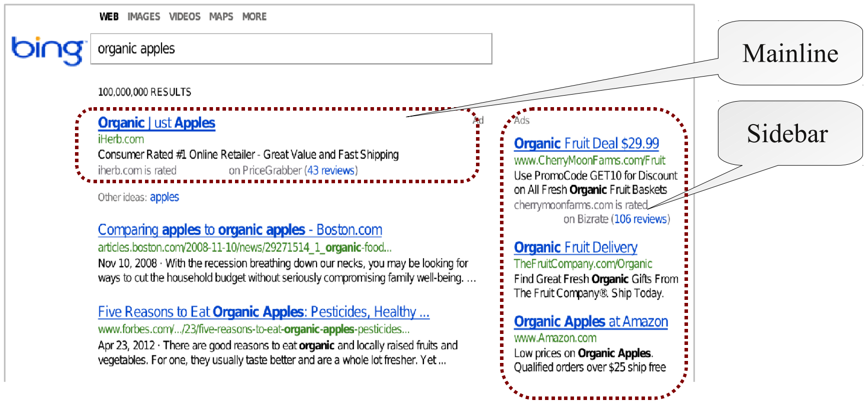

Figure 1: Mainline and sidebar ads on a search result page. Ads placed in the mainline are more likely to be noticed, increasing both the chances of a click if the ad is relevant and the risk of annoying the user if the ad is not relevant.

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):

- 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 $a_1 \ldots a_n$ and the corresponding bids $b_1 \ldots b_n$ 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 $\mathcal{L}$ 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

$$\max_{L\in\mathcal{L}} \max_{i_1,i_2,\ldots} \sum_{p\in L} r_{i_p,p}(x)\,, \tag{1}$$

subject to reserve constraints

$$\forall p \in L,\ \ r_{i_p,p}(x) \geq R_p(x)\,, \tag{2}$$

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.1.)

- 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 excessive numbers 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 advertisement messages 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 demand 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.

For instance, the empirical comparison of certain kidney stone treatments illustrates this difficulty (Charig et al., 1986). Table 1 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.

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.

	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)

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.

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[1] 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 reports the click counts and frequencies observed on the second mainline ad in each group. Although the overall

1. http://bing.com

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.

	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% (44/500)

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 these 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 (e.g., Breiman, 2001). Better models are sometimes obtained by deliberately avoiding to reproduce the true mechanisms (e.g., Vapnik, 1982, section 8.6). We can solve the manipulability puzzle by viewing causation as a component of a reasoning model (Bottou, 2011) rather than a property of the world. In this perspective, causes and effects are only the pieces of reasoning games played in our minds. What makes a collection of causal statements valid is simply the accuracy of the conclusions we reach when we reason 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.

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

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

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 o … sent independent s … es.

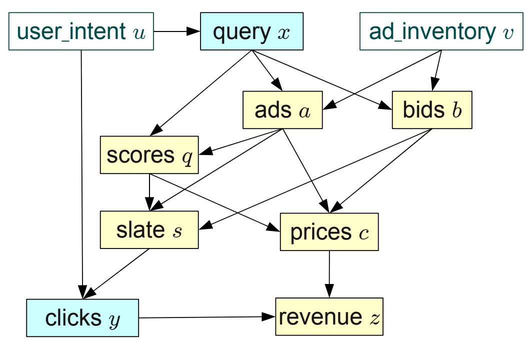

Figure 3: Causal graph associated with the ad placement structural equation model (figure 2). Nodes with yellow (resp. blue) background indicate bound variables with known (resp. unknown) functional dependencies. The mutually independent noise variables are implicit.

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*. 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.

- *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. We no longer 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 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

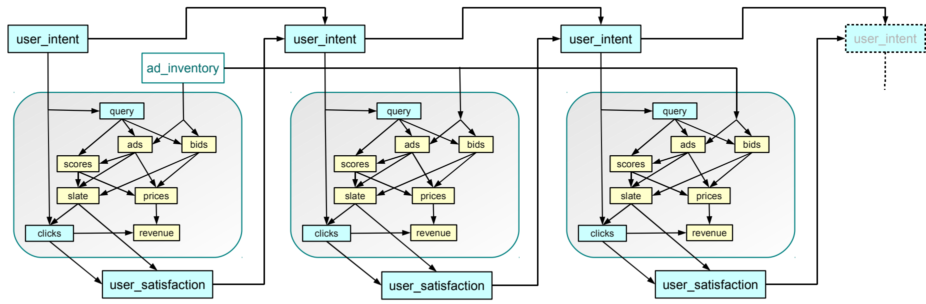

Figure 4: Conceptually unrolling the user feedback loop by threading instances of the single page causal graph (figure 3). Both the ad slate s_t and user clicks y_t have an indirect effect on the user intent u_{t+1} associated with the next query.

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.[2] 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 *isolation assumption* that expresses the idea that the oucome of one trial cannot affect the following trials. This can be achieved by assuming that *the exogenous variables are 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 useful to represent the conditional distribution $\mathrm{P}(\mathsf{effect}\,|\,\mathsf{causes})$ using the equation $\mathsf{effect} = f(\mathsf{causes}, \varepsilon)$. We therefore 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 model shown in figure 2, we assume that the joint distribution of the exogenous variables factorizes as

$$\mathrm{P}(u, v, \varepsilon_1, \ldots, \varepsilon_8) = \mathrm{P}(u, v)\,\mathrm{P}(\varepsilon_1)\ldots\mathrm{P}(\varepsilon_8)\,. \tag{3}$$

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

2. See also the discussion on reinforcement learning, section 3.4.
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.

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

F

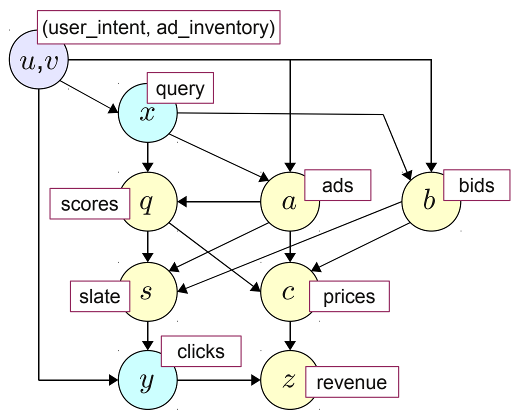

Figure 6: Bayesian network associated with the Markov factorization shown in figure 5.

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 measurements 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 (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 models the joint probability distribution followed by the variables under the isolation assumption.

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 statistical model representing a class of joint probability distributions, and, as such, does not support algebraic manipulations. However its Markov factorization is an algebraic object, essentially equivalent to the structural equation model.

Consider a causal system represented by a structural equation model with some unknown functional dependencies. We can collect statistical data during experiments involving different interventions on the causal system. These interventions can be represented as algebraic manipulations of the structural equation. Each intervention leads to a different Bayesian network representing the joint probability distribution of the data collected during the corresponding experiment. However, the Markov factorizations of all these Bayesian networks share factors with the original Markov factorization. If one experiment allows us to discover some aspect of one of these shared factors, we can transfer this discovery into the Bayesian networks describing the statistical properties of other experiments. The causal modeling framework presented in this section is therefore a powerful transfer learning scheme characterized by a family of statistical models endowed with an algebraic structure. Such a scheme is sometimes called a *reasoning model* (Bottou, 2011).

3.4 Special Cases

Three special cases of causal models with increasing generality are particularly relevant.

- 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 (e.g., 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

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

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.

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

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

$$\begin{array}{rcll} \boldsymbol{a}_t & = & \boldsymbol{\pi}(\boldsymbol{s}_{t-1},\ \boldsymbol{\varepsilon}_t) & \text{Action} \\ \boldsymbol{y}_t & = & \boldsymbol{r}(\boldsymbol{s}_{t-1},\ \boldsymbol{a}_t,\ \boldsymbol{\varepsilon}'_t\,) & \text{Reward } r_t \in \mathbb{R} \\ \boldsymbol{s}_t & = & \boldsymbol{s}(\boldsymbol{s}_{t-1},\ \boldsymbol{a}_t,\ \boldsymbol{\varepsilon}''_t\,) & \text{Next state} \end{array}$$

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.

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 5.1 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 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 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^* ?*"

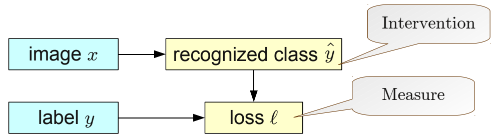

Figure 10: Causal graph for an image recognition system. We can estimate counterfactuals by replaying data collected in the past.

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?*"

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 counterfactuals statements to reason about cause 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.

4.1 Replaying a Data Set

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 labeled images, run the classifier on each image, and average the resulting losses.

We can then *replay* the data set at will to estimate what (counterfactual) performance would have been observed if we had used a different classifier. We can then select in

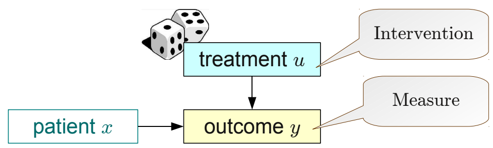

Figure 11: Causal graph for a randomized experiment. We can estimate certain counterfactuals by reweighting data collected in the past.

