LOSS FUNCTIONS FOR PREDICTED CLICK-THROUGH RATES IN AUCTIONS FOR ONLINE ADVERTISING

PATRICK HUMMEL* R. PRESTON MCAFEE*

OCTOBER 11, 2013

ABSTRACT. We consider the problem of the optimal loss functions for predicted click-through rates in auctions for online advertising. While standard loss functions such as mean squared error or the log likelihood loss function severely penalize large mispredictions while imposing little penalty on smaller mistakes, we find that a loss function reflecting the true underlying economic loss resulting from mispredictions would impose significant penalties for small mispredictions while only imposing slightly larger penalties on large mispredictions. We illustrate that using such a loss function can significantly improve economic efficiency and revenue from online auctions if one is trying to fit a model that is misspecified even when one has an arbitrarily large amount of training data.

Keywords: Loss functions; Predicted click-through rates; Auctions; Online advertising

1. INTRODUCTION

A loss function represents the loss incurred from making an error in estimation. There is widespread consensus that the choice of loss function should reflect the actual costs of misestimation. For example, Moyé (2006) writes "The Bayesian constructs the loss function to reflect the true consequences of correct and incorrect decisions." In practice, however, the quadratic loss (mean squared error) and log likelihood loss appear to dominate applications, with other loss functions such as hinge loss and the linex loss (Varian (1974)) as distant thirds. For none of these loss functions is the selection closely guided by the application. This paper develops loss functions for predicted click-through rates in Internet advertisement auctions that reflect the true consequences of estimation errors.

*Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA.

Internet advertisement auctions are used to place advertisements on the results pages of Google, Bing and Yahoo. In addition, many web publishers, such as CNN, the New York Times, and Fox Interactive, sell a portion of their advertising inventory through online auctions, run by Google (AdX and AdSense), Yahoo (Right Media), Facebook (FBX), AppNexus, OpenX and others. While many web publishers still rely on negotiated contracts for the majority of their revenue, Internet auctions are growing rapidly as a mechanism for placing advertisements on web pages.

In a majority of these Internet advertisement auctions, advertisers pay only if a user clicks on the advertisement, which is known as cost-per-click (CPC) pricing, in contrast to cost-per-one-thousand-impressions (CPM) advertising. In order to identify the highest revenue advertisement, it is necessary to forecast the probability that a visitor on a web page will click on the ad. For example, one ad might have a CPC bid of \$2, which is the amount the advertiser pays if the user clicks on the ad, while another ad might have a CPC bid of \$4. The ad with a \$2 bid is more valuable if this ad is at least twice as likely to receive a click as the ad with a \$4 bid. The product of a CPC bid and the probability of a click is known as the expected-cost-per-mille (eCPM), which represents the estimated value of the impression (in thousands).

The forecasts of click probabilities are created using enormous machine learning systems involving potentially hundreds of millions or even billions of variables.[1] The proliferation of variables arises because click rates for a 'woman on a celebrity site with a minivan ad' might differ substantially from 'woman on a celebrity site with an economy car ad' or 'woman on the Wall Street Journal site with a minivan ad.' Intersections of ad characteristics, user characteristics and website characteristics produce a plethora of potentially relevant variables. Moreover, a history involving billions of data points per day on users visiting pages with advertisements is available (McMahan *et al.* (2013)). The standard technique

[1]For instance, in Yahoo!'s description of their display advertising inventory allocation optimization, they write "Scalability of the optimizer is a huge challenge because the problems being solved in real applications are typically of large scale, involving tens to hundreds of millions of variables and hundreds of thousands to millions of constraints" and McMahan *et al.* (2013) develops techniques for forecasting click-through rates using models with billions of different coefficients. See http://labs.yahoo.com/project/display-inventory-allocation-optimization/ for further details on Yahoo!'s display advertising inventory allocation optimization.

is a logistic regression (Hilbe (2009)) which is typically estimated by maximizing the log of the likelihood function.[2] It is this maximization that we propose to change by tailoring the objective function to match the economic losses that result from misestimates in the cases where the estimates are employed.[3]