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).

This procedure 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.

4.2 Randomized Experiments

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.

The *common cause principle* (Reichenbach, 1956) formalizes a fundamental intuition about causation: if two events are correlated, then the first event causes the second event, or the second event causes the first event, or the two events have common causes. 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 was a roll of the dices, we can reject two of the three cases enumerated by the common cause principle. Having eliminated the possibility of a confounding common cause, we can simply *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 discussed in section 4.2 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. The click yield is therefore a very useful metric when we reason with an isolation assumption that ignores the advertiser reactions to pricing changes.

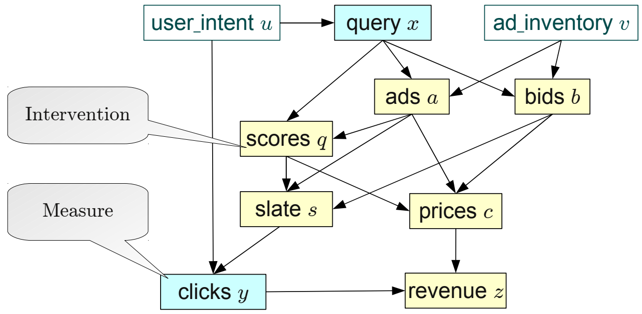

Figure 12: Estimating which average number of clicks per page would have been observed if we had used a different scoring model.

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

$$\begin{aligned} \mathrm{P}(\omega) \;=\;& \mathrm{P}(u,v)\,\mathrm{P}(x\,|\,u)\,\mathrm{P}(a\,|\,x,v)\,\mathrm{P}(b\,|\,x,v)\,\mathrm{P}(q\,|\,x,a) \\ & \times\,\mathrm{P}(s\,|\,a,q,b)\,\mathrm{P}(c\,|\,a,q,b)\,\mathrm{P}(y\,|\,s,u)\,\mathrm{P}(z\,|\,y,c)\;. \end{aligned} \tag{4}$$

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.

$$\begin{aligned} \mathrm{P}^*(\omega) \;=\;& \mathrm{P}(u,v)\,\mathrm{P}(x\,|\,u)\,\mathrm{P}(a\,|\,x,v)\,\mathrm{P}(b\,|\,x,v)\,\mathbf{P^*(\boldsymbol{q}\,|\,\boldsymbol{x},\boldsymbol{a})} \\ & \times\,\mathrm{P}(s\,|\,a,q,b)\,\mathrm{P}(c\,|\,a,q,b)\,\mathrm{P}(y\,|\,s,u)\,\mathrm{P}(z\,|\,x,c)\;. \end{aligned} \tag{5}$$

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

$$Y^* \;=\; \int_\omega y\;\mathrm{P}^*(\omega) \;=\; \int_\omega y\;\frac{\mathrm{P}^*(q\,|\,x,a)}{\mathrm{P}(q\,|\,x,a)}\;\mathrm{P}(\omega) \;\approx\; \frac{1}{n}\sum_{i=1}^{n} y_i\;\frac{\mathrm{P}^*(q_i\,|\,x_i,a_i)}{\mathrm{P}(q_i\,|\,x_i,a_i)}\;, \tag{6}$$

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 $\ell(\omega)$:

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

with weights

$$w_i \;=\; w(\omega_i) \;=\; \frac{\mathrm{P}^*(\omega_i)}{\mathrm{P}(\omega_i)} \;=\; \frac{\text{factors appearing in } \mathrm{P}^*(\omega_i) \text{ but not in } \mathrm{P}(\omega_i)}{\text{factors appearing in } \mathrm{P}(\omega_i) \text{ but not in } \mathrm{P}^*(\omega_i)}\,. \tag{8}$$

Equation (8) 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 (8).

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 (8) 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

$$Y^* = \int_\omega \ell(\omega)\, w(\omega)\; \mathrm{P}(\omega) \quad\approx\quad \frac{1}{n}\sum_{i=1}^{n} \ell(\omega_i)\, w_i\,. \tag{9}$$

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.

In other words, 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. We can in fact obtain better estimates by containing the importance of the 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 $\ell(\omega)$:

$$\forall \omega \quad 0 \leq \ell(\omega) \leq M \tag{10}$$

We control the importance of these noise-inducing samples by replacing the importance sampling weights $w(\omega)$ by capped weights $\bar{w}(\omega)$. Let R be the maximum weight value deemed acceptable, and $\mathbb{1}\{c\}$ be the indicator function, which is equal to 1 when condition c is true and 0 otherwise. *Zero-capped* weights

$$\bar{w}(\omega) = w(\omega)\,\mathbb{1}\{w(\omega) < R\}$$

eliminate the contribution of poorly explored domains, whereas *max-capped* weights

$$\bar{w}(\omega) = \min\{w(\omega), R\}$$

merely contain their importance. In practice, although the theory calls for choosing the constant R before observing the data, we have obtained very consistent results using zero-capped weights and choosing R equal to the fifth largest reweighting ratio observed on the empirical data.

The capped expectation

$$\bar{Y}^* = \int_\omega \ell(\omega)\,\bar{w}(\omega)\ \mathrm{P}(\omega) \quad \approx \quad \widehat{Y}^* = \frac{1}{n}\sum_{i=1}^{n} \ell(\omega_i)\,\bar{w}(\omega_i)\,. \tag{11}$$

is much easier to estimate than (9) because the magnitude of the capped weights $\bar{w}(\omega)$ is bounded by R. Much better confidence intervals are easily obtained using either the central limit theorem or empirical Bernstein bounds (see appendix A.2 for details).

The estimation error $Y^* - \widehat{Y}^*$ can be split into two components $\bar{Y}^* - \widehat{Y}^*$ and $Y^* - \bar{Y}^*$ that can be bounded using confidence intervals that we call *outer confidence intervals* and *inner confidence intervals*.

The outer confidence interval has the form:

$$\mathbb{P}\Big\{\ \widehat{Y}^* - \epsilon_R \ \leq\ \bar{Y}^* \ \leq\ \widehat{Y}^* + \epsilon_R\ \Big\} \ \geq\ 1 - \delta\,, \tag{12}$$

where we use symbol $\mathbb{P}$ instead of P to emphasize that this probability is not associated with the Markov factorization but represents a random draw of a sample $\omega_1 \ldots \omega_n$.

The difference $Y^* - \bar{Y}^*$ then represents the contribution of the poorly explored domains to the expectation Y^*. Since the samples do not provide reliable information about such domains, we bound this difference using assumption (10) :

$$0 \ \leq\ Y^* - \bar{Y}^* \ =\ \int_\omega \ell(\omega)\,\big[w(\omega) - \bar{w}(\omega)\big]\ \mathrm{P}(\omega) \ \leq\ M\int_\omega \big[w(\omega) - \bar{w}(\omega)\big]\ \mathrm{P}(\omega)\,. \tag{13}$$

In order to estimate these bounds using the empirical data, observe that

$$\int_\omega \big[w(\omega) - \bar{w}(\omega)\big]\ \mathrm{P}(\omega) \ =\ \int_\omega \frac{\mathrm{P}^*(\omega)}{\mathrm{P}(\omega)}\ \mathrm{P}(\omega) - \int_\omega \bar{w}(\omega)\ \mathrm{P}(\omega) \ =\ 1 - \bar{W}^*\ .$$

where the quantity

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

is easy to estimate because the magnitude of the capped weight $\bar{w}(\omega)$ is conveniently bounded. We can then rewrite inequality (13) as

$$0 \;\leq\; Y^* - \bar{Y}^* \;\leq\; M\left(1 - \bar{W}^*\right). \tag{14}$$

The standard techniques then yield an inner confidence interval of the form

$$\mathbb{P}\Big\{\; \bar{Y}^* \;\leq\; Y^* \;\leq\; \bar{Y}^* + M(1 - \widehat{W}^* + \xi_R) \;\Big\} \;\geq\; 1 - \delta\,. \tag{15}$$

Putting (12) and (15) together yields our final confidence interval:

$$\mathbb{P}\Big\{\; \widehat{Y}^* - \epsilon_R \;\leq\; Y^* \;\leq\; \widehat{Y}^* + M(1 - \widehat{W}^* + \xi_R) + \epsilon_R \;\Big\} \geq 1 - 2\delta\,. \tag{16}$$

Replacing the unbiased importance sampling estimator (9) by a capped importance sampling estimator (11) therefore leads to improved confidence intervals. This replacement also eliminates the need for the assumption $\forall\omega,\ \mathrm{P}(\omega) > 0$ because we can define $\bar{w}(\omega)$ by continuity whenever $\mathrm{P}(\omega) = 0$. The only difference is the need to consider the two cases $\mathrm{P}(\omega) = 0$ and $\mathrm{P}(\omega) > 0$ when deriving inequality (14).

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. We argue in this section that 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.