There are two main reasons for using a correctly specified loss function: to improve small sample performance, and to improve performance under misspecification. The optimal loss function is the same for both purposes; it is the loss function that correctly measures the actual losses. In assessing the importance of using the correct loss function, we focus on misspecification for the following reason. In most settings employing machine learning, there is an enormous amount of data. In search advertising, our primary application, there can be trillions of observations. The large number of observations ensures that one can accurately estimate the coefficients of the most common variables that are the most critical in optimizing the objective. While there may be some variables that only rarely appear in the data, misestimates on these variables will not matter much because they occur so rarely. Thus small sample performance problems should be significantly mitigated by the large number of observations in advertising auctions.

By contrast, when the model that one is trying to fit is misspecified, no amount of data eliminates the importance of the loss function, because even in the limit of an arbitrarily large amount of data, predictions and truth will be distinct. Misspecification is almost surely important in the advertising auction framework because essentially no attention has been paid to specification and it would be nearly impossible for anyone to perfectly specify a model with such a large number of explanatory variables. Instead, the approach is one of over-inclusiveness of extant variables, with the expectation that the system will eventually learn the correct coefficients. Problems of endogeneity, correlation of explanatory variables with the error terms, missing interaction terms, and other misspecification errors are necessarily

[2]If p denotes the actual click-through rate of an ad and q denotes the predicted click-through rate, then this log likelihood loss function is given by $p \log q + (1-p)\log(1-q)$.

[3]Another potentially interesting decision problem is whether to show an ad at all, taking into account the fact that low quality ads may lead to poor user experience. Though using appropriate loss functions may also be important in this setting, in this paper we simply focus on the issue of what loss functions to use in deciding which ad to show.

ignored given the high dimensionality of the problems. Since the models used in online auctions are likely to be misspecified, choosing the loss function to match the actual economic losses from misestimates is likely to be important despite the enormous amount of data that can be used to fit the models.

We begin by characterizing the economic efficiency loss incurred from a misestimate. When we overestimate the probability of a click, also known as the click-through rate or CTR, we will sometimes run an ad that is not best, thereby incurring an loss in economic welfare. Similarly, when we underestimate the CTR, we may fail to run the best ad, opting instead for an alternative. In both cases, the welfare loss is determined by whether alternative ads are close, and as a result, the distribution of alternatives plays a key role in the analysis. In ad auctions, the distribution of alternatives is often approximated by a lognormal distribution, but there is no theoretical or practical impediment to using the empirical distribution of alternatives. Because of its close connection to the empirical distribution of alternatives, we call our construct the *empirical loss function*.

One prominent feature of using the empirical loss function is that misestimates outside the range of the data on alternatives incur a small marginal penalty. Suppose, for example, that most eCPMs drawn from the distribution of alternatives fall below 2¢. If an advertiser then makes a CPC bid of \$1, there is little difference between the likely auction outcomes that result from predicting a CTR of 0.02 or predicting a CTR of 0.03, as the CPC advertiser will win almost every auction regardless of which of these predicted click-through rates is used. Thus, in contrast to the log likelihood or mean squared error loss functions, which penalize larger losses at an increasing rate, the empirical loss function only imposes slightly larger penalties on mispredictions beyond a certain level of inaccuracy.[4]

There are two significant properties of both the mean squared error and the log likelihood losses. Both are unbiased (also called calibrated), in the sense that they are minimized by predicting a click-through rate equal to the actual CTR, and they are convex, which ensures that an iterative process like Newton's Method will find the minimum loss in a

[4]This feature of the empirical loss function is also shared by other robust statistics such as Tukey's biweight (Maronna *et al.* (2006)).

straightforward, hill-climbing way. The empirical loss function is similarly unbiased but it is not convex. Because of the scale of the click-through rate estimation problem, it would be exceedingly difficult to optimize non-convex loss functions in practice. For this reason, we construct a best convex loss function based on the empirical loss function that can be more easily used in practice.

Finally we investigate whether using the loss functions that we derive can improve the economic performance of online auctions when the model that one is trying to fit is misspecified. As we have noted previously, when the model is misspecified, even with an arbitrarily large amount of data, predictions and truth will be distinct, and the choice of loss function may matter for economic welfare. We illustrate that using the empirical loss function rather than standard loss functions such as mean squared error or log likelihood can have a substantial effect on the economic losses that result when one has an arbitrarily large amount of training data through some simple simulations on misspecified models.

Our paper relates to two distinct strands of literature. First, our paper relates to the literature on loss functions. There is an extensive literature in economics, machine learning, and statistics that addresses questions related to using alternative loss functions (see *e.g.* Arrow (1959), Altun *et al.* (2003), Bartlett *et al.* (2006), Manski (2004), Reid and Williamson (2010, 2011), Skalak *et al.* (2007), Steinwart (2007), and Zhang (2004)). While some of these papers are motivated by specific applications, such as natural language processing or a utilitarian social planner seeking to apply treatment rules for heterogeneous populations, these papers primarily focus on questions related to loss functions in general statistical settings, and none of these papers considers the question of the best loss function to use in the context of online auctions for advertising.

Our work also relates to work within economics on dealing with misspecified models. Within this literature, White (1982) has considered the problem of maximum likelihood estimation of misspecified models, Sawa (1978) and Vuong (1989) present methodology for selecting amongst several misspecified models, and Manski (2006, 2009) presents methodology for partial identification of models that results in set-valued estimates rather than point

estimates. We instead focus on the problem of choosing the best possible loss function for misspecified models in online auctions.

More broadly our paper relates to literature on the design and performance of systems for auctions for online advertising. This literature has considered questions such as mechanisms for selling advertising in position auctions (Aggarwal *et al.* (2006, 2008), Athey and Ellison (2011), Edelman and Ostrovsky (2007), Edelman *et al.* (2007), Edelman and Schwarz (2010), Even-Dar *et al.* (2008), Gonen and Vassilvitskii (2008), Ostrovsky and Schwarz (2009), and Varian (2007)), mechanisms for re-ranking advertisers using non-standard ranking methods (Bax *et al.* (2012), Lahaie and McAfee (2007), and Lahaie and Pennock (2011)), ways to optimally explore the advertisers in an auction (Li *et al.* (2010)), and mechanisms for ensuring that the advertisers obtain a representative allocation of display advertising opportunities (Beck and Milgrom (2012) and McAfee *et al.* (2013)). However, none of these papers has considered questions related to appropriate loss functions in auctions for online advertising. Our paper fills this gap in the literature.

2. The Model

In each auction, a machine learning system must predict a click-through rate for an advertiser who has submitted a CPC bid into an auction for an advertising opportunity where advertisers are ranked according to their eCPM bids. Thus if this CPC bidder submits a bid per click that is equal to b and the machine learning system predicts that the click-through rate of this ad is q, then the eCPM bid for this advertiser that is entered into the auction is equal to bq.

While the machine learning system predicts a click-through rate of the ad that is equal to q, the actual click-through rate of the ad may be different from the predicted click-through rate. We let p denote the actual click-through rate of the CPC ad in question. We also assume that in each auction, the highest competing eCPM bid that this advertiser faces is a random draw from the cumulative distribution function $G(\cdot|b)$ with corresponding probability density function $g(\cdot|b)$, where the dependence of $G(\cdot|b)$ on b indicates that we explicitly allow for the

possibility that the highest competing eCPM bid may be correlated with b. We also let A denote this highest competing eCPM bid that the advertiser faces.

3. Preliminaries

We first address the question of the appropriate choice of a loss function when the goal of the machine learning system is to predict click-through rates that would maximize total economic efficiency in the auction. Throughout when we refer to efficiency we mean the resulting economic surplus or economic welfare rather than the statistical efficiency of the estimator in terms of how quickly the estimator converges. For the application we consider, we are primarily concerned with how well our estimates perform in misspecified models with infinite training samples rather than any finite sample properties, so it makes sense to focus on economic welfare in the limit of an arbitrarily large amount of training data rather than statistical efficiency of the estimator.