Let the set Ω_R contain all the ω excluded from the computation of the zero-capped expectation $\bar{Y}$, that is, $\Omega_R = \{\omega \;:\; \mathrm{P}^*(\omega) > R\,\mathrm{P}(\omega)\}$. The size of the inner confidence interval always exceeds the gap G_R separating the upper and lower bound of inequality (14).

$$G_R \;=\; M(1 - \bar{W}^*) \;=\; M \int_{\omega\in\Omega_R} w(\omega)\,\mathrm{P}(\omega)$$

Since the gap is a positive decreasing function of R,

$$G_R \;\xrightarrow{R\to\infty}\; G_\infty \;=\; M \int_{\omega\in\Omega_\infty} w(\omega)\,\mathrm{P}(\omega) \;\geq\; 0\,,$$

where $\Omega_\infty = \bigcap_{R>0} \Omega_R = \{\omega \;:\; \mathrm{P}^*(\omega) > 0 \text{ and } \mathrm{P}(\omega) = 0\}$. Therefore the inner confidence interval size always exceeds the quantity G_∞ which represents the impact of unexplored domains on the counterfactual expectation Y^*. Since increasing the sample cannot probe domain Ω_∞, the only way to reduce the inner confidence below the minimum gap is to collect data using a different distribution.

In order to obtain a good confidence interval (16), we ideally would like to select a capping bound R that achieves a good compromise between the non-increasing gap G_R and the uncertainties ξ_R and ϵ_R resulting from the sample size n and from the non-decreasing variances $\mathrm{var}[\bar{w}(\omega)]$ and $\mathrm{var}[\bar{w}(\omega)\ell(\omega)]$.

This is relatively easy when the weights (8) can be expressed as the ratio of discrete distributions defined on domains with relatively small cardinality. The gap and the variances are then piecewise constant functions of R. A single well chosen value of R can usually handle a broad range of realistic sample sizes. The innner confidence interval then measures the uncertainty associated with the insufficiently explored domain $\Omega_R \supset \Omega_\infty$.

The situation is far more complex with continuous distributions. One one hand, we can select distributions $\mathrm{P}(\omega)$ that cover large domains and therefore ensure that G_∞ is zero or close to zero. On the other hand, since the integral of the density sums to the unity, such distributions tend to have small densities, leading to potentially very large weights $w(\omega)$ and therefore extremely large variances[4] that cannot be countered with realistically large sample sizes. Furthermore, when R reaches large values, the variances $\mathrm{var}[\bar{w}(\omega)]$ and $\mathrm{var}[\bar{w}(\omega)\ell(\omega)]$ are even harder to estimate than the capped expectation $\bar{Y}$. This means in practice that we must conservatively pick a value of R that is well below the theoretically optimal choice. With such a setup, the inner confidence interval is an approximation of the gap G_R which measures the uncertainty associated with the insufficiently explored domain Ω_R.

In both cases, assuming that the capping bound R has been chosen competently, the two components of the zero-capped confidence intervals (16) provide precious guidelines:

- The *inner confidence interval* (15) 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.

This interpretation is less obvious in the case of max-capped weights. It nevertheless remains useful because both capping methods usually produce comparable numerical results.

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

4. Consider for instance two normal distribution with unit variance and means respectively equal to zero and m. The ratio of their densities follows a log-normal distribution of variance $e^{m^2} - 1$.

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[5] 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, 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 green 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.

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

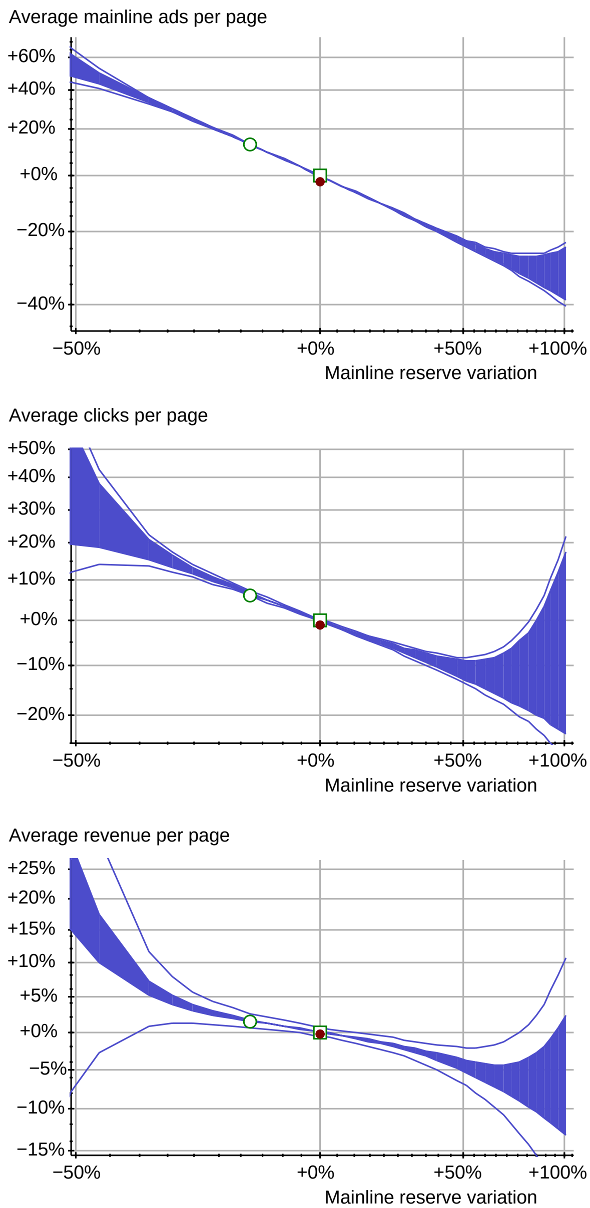

Figure 13: Estimated variations of three performance metrics in response to mainline reserve changes. The curves delimit 95% confidence intervals for the metrics we would have observed if we had increased the mainline reserves by the percentage shown on the horizontal axis. The filled areas represent 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.

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.

- *Point-wise 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 point-wise 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 advertisement 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 difficult it is to collect data that effectively explores all these dimensions. The next section presents a portfolio of methods to work around this problem.

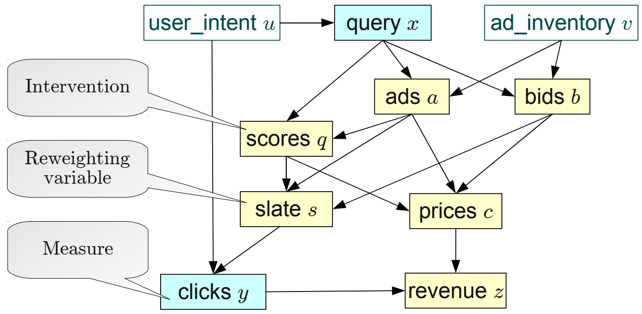

Figure 14: The reweighting variable(s) must intercept all causal paths from the point of intervention to the point of measurement.

5. Importance Sampling Toolbox

This section describes techniques that improve our ability to give answers to counterfactual questions. Leveraging the structure of the causal graph provides opportunities to improve the inner confidence intervals. Exploiting an invariant prediction function can improve outer confidence intervals when comparing the expectations of a same variable under two different counterfactual distributions. Computing the derivatives of a counterfactual expectation with respect to parameters describing the intervention provides directional answers.

5.1 Better Reweighting Variables

Many search result pages come without eligible ads. Regardless of the reserves, we know with certainty that such pages will have zero mainline ads, receive zero clicks, and generate zero revenue. The results shown in figure 13 were in fact helped by a little optimization: when a sample ω_i describes a page without ads, the weights $w(\omega_i)$ are forced to 1. This does not change the estimate since these weights are multiplied by zero, but this reduces the variance of the weights and therefore improves the inner confidence intervals.

Such prior knowledge is in fact encoded in the structure of the causal graph. 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 (see figure 14).

Because the causal graph has this special structure, we can simplify both the actual and counterfactual Markov factorizations (4) (5) 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.

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

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

$$\begin{aligned} Y^* &= \int_{\dots} y\ \mathrm{P}^*(u,v,x,a,b,s,y) \ = \ \int_{\dots} y\ \frac{\mathrm{P}^*(\,s\,|\,x,a,b)}{\mathrm{P}(\,s\,|\,x,a,b)}\ \mathrm{P}(u,v,x,a,b,s,y) \\ &\approx \frac{1}{n}\sum_{i=1}^{n}\ y_i\,\frac{\mathrm{P}^*(\,s_i\,|\,x_i,a_i,b_i)}{\mathrm{P}(\,s_i\,|\,x_i,a_i,b_i)}\ . \end{aligned} \tag{17}$$

Comparing (6) and (17) 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 14, we now distinguish the reweighting variable from the intervention.

In general, the algebraic manipulation described above is possible whenever the reweighting variable (or variables) intercepts all the causal paths connecting the point of intervention to the measurement point. The numerator and the denominator of the reweighting ratio are then computed by collapsing all the factors connecting the intervention point to the intercepting reweighting variable along each of these causal paths. In order to be able to evaluate the weights, these factors must of course be known. We can only improve the confidence intervals by using reweighting variables that are closer to the measurement point and respect these constraints.