Given our estimate of the predicted click-through rate of the CPC ad, we will select the ad in question when $bq \geq A$, and we will select the ad with the highest competing eCPM bid otherwise. In the first case, the expected total welfare generated is bp, where p denotes the actual click-through rate of the CPC ad, but in the second case the total welfare generated is A. Thus the expected amount of welfare generated when the machine learning system predicts a click-through rate of q is $bpPr(bq \geq A) + E[A|bq < A]Pr(bq < A)$.[5]

To derive an appropriate loss function, it is solely necessary to compare the expected amount of welfare generated when the machine learning system predicts a click-through rate of q with the total expected welfare that would be generated when the machine learning system predicts the click-through rates of the ads perfectly. We refer to this loss function as the *empirical loss function* because it reflects the true empirical loss that results from

[5]If the highest competing ad is a CPC bidder with an uncertain click-through rate, then the actual amount of welfare generated upon showing such an ad may be different from A. However, as long as the predicted click-through rates are unbiased on average, the expected amount of welfare that would arise in this case would be A, and all the analysis in our paper would continue to hold.

misestimates of the predicted click-through rates. We derive the appropriate empirical loss function for machine learning for advertising in Theorem 1:

Theorem 1. *The loss function that maximizes total economic efficiency for a machine learning system from predicting that an ad has a click-through rate of q when this ad actually has a click-through rate of p is $\int_{bp}^{bq}(bp - A)g(A|b)\ dA$.*

While this loss function may appear to be substantially different from standard loss functions that are used in statistics, it is worth noting that minimizing the magnitude of this loss function will be equivalent to selecting the maximum likelihood estimator when the errors are distributed according to some carefully chosen distribution. Much like minimizing mean-squared is equivalent to choosing the maximum likelihood estimator when the likelihood of a given error is distributed according to a normal distribution, or proportional to $e^{-(q-p)^2/2}$, optimizing the empirical loss function is equivalent to choosing the maximum likelihood estimator when the likelihood of a given error is proportional to $e^{\int_{bp}^{bq}(bp-A)g(A|b)\ dA}$. Thus known properties of maximum likelihood estimators will extend to the estimators that are derived by minimizing the magnitude of the empirical loss function, including both finite sample and asymptotic properties.

As we have noted in the introduction, in most machine learning systems it is standard to use simple loss functions such as mean squared error or the log likelihood loss function. Given the result in Theorem 1, it is natural to ask whether these standard loss functions are indeed compatible with the empirical loss function that we have identified as reflecting the actual economic efficiency loss that results from making an inaccurate prediction. Our next result illustrates that it is indeed theoretically possible to come up with distributions of the highest competing eCPM bids such that the empirical loss function given in Theorem 1 will be compatible with one of these standard loss functions.

Theorem 2. *Suppose that the highest competing eCPM bid is drawn from a uniform distribution that is independent of b. Then the empirical loss function in Theorem 1 is equivalent to the mean squared error loss function.*

While minimizing the empirical loss function in Theorem 1 is equivalent to minimizing mean squared error when the highest competing eCPM bid is drawn from a uniform distribution, it is worth noting that this result will not extend under other distributions. Empirically the uniform distribution is unlikely to be a good representation of the distribution of highest competing eCPM bids, as these distributions are generally thought to be better matched by a log-normal distribution. This holds, for example, in the context of bidding in sponsored search auctions on Yahoo!, where both Ostrovsky and Schwarz (2009) and Lahaie and McAfee (2011) have noted that these distributions can be modeled well by a log-normal distribution.

Under more realistic distributions of the highest competing eCPM bid such as the log-normal distribution, it is no longer the case that the empirical loss function in Theorem 1 will be equivalent to minimizing either mean squared error or the log likelihood loss function. This can be readily seen in Figure 1, where we plot mean squared error (in a solid black line), the log of the likelihood function (in long red dotted lines), and the empirical loss function (in short blue dotted lines) that results from predicting a click-through rate of q when the actual click-through rate of the ad is $p = 0.019$ and the highest competing eCPM bid is drawn from a lognormal distribution with parameters μ and σ^2. All of these loss functions have been normalized in such a way that the biggest loss that ever results in the range considered is -1.