We have reproduced the experiments described in section 4.6 with the counterfactual estimate (17) instead of (6). 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 15 shows counterfactual estimates obtained using the same data as figure 13. The obvious improvement of the inner confidence intervals significantly extends the range of mainline reserve multipliers for which we can compute accurate counterfactual expectations using this same data. The figure 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. Although reweighting according to both the ad slate s and the click prices c does not bring much benefit, we have obtained nice counterfactual revenue estimates by reweighting according to both the ad slate s and a new variable representing solely the click prices of those ads that actually received clicks.

Figure 16 shows how this approach can be extended to the randomization of all the scores using a collection of independent log-normal multipliers. The weights are then computed as the ratio of the probabilities of the observed ad slate under the counterfactual and actual multiplier distributions. Details will be provided in a forthcoming publication.

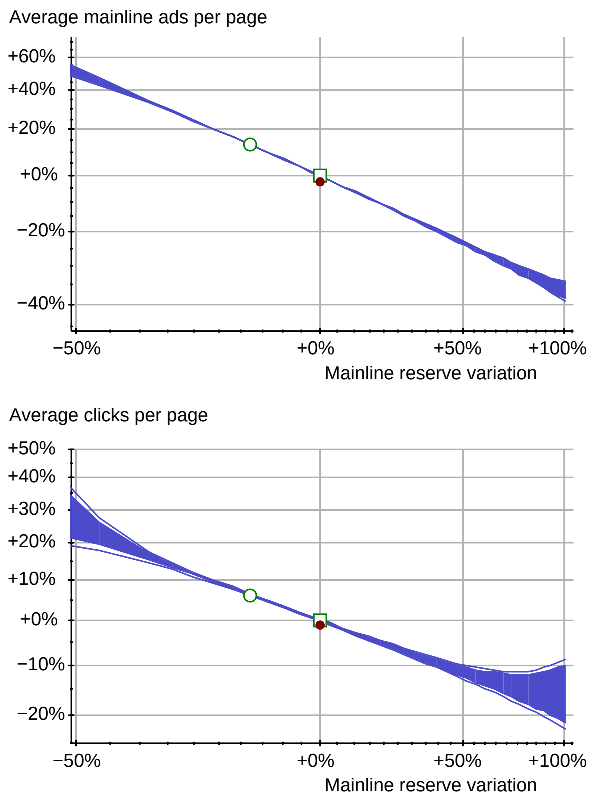

Figure 15: Estimated variations of two performance metrics in response to mainline reserve changes. These estimates were obtained using the ad slates s as reweighting variable. Compare the inner confidence intervals with those shown in figure 13.

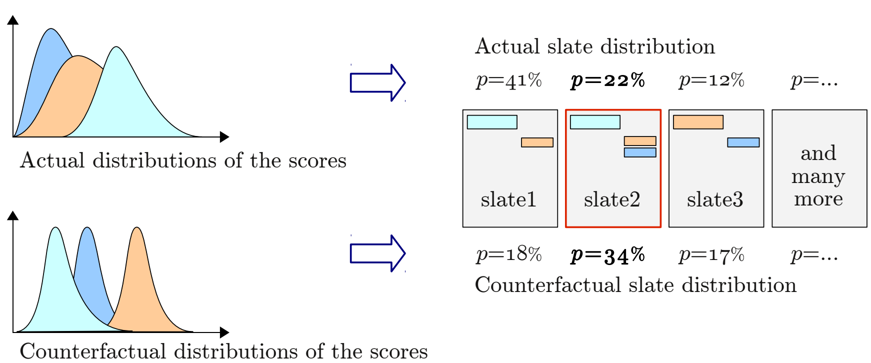

Figure 16: A distribution on the scores q induce a distribution on the possible ad slates s. If the observed slate is `slate2`, the reweighting ratio is 34/22.

5.2 Counterfactual Differences

The designer of a learning system often has to evaluate which of two interventions is most likely to improve the system performance. This can be achieved by estimating the difference $Y^+ - Y^*$ of 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. We can then expect that these variables affect both Y^+ and Y^* in similar ways. This provides an opportunity to 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:

$$\begin{aligned} Y^* \;=\; \int_\omega \ell(\omega)\; \mathrm{P}^*(\omega) \;&=\; \int_\upsilon \zeta(\upsilon)\; \mathrm{P}^*(\upsilon) \;+\; \int_\omega \left[\ell(\omega) - \zeta(\upsilon)\right]\; \mathrm{P}^*(\omega) \\ &=\; \int_\upsilon \zeta(\upsilon)\; \mathrm{P}(\upsilon) \;+\; \int_\omega \left[\ell(\omega) - \zeta(\upsilon)\right]\; w(\omega)\; \mathrm{P}(\omega)\;. \end{aligned} \tag{18}$$

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

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

$$\text{with} \qquad \Delta w(\omega) \;=\; \frac{\mathrm{P}^+(\omega)}{\mathrm{P}(\omega)} - \frac{\mathrm{P}^*(\omega)}{\mathrm{P}(\omega)} \;=\; \frac{\mathrm{P}^+(\omega) - \mathrm{P}^*(\omega)}{\mathrm{P}(\omega)}\,. \tag{19}$$

The construction of confidence intervals for such an estimator of the difference $Y^+ - Y^*$ demands additional bookkeeping because both the weights $\Delta w(\omega_i)$ and the integrand $\ell(\omega) - \zeta(\upsilon)$ can now be positive or negative. 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. Details are provided in appendix A.3.2.

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.

5.3 Estimating 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 (18) immediately gives

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

$$\text{with} \quad w_\theta(\omega) = \frac{\mathrm{P}^\theta(\omega)}{\mathrm{P}(\omega)} \qquad \text{and} \quad w'_\theta(\omega) = \frac{\partial w_\theta(\omega)}{\partial \theta} = w_\theta(\omega)\, \frac{\partial \log \mathrm{P}^\theta(\omega)}{\partial \theta}\,. \tag{20}$$

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 (8). 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,

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

$$\text{with} \quad w''_{ij}(\omega) = \frac{\partial^2 w_\theta(\omega)}{\partial \theta_i\, \partial \theta_j} = w_\theta(\omega)\, \frac{\partial \log \mathrm{P}^\theta(\omega)}{\partial \theta_i}\, \frac{\partial \log \mathrm{P}^\theta(\omega)}{\partial \theta_j} + w_\theta(\omega)\, \frac{\partial^2 \log \mathrm{P}^\theta(\omega)}{\partial \theta_i\, \partial \theta_j}\,. \tag{21}$$

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

5.4 Infinitesimal Interventions

Expression (20) 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* (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.

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

$$\text{where } \omega_i \text{ are sampled i.i.d. from } \mathrm{P}^\theta \text{ and } w'_\theta(\omega) \;=\; \frac{\partial \log \mathrm{P}^\theta(\omega)}{\partial \theta}\,. \tag{22}$$

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 $\zeta = \dfrac{\mathrm{cov}[\ell w'_\theta, w'_\theta]}{\mathrm{var}[w'_\theta]} = \dfrac{\mathbb{E}[\ell {w'_\theta}^2]}{\mathbb{E}[{w'_\theta}^2]}$.

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.

5.5 Off-Policy Derivatives

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 (20) can take very large values. However it is comparatively easy to estimate the derivatives of lower and upper bounds computed using max-capped weights.

Let $\bar{w}^{\mathrm{Z}}_\theta$ and $\bar{w}^{\mathrm{M}}_\theta$ be respectively the zero-capped and max-capped weights,

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

We assume 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.

Proceeding as in section 4.4, we define the quantities

$$\bar{Y}^\theta = \int_\omega \ell(\omega)\, \bar{w}^{\mathrm{M}}_\theta(\omega)\, \mathrm{P}(\omega) \quad \text{and} \quad \bar{W}^\theta = \int_\omega \bar{w}^{\mathrm{M}}_\theta(\omega)\, \mathrm{P}(\omega) \tag{23}$$

and obtain the inequality

$$\bar{Y}^\theta \;\leq\; Y^\theta \;\leq\; \bar{Y}^\theta + M(1 - \bar{W}^\theta)\ . \tag{24}$$

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

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

and, thanks to the regularity assumptions, we can write

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

Estimating these derivatives is considerably easier than using approximation (20) because they involve the bounded quantity $\bar{w}^{\mathrm{Z}}_\theta(\omega)$ instead of the potentially large ratio $w_\theta(\omega)$. It

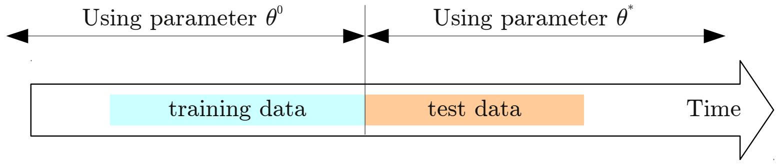

Figure 17: Single design – A preferred parameter value θ^* is determined using randomized data collected in the past. Test data is collected after loading θ^* into the system.

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. Maximizing the estimate obtained using approximation (7) is unwise because it could reach its 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 reveals the uncertainty associated with insufficient exploration. Therefore, maximizing an estimate of the lower bound (24) ensures that the optimization algorithm finds a trustworthy answer.

6. Learning

Optimizing of a counterfactual estimate fundamentally is a learning procedure. The main purpose of this section is to demonstrate that optimizing a counterfactual expectation estimate is a sound learning principle and to outline its relation to well known learning methods for bandit and reinforcement learning problems.

6.1 A Learning Principle

We consider the simple learning setup illustrated in figure 17. Training data is collected during a single data collection experiment designed 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.

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 could have been observed during the training data collection period if the system had been using parameter θ^* all along.

Therefore we can state this problem as the optimization of the expectation Y^θ of the reward $\ell(\omega)$ with respect to the distribution $\mathrm{P}^\theta(\omega)$,

$$\max_\theta \; Y^\theta \;=\; \int_\omega \ell(\omega)\, \mathrm{P}^\theta(\omega) \;, \tag{25}$$

on the basis of a finite set of training examples $\omega_1, \ldots, \omega_n$ sampled from distribution $\mathrm{P}(\omega)$. Following section 5.5, we propose to leverage inequality (24) and, as a learning principle, to optimize an empirical estimate $\widehat{Y}^\theta$ of the lower bound $\bar{Y}^\theta$:

$$\theta^* \; = \; \arg\max_\theta \widehat{Y}^\theta \; . \tag{26}$$

We shall now discuss the statistical basis of this learning principle.[6]

6.2 Uniform Confidence Intervals

As discussed in section 4.4, inequality (24),

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

where

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

leads to confidence intervals (16) of the form

$$\forall\delta>0,\ \forall\theta \quad \mathbb{P}\Big\{\; \widehat{Y}^\theta - \epsilon_R \;\leq\; Y^\theta \;\leq\; \widehat{Y}^\theta + M(1-\widehat{W}^\theta+\xi_R) + \epsilon_R \;\Big\} \geq 1-\delta \, . \tag{27}$$

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 A.2.)

Such confidence intervals are insufficient to provide guarantees for a parameter value θ^* that depends on the sample. In fact, the optimization (26) 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 $\mathcal{F}$, applying the union bound to the ordinary intervals (27) immediately gives the uniform confidence interval :

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

- 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 $\mathcal{F}$ is infinite. For instance, appendix A.4 leverages uniform empirical Bernstein bounds (Maurer and Pontil, 2009) and obtains the uniform confidence interval

$$\mathbb{P}\Big\{\; \forall\theta\in\mathcal{F},\ \widehat{Y}^\theta - \epsilon_R \leq Y^\theta \leq \widehat{Y}^\theta + M(1-\widehat{W}^\theta+\xi_R)+\epsilon_R \;\Big\} \geq 1-\mathcal{M}(n)\,\delta \, , \tag{28}$$

6. 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 this setup results from the active system randomization during the offline data collection. See also section 6.4.

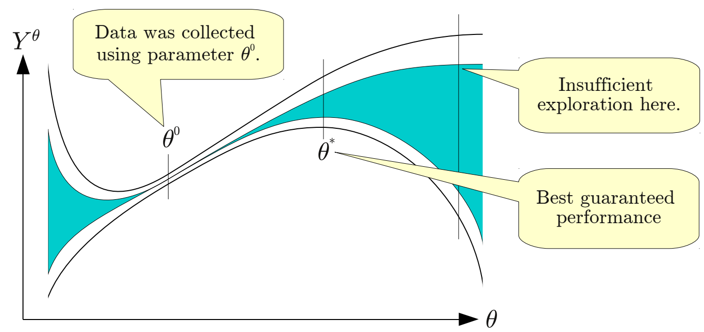

Figure 18: The uniform inner confidence interval reveals where the best guaranteed Y^θ is reached and where additional exploration is needed.

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

$$\{ \ f_\theta : \omega \mapsto \ell(\omega)\bar{w}^{\mathrm{M}}_\theta(\omega) \ , \quad g_\theta : \omega \mapsto \bar{w}^{\mathrm{M}}_\theta(\omega) \ , \quad \forall \theta \in \mathcal{F} \ \} \ . \tag{29}$$

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

The intepretation of the inner and outer confidence intervals (section 4.5) also applies to the the uniform confidence interval (28). 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 (24).

Therefore, Y^{θ^*} is close to the maximum of the lower bound of inequality (24) which essentially represents the best performance that can be guaranteed using training data sampled from $\mathrm{P}(\omega)$. Meanwhile, the upper bound reveals which values of θ could potentially offer better performance but have been insufficiently probed by the sampling distribution. Both bounds are estimated by the bounds of the inner confidence interval (figure 18).

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.

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.1), they redefine the rank-score $r_{ip}(x)$ as:

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

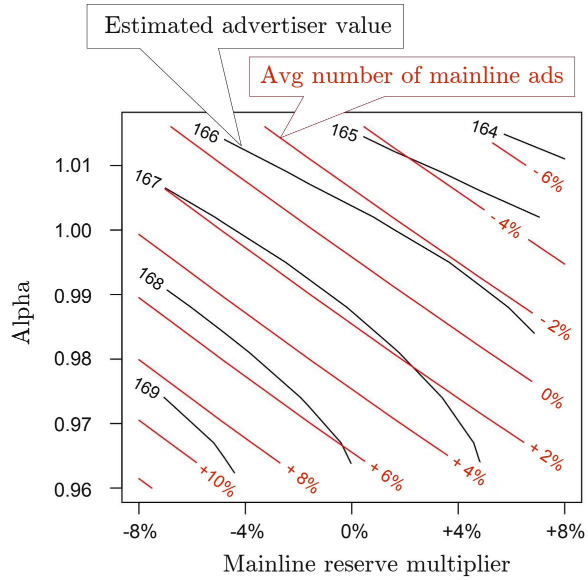

Figure 19: Level curves associated with the average number of mainline ads per page (red curves, from -6% to $+10\%$) and the average estimated advertisement value generated per page (black curves, arbitrary units ranging from 164 to 169) that would have been observed for a certain query cluster if we 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 $q_{i,p}(x)$.

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 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[7] 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

7. 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.

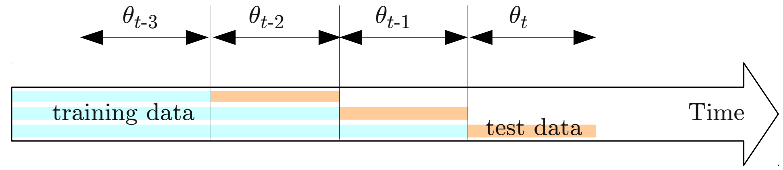

Figure 20: Sequential design – The parameter θ_t of each data collection experiment is determined using data collected during the previous experiments.

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 the off-policy derivatives of the lower bound of the counterfactual expectations of the advertiser value and of the counterfactual expectation of the number of mainline ads per page. Figure 19 shows the corresponding level curves for a particular query cluster. We did not use an upper bound for the number of mainline ads because this estimate is always much more accurate that the estimate of the value (see also figure 13). The ability to estimate these derivatives is then sufficient to run a simple optimization algorithm and determine the optimal auction tuning parameters.

The obvious alternative (see Charles and Chickering, 2012) consists in replaying the auctions with different parameters and simulate 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 peformance 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 definitive 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 18) 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 20).

Although it is relatively easy to construct convergent algorithms for the design of sequential experiments, achieving the best learning performance is notoriously difficult (e.g., 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 exploration exploitation trade-off is well understood in the case of multi-armed bandits (Gittins, 1989; Auer et al., 2002; Audibert et al., 2007; Seldin et al., 2012) because the analysis can leverage an essential property of multi-armed bandits: the outcome observed after performing a particular action brings no information about the value of other actions. In practice, 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.

Although little theoretical guidance is available in the general case, experience suggests that simple exploration heuristics perform surprisingly well. In fact, even in the simple case of multi-armed bandits, excellent empirical results have been obtained using Thompson sampling (Chapelle and Li, 2011) or using fixed exploration strategies (Vermorel and Mohri, 2005; Kuleshov and Precup, 2010). Therefore, it is often practical to simply set up each experiment by maximizing $\widehat{Y}^{\theta}$ as described in section 6.1, subject to additional ad-hoc constraints ensuring that each successive experiment guarantees 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 a same user during a thirty minute web session. On the other hand, it would be challenging 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 a third way to take feedback loop into account. Using data collected while the system was at equilibrium, we describe empirical methods to determine how an infinitesimal intervention would have displaced the equilibrium:

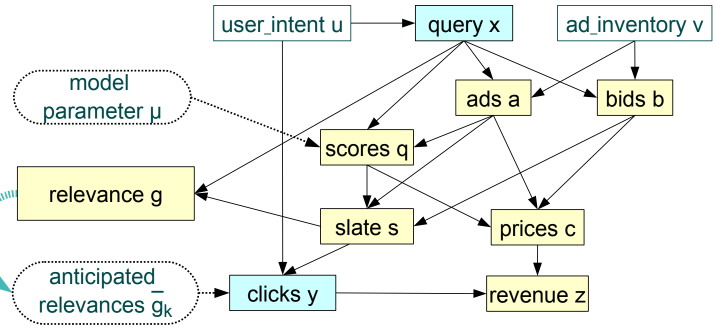

Figure 21: The new variable g is a measure of the relevance of the displayed ads. We model the user feedback loop by letting the clicks y depend on new parameters $\bar{g}_k$ representing the average relevance experienced by each user in the past.

"*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.*"

We first outline the main idea using the example of the user reactions to interventions on the ad placement system. We then describe a more sophisticated framework modeling the reactions of rational advertisers to such interventions.

7.1 Modeling User Reactions

Since displaying irrelevant advertisement messages has known negative effects, various ways to measure the ad relevance have been designed. For instance, human labelers can be asked to score the ads displayed in response to a particular query. The new variable g in figure 21 represents the measured relevance of the ad slate s displayed in response to query x.

Besides the parameter θ controlling the conditional probability distribution $P^{\theta}(s|x,a)$ that represents the scoring models, we introduce new parameters $\bar{g}_1 \ldots \bar{g}_K$ that represent the average relevances anticipated by each user on the basis of their past experience, and we assume that each of these parameters affects the conditional click probability $\mathrm{P}^{\bar{g}}(y|s,u)$ of the corresponding user. This model relies on the assumption that the relevance measure is good enough to capture how the past experience of the users affects their ad clicks.[8]

Let p_k denote the probability $\mathrm{P}\{\mathsf{user}=k\}$ that the web page is served to user k. The following counterfactual expectations functions then express the expected performance and

8. Although a poor experience can also drive users to competing web sites, modeling how the publisher loses users (e.g. because of poor relevance) or acquires new users (e.g. with marketing initiatives) is beyond the scope of this section.

relevance as a function if the user anticipated relevances.

$$
\begin{aligned}
Y_k(\theta, \bar{g}_k) &= \int_\omega \ell(\omega)\ \mathrm{P}^{\theta, \bar{g}_{\mathsf{user}}}(\omega \,|\, \mathsf{user}\!=\!k)\ , \\
Y(\theta, \bar{g}_1 \dots \bar{g}_K) &= \int_\omega \ell(\omega)\ \mathrm{P}^{\theta, \bar{g}_{\mathsf{user}}}(\omega)\ =\ \sum_k p_k\, Y_k(\theta, \bar{g}_k)\ , \\
G_k(\theta, \bar{g}_k) &= \int_\omega g(\omega)\ \mathrm{P}^{\theta, \bar{g}_{\mathsf{user}}}(\omega \,|\, \mathsf{user}\!=\!k)\ , \\
G(\theta, \bar{g}_1 \dots \bar{g}_K) &= \int_\omega g(\omega)\ \mathrm{P}^{\theta, \bar{g}_{\mathsf{user}}}(\omega)\ =\ \sum_k p_k\, G_k(\theta, \bar{g}_k)\ .
\end{aligned}
$$

Immediately after an intervention on the ad placement engine, the user anticipations $\bar{g}_k$ are incorrect because they are based on experiences that no longer represent the system performance. After a certain time, the user relevance anticipations $\bar{g}_k$ match the actual expectations $G_k(\theta, \bar{g}_k)$ and the system returns to equilibrium.

Our analysis relies on a *quasi-static* assumption: we assume that the publisher changes the parameter θ so slowly that the system remains at equilibrium at all times. Therefore, in response to an infinitesimal change $\mathrm{d}\theta$ of the scoring model parameter,[9]

$$
\forall k \quad \mathrm{d}\bar{g}_k\ =\ \mathrm{d}G_k\ =\ \frac{\partial G_k}{\partial \theta}\,\mathrm{d}\theta + \frac{\partial G_k}{\partial \bar{g}_k}\,\mathrm{d}\bar{g}_k\ =\ \frac{\partial G_k}{\partial \theta}\,\mathrm{d}\theta\ . \tag{30}
$$

We can then express the variations of the counterfactual expectation $Y(\theta, \bar{g}_1 \dots \bar{g}_K)$.

$$
\mathrm{d}Y\ =\ \frac{\partial Y}{\partial \theta}\,\mathrm{d}\theta + \sum_k \frac{\partial Y}{\partial \bar{g}_k}\,\mathrm{d}\bar{g}_k\ =\ \left(\frac{\partial Y}{\partial \theta} + \sum_k p_k \frac{\partial Y_k}{\partial \bar{g}_k} \frac{\partial G_k}{\partial \theta} \right) \mathrm{d}\theta
$$

Each partial derivative $\frac{\partial Y_k}{\partial \bar{g}_k}$ describes how the average click probability of a single user changes with his or her anticipated ad relevance. In order to permit the estimation of these derivatives, we make the additional assumption that all users respond in the same manner:

$$
\frac{\partial Y_1}{\partial \bar{g}_1} = \dots = \frac{\partial Y_K}{\partial \bar{g}_K}\ \triangleq\ \frac{\partial Y}{\partial \bar{g}}\,. \tag{31}
$$

We then obtain our final answer

$$
\mathrm{d}Y\ =\ \left(\frac{\partial Y}{\partial \theta} + \frac{\partial Y}{\partial \bar{g}} \frac{\partial G}{\partial \theta} \right) \mathrm{d}\theta\ . \tag{32}
$$

This expression describes how the expectation Y changes when the publisher applies an infinitesimal change $\mathrm{d}\theta$ to the scoring parameter θ and the users adjust their relevance anticipations $\bar{g}_k$ in response. Therefore, if we can empirically estimate the three partial derivatives appearing in equation (32), we can estimate how infinitesimal changes of the scoring model parameter θ affects the performance of the ad placement engine measured *after incurring the user reaction.*

9. The specific structure of the user feedback loop (see figure 21) ensures that the relevance g does not depend on the relevance anticipations $\bar{g}_k$. The resulting simplification $\partial G_k / \partial \bar{g}_k = 0$ facilitates the derivation but is not an essential requirement of the proposed method.

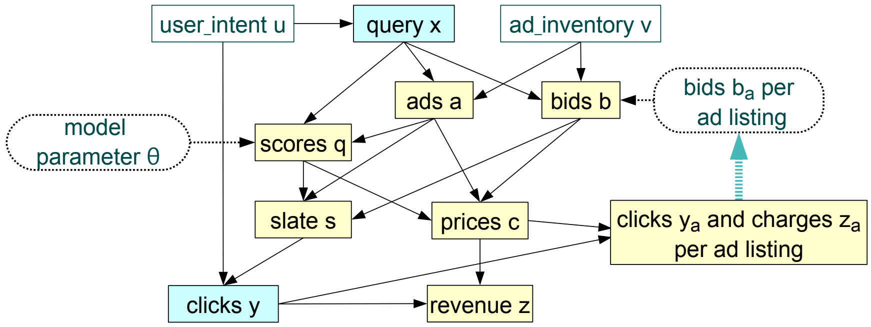

Figure 22: Advertisers select the bid amounts b_a on the basis of the past number of clicks y_a and the past prices z_a observed for the corresponding ads.

- Estimating the partial derivatives $\frac{\partial Y}{\partial \theta}$ and $\frac{\partial G}{\partial \theta}$ is a straightforward application of the policy gradient method discussed in section 5.4.
- Estimating $\frac{\partial Y}{\partial \bar{g}}$ is less direct because we cannot algorithmically select $\bar{g}$ randomly before each auction. However, we can organize a data collection experiment in two successive phases. During the priming phase, randomly selected users are exposed to ads selected by a slightly degraded version of the ad placement engine. During the second phase, all users are again treated identically. However, the ad relevance anticipations of the users selected during the priming phase still reflects the lower relevance experienced during the priming phase. Their lower click probabilities then reveal the partial derivatives of intest. This experiment also reveals how long users take to recover after being exposed to less relevant ads.

The method described in this section can of course be refined in many ways. For instance, we could consider multiple relevance signals, multiple averaging methods, and multiple clusters of users assumed to respond identically to relevance changes. The principle remains the same.

7.2 Modeling 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.

The previous section relies on assumption (31) and on ad-hoc experiments to determine how users react to changes in anticipated relevance. This section assumes that the advertisers are *rational* and therefore always maximize their economic interests. Although there are more realistic ways to model advertisers, this exercise is interesting because the same assumption underlies auction theory (see section 2.1). This approach therefore provides a framework to seamlessly integrate auction theory and machine learning.

As illustrated in figure 22, 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. Each variable y_a in the structural equation model represents the number of clicks received by ads associated with bid b_a. Each variable z_a represents the amount charged for these clicks to the corresponding advertiser.

Rational advertisers seek to maximize their surplus, that is, the difference between the value they see in the clicks and the price they pay to the publisher (figure 23). Therefore, each advertiser selects bids b_a according to their anticipated impact on the number of resulting clicks y_a and on their cost z_a.

Let V_a denote the value of a click for the advertiser. Following the pattern of the perfect information assumption (see section 2.1), we assume that the advertisers eventually acquire full knowledge of the expectations

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

and reach a Nash equilibrium

$$\forall a \quad b_a \in \operatorname*{Arg\,Max}_b \ U_a^\theta(b_1, \ldots, b_{a-1}, b, b_{a+1}, \ldots, b_A) \ . \tag{33}$$

with utility functions

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

The existence of such a Nash equilibrium is not obvious.[10] However, we do not strictly need to establish that this equilibrium exists for any combination of advertiser values V_a. Since the true values are unknown, we shall use data collected when the system is stationary to estimate advertiser values V_a that are consistent with a Nash equilibrium. We shall then estimate how a small change of the model parameters θ displaces this posited equilibrium.

The injection of smooth random noise into the auction mechanism changes the discrete problem into a continous problem amenable to well known differential methods. Therefore we assume that the densities $\mathrm{P}^{b_\star}(b|x,v)$ and $\mathrm{P}^\theta(q|x,a)$ are smooth enough to ensure that the expectations Y_a and Z_a are continuously differentiable functions of the parameters $b_\star$ and θ. The equilibrium (33) then satisfies the necessary Kuhn-Tucker conditions

$$\forall a \qquad V_a \, \frac{\partial Y_a}{\partial b_a} - \frac{\partial Z_a}{\partial b_a} \ \begin{cases} \leq 0 & \text{if } b_a = 0, \\ \geq 0 & \text{if } b_a = b_{\max}, \\ = 0 & \text{if } 0 < b_a < b_{\max}. \end{cases} \tag{34}$$

If the corresponding ad is displayed with sufficient frequency, we can estimate the value of the partial derivatives appearing in (34) by randomizing the bids and computing policy gradient as explained in section 5.4.

However, 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. Fortunately, 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

10. Subject to certain assumptions on the utility functions, classic results ensure the existence of a Nash equilibrium for arbitrary values V_a. For instance, if we assume that the marginal click prices increase with the click volume, the pricing curves are convex (figure 23), the utilities $U_a^\theta(b)$ are diagonally quasiconcave, and Friedman's theorem (1977) establishes the existence of a Nash equilibrium.

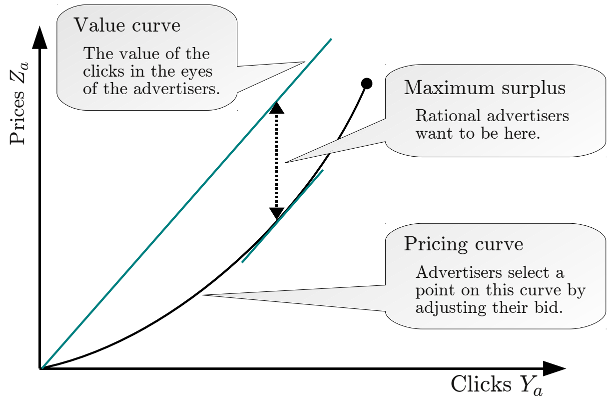

Figure 23: Advertisers control the expected number of clicks Y_a and expected prices Z_a by adjusting their bids b_a. Rational advertisers select bids that maximize the difference between the value they see in the clicks and the price they pay.

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 collect data as if the multiplier was applied to the bid, and charge prices computed as if the multiplier was applied to the estimated click probabilities.

We can then estimate the advertiser values V_a by solving the equilibrium equations. 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 the value of a click 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 (34). 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.

Let $\mathcal{A}$ 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 (34):

$$\forall a' \in \mathcal{A}, \quad 0 \;=\; \left(V_{a'} \frac{\partial^2 Y_{a'}}{\partial b_{a'}\,\partial\theta} - \frac{\partial^2 Z_{a'}}{\partial b_{a'}\,\partial\theta} \right) \mathrm{d}\theta + \sum_{a\in\mathcal{A}} \left(V_{a'} \frac{\partial^2 Y_{a'}}{\partial b_{a'}\,\partial b_a} - \frac{\partial^2 Z_{a'}}{\partial b_{a'}\,\partial b_a} \right) \mathrm{d}b_a \;. \qquad (35)$$

The partial second derivatives must be estimated as described in section 5.3.

Solving the system (35) yields expressions of the form

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

We can then 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 $\mathcal{A}$ rationally adjust their bids b_a in response:

$$\mathrm{d}Y = \left(\frac{\partial Y}{\partial\theta} + \sum_a \Xi_a \frac{\partial Y}{\partial b_a} \right) \mathrm{d}\theta \;. \qquad (36)$$

Such derivatives can of course drive learning algorithms (section 6).

Although we only can estimate the reaction of the active advertisers $\mathcal{A}$, expression (36) provides a useful and well characterized point of reference. We know that it does not include the potentially positive reaction of advertisers who did not bid but could have. We also know that advertisers placing unrealistically high bids are modeled pessimistically because we only can estimate a lower bound of their values.

To alleviate these issues, we could alter the auction mechanism in ways that force these advertisers to reveal more information. 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 quasi-static approach provides a generic framework to take such aspects into account.

7.3 Dealing with multiple feedback loops

Using the quasi-static methodology familiar to physicists, we have described how estimate the derivatives of counterfactual expectations that describe the system performance when it reaches an equilibrium induced by a causal feedback loop. A natural extension of this approach handles multiple simultaneous feedback loops: we simply write the derivatives of all the equilibrium equations and solve the resulting linear system. This flexibility provides countless refinement opportunities.

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, thanks to a sharp distinction between the learning algorithms and the extraction of the signals that drive them, these methods are applied to causal models with different structures, offering, for instance, a fresh viewpoint on known reinforcement learning algorithms. Second, maybe unsurprisingly, the mathematical and philosophical tools developed for the analysis of physical systems appear very effective for the analysis of causal information system and of their equilibria. With such themes, this work is also a vindication of cybernetics (Wiener, 1948).

Acknowledgements

We would like to acknowledge extensive discussions with our colleagues Susan Athey, John Langford, Lihong Li, and Chris Meek. We also would like to thank the Microsoft adCenter RnR department for giving us the precious opportunity to try these ideas at scale.

Appendix

A.1 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

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

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 $\mathcal{I}_s = \{i : S_i = s\}$ gathering the ads belonging to the same advertiser s. The ads that maximize the product $b_i\beta_i(x)$ within set $\mathcal{I}_s$ 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 $\mathcal{R}_L$ 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 $\mathcal{R}_L$.

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

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

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

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

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

It is easy to see that I^*_k intersects $\{i_k,\dots,i_N\}$ because, were it not the case, replacing i_k by any element of I^*_k would increase $\mathcal{R}_L$ 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 $\mathcal{R}_L$ without violating any of the constraints.

Therefore, if the inner maximization problem admits a solution, we can compute a solution by recursively picking $i_1,\dots,i_N$ from $I^*_1, I^*_2, \dots, I^*_N$. 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).

A.2 Confidence Intervals

Section 4.4 explains how to obtain improved confidence intervals by replacing the unbiased importance sampling estimator (9) by the capped importance sampling estimator (11). This appendix provides details that could have obscured the main message.

A.2.1 Outer confidence interval

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

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

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

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

with

$$\epsilon_R \ = \ \mathrm{erf}^{-1}(1-\delta)\ \sqrt{2\,\bar{V}}\ . \tag{37}$$

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 (37) 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 $1 - \delta$,

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

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 $\bar{w}(\omega)$.

$$\mathbb{P}\Big\{\; \widehat{Y}^* - \epsilon_R \leq \bar{Y}^* \leq \widehat{Y}^* + \epsilon_R \;\Big\} \geq 1 - 2\delta$$

where

$$\epsilon_R \;=\; \sqrt{\frac{2\,\widehat{V} \log(2/\delta)}{n}} + M\,R\,\frac{7 \log(2/\delta)}{3(n-1)}. \tag{38}$$

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 the the Bernstein bound, they remains reasonably competitive, and they provide much stronger guarantee.

A.2.2 Inner confidence interval

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

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

The constant M is defined by assumption (10). 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 capped weights

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

Replacing in inequality (14) gives the outer confidence interval

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

with

$$\xi_R \;=\; \sqrt{\frac{2\,\widehat{V}_w\log(2/\delta)}{n}} + R\,\frac{7\log(2/\delta)}{3(n-1)}\;. \tag{39}$$

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

Putting together the inner and outer confidence intervals,

$$\mathbb{P}\Big\{\;\widehat{Y}^*-\epsilon_R \leq Y^* \leq \widehat{Y}^* + M(1-\widehat{W}^*+\xi_R)+\epsilon_R\;\Big\} \geq 1-3\delta\,, \tag{40}$$

with ϵ_R and ξ_R computed as described in expressions (38) and (39).

A.3 Using Invariant Variables and Predictors

This appendix describes useful techniques leveraging invariant variables and invariant predictors to construct better confidence intervals.

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 $\mathrm{var}[\ell(\omega)]$.

A.3.1 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 (10), we can use bounds that depend on invariant variables:

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

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

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

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

$$\mathcal{B}_{\text{lo}} \ \le \ Y^* - \bar{Y}^* \ \le \ \mathcal{B}_{\text{hi}}$$

With the notations

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

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

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

two applications of theorem 1 give the inner confidence interval:

$$\mathbb{P}\Big\{\ \bar{Y}^* + \widehat{\mathcal{B}}_{\text{lo}} - \xi_{\text{lo}} \ \le \ Y^* \ \le \ \bar{Y}^* + \widehat{\mathcal{B}}_{\text{hi}} + \xi_{\text{hi}}\ \Big\} \ \ge \ 1 - 2\delta \ .$$

A.3.2 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 (section 5.2):

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

Let us define the reweigthing ratios $w^+(\omega) = \mathrm{P}^+(\omega)/\mathrm{P}(\omega)$ and $w^*(\omega) = \mathrm{P}^*(\omega)/\mathrm{P}(\omega)$, their capped variants $\bar{w}^+(\omega)$ and $\bar{w}^*(\omega)$, and the capped 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 A.2.1 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

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

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

$$\forall\omega \qquad -M \ \le \ -\zeta(\upsilon) \ \le \ \ell(\omega) - \zeta(\upsilon) \ \le \ M - \zeta(\upsilon) \ \le \ M \ .$$

This can be achieved as shown in section A.3.1.

A.4 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 (28).

The first step is to characterize the size of a family $\mathcal{F}$ of functions mapping a space $\mathcal{X}$ into the interval $[a, b] \subset \mathbb{R}$. Given n points $\mathbf{x} = (x_1 \dots x_n) \in \mathcal{X}^n$, the trace $\mathcal{F}(\mathbf{x}) \in \mathbb{R}^n$ is the set of vectors $(f(x_1), \dots, f(x_n))$ for all functions $f \in \mathcal{F}$.

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

$$\forall v \in \mathcal{F}(\mathbf{x}) \quad \exists c \in C \qquad \max_{i=1\dots n} |v_i - c_i| \leq \varepsilon \ ,$$

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

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

Thanks to a famous combinatorial lemma (Vapnik and Chervonenkis, 1968, 1971; Sauer, 1972), for many usual parametric families $\mathcal{F}$, the growth function $\mathcal{N}(n, \varepsilon, \mathcal{F})$ increases at most polynomially[11] 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 $X, X_1, \dots, X_n$ be i.i.d. random variables with values in $\mathcal{X}$. Let $\mathcal{F}$ be a set of functions mapping $\mathcal{X}$ into $[a, b] \subset \mathbb{R}$ and let $\mathcal{M}(n) = 10\,\mathcal{N}(2n, \mathcal{F}, 1/n)$. Then we probability at least $1 - \delta$,

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

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 $\mathcal{M}(n) = 10\,\mathcal{N}(2n, \mathcal{F}, 1/n)$ grows polynomially with n.

Let us then define the family of functions

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

and use the uniform empirical Bernstein bound to derive an outer inequality similar to (38) and an inner inequality similar to (39). The theorem implies that, with probability $1 - \delta$, both inequalities are simultaneously true for all values of the parameter θ. The uniform confidence interval (28) then follows directly.

11. 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 $(x_i, S_k) \mapsto \mathbb{1}\{f(x_i) \in S_k\}$.

A.5 Bibliographical Notes

This appendix expands the references discussed in the main text.

A.5.1 Notes for section 2

The presentation of the ad placement problem summarizes experience acquired at Microsoft adCenter in 2010 and 2011. Generalized second order auctions for position auctions were discussed by Varian (2007) and Edelman et al. (2007). Simpson's paradox (Simpson, 1951) has been abundantly discussed, for instance by Pearl (2000). The example of Simpson's paradox in ad placement has not been previously published.

A.5.2 Notes for section 3

Despite burning philosophical debates, the manipulability theory of causation (e.g., von Wright, 1971; Woodward, 2005) has gained acceptance in the statistical community. Rubin (1986) argues that there is "*no causation without manipulation*". Viewing causation as a reasoning model (Bottou, 2011) gives statistical meaning to causal statements that do not lend themselves to manipulation.

Structural equation models (Wright, 1921) replicate the methods of classical physics developed and rationalized during the age of enlightenment. Reichenbach (1956) gives a thorough discussion of causation in physics. Wiener (1948) demonstrates how the methods of physics can be applied to the treatment of information in general, even when this information does not describe physical quantities. The discussion of the isolation assumption in section 3.2 is clearly inspired by this connection.

Pearl (2009) gives a clear and concise review of the connection between structural equation models and probabilistic models and of its implications for causal inference. Pearl (2000) contains considerably more details and references. Spirtes et al. (1993) focus on the discovery of causal relations.

A.5.3 Notes for section 4

Many of the estimation techniques described in sections 4 and 5 have known counterparts in the special cases of reinforcement learning, contextual bandits, and multi-armed bandits.

Monte-Carlo methods for reinforcement learning (Sutton and Barto, 1998, chapter 5) are essentially reweighted counterfactual estimates. Reinforcement learning research traditionally focuses on control problems with relatively small state spaces and long sequences of observations. This reduces the need for characterizing exploration with tight confidence intervals. For instance, Sutton and Barto suggest to normalize the estimator (7) by $1/\sum_i w(\omega_i)$ instead of $1/n$. This works poorly when parts of the state space are left unexplored.

A series of papers on the evaluation of contextual bandits algorithms appeared while our ad placement work was under way. Strehl et al. (2010) describe a capped estimator for contextual bandits, unfortunately in a setup that subtly assumes the absence of unknown confounding variables. Li et al. (2011) evaluate contextual bandit policies using randomization and reweighted estimates.

A.5.4 Notes for section 5

Hesterberg (1988) describes variance reduction methods for importance sampling. These methods are designed to improve the confidence interval of the normal average by constructing a sampling distribution that reduces the variance. Our situation is less favorable because we want to use a same sampling distribution to obtain estimates for many counterfactual distributions. Dudík et al. (2011) propose a variance reduction method that extends the invariant predictor method to any predictor whose conditional expectation $\zeta^*(\upsilon) = \int_{\omega\setminus\upsilon} \zeta(\omega)\,\mathrm{P}^*(\omega\,|\,\upsilon)$ can be computed efficiently.

The policy gradient estimation method has been rediscovered multiple times (Aleksandrov et al., 1968; Glynn, 1987; Williams, 1992) in slightly different contexts. Variance reduction with a baseline has been analyzed by Greensmith et al. (2002).

A.5.5 Notes for section 6

The study of learning principles using uniform confidence intervals has been pioneered by Vapnik and Chervonenkis (1968) and has been the object of numerous developments in statistics and in machine learning.

Squashing the click probability estimates in ad placement has been described by Lahaie and McAfee (2011). The general auction tuning scheme is due to Charles and Chickering (2012).

Both Wald (1945) and Robbins (1952) give an overview of early works on sequential design. Robbins formulates the two-armed bandit problem as one of the simplest instance of a problem involving the explore/exploit trade-off. The k-armed bandit problem has received an elegant Bayesian solution (Gittins, 1989). Computationally efficient algorithms verify regret bounds that grow at the optimal rate (Auer et al., 2002; Audibert et al., 2007; Seldin et al., 2012) but are often outperformed by algorithms relying on Gittins indices or on the Thompson sampling heuristic (Chapelle and Li, 2011). Simple heuristics perform suprisingly well (Vermorel and Mohri, 2005; Kuleshov and Precup, 2010). The design of computationally efficient algorithms to optimally balance exploration and exploitation in arbitrarily complex learning systems is still considered an open problem.

A.5.6 Notes for section 7

Although the differential analysis of equilibrium has been used for centuries in classical mechanics, we are not aware of previous uses of this technique to deal with long-term feedback in learning systems.

References

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

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.

Léon Bottou. From machine learning to machine reasoning. arXiv: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.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proc. 28th International Conference on Machine Learning (ICML 2011)*, pages 1097–1104. Omnipress, 2011.

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.

James W. Friedman. *Oligopoly and the Theory of Games*. North-Holland, 1977.

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.

Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. In *Advances in Neural Information Processing Systems 14*, pages 1507–1514. MIT Press, 2002.

David J. Hand and David J. Weston. Statistical techniques for fraud detection, prevention, and assessment. In *Mining Massive Data Sets for Security*, pages 257–270. IOS Press, Amsterdam, 2008.

Tim Hesterberg. *Advances in Importance Sampling*. PhD thesis, Stanford University, 1988.

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.

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.

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

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.

Hans Reichenbach. *The Direction of Time*. University of California Press, Berkeley, 1956.

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

Donald B. Rubin. Comment: Which ifs have causal answers. *Journal of the Americal Statistical Association*, 81(396):961–962, 1986.

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

Yevgeny Seldin, Nicolò Cesa-Bianchi, Peter Auer, François Laviolette, and John Shawe-Taylor. PAC-Bayesian inequality for martingales and its application to multiarmed bandits. *JMLR Workshop and Conference Proceedings*, 26:98–111, 2012.

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

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction and Search*. Springer, 1993. 2nd edition: MIT Press, 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.

Alex Strehl, John Langford, Lihong Li, and Sham Kakade. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems 23*, pages 2217–2225. NIPS Foundation, 2010.

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.* The Technology Press, Cambridge, MA, 1948. 2nd Edition: MIT Press, 1965.

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.