This figure indicates that the shape of the empirical loss function differs dramatically from the shapes of the loss functions represented by mean squared error and the log of the likelihood function when the highest competing eCPM bid is drawn from a log-normal distribution. Both mean squared error and the log of the likelihood function impose significantly larger penalties for predictions that are far off the mark than they do for predictions that are only somewhat inaccurate. However, the empirical loss function imposes nearly identical losses for predictions that are off by more than a certain amount.

The reason for this is as follows. If a prediction is off by more than a certain amount, additional errors in the prediction are unlikely to affect the outcome of the auction. If one

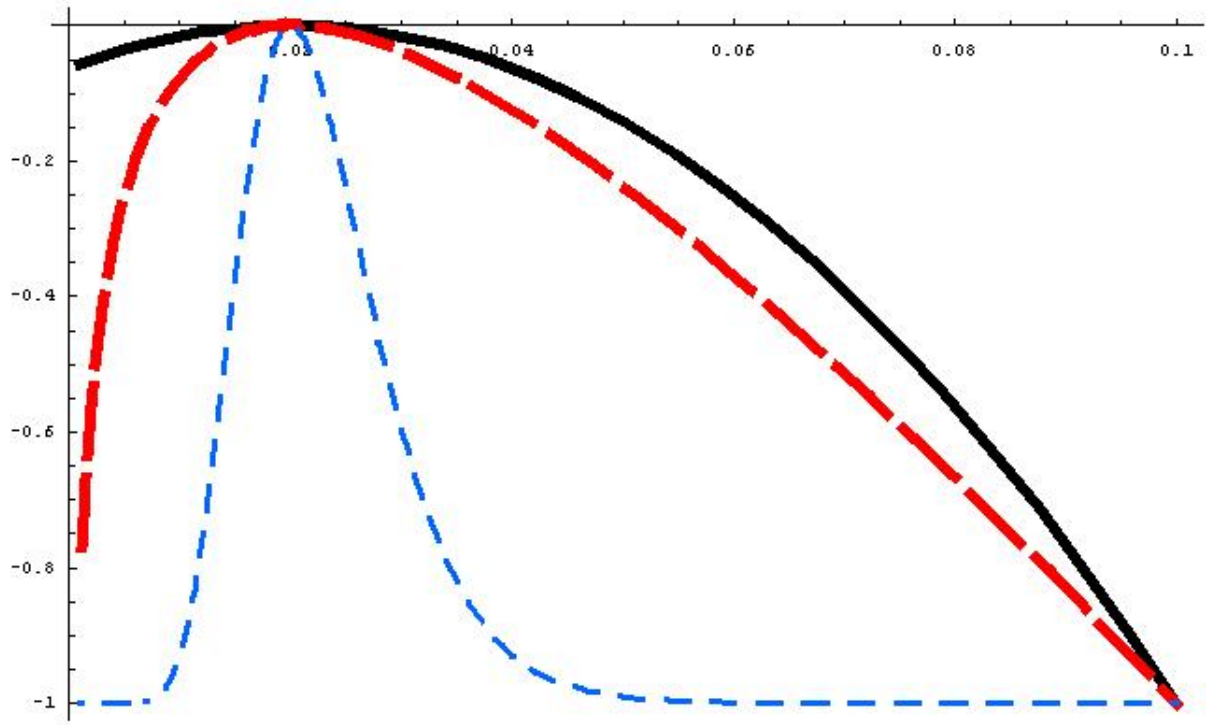

FIGURE 1. In this figure, the blue curve (short dotted lines) represents the shape of the empirical loss function, the red curve (long dotted lines) represents the shape of the log likelihood loss function, and the black curve (solid line) represents the shape of the mean squared error loss function. The magnitudes of all three loss functions are minimized by making a prediction equal to the true click-through rate of the ad. However, the empirical loss function imposes similar penalties on mispredictions that are off by more than some fixed amount, whereas the log likelihood and mean squared error loss functions impose significantly greater penalties for large mispredictions than for small mispredictions. These curves have been normalized so that the largest possible loss is -1. If the curves were instead normalized so that the second derivatives were equal at the optimum, then the log likelihood loss function would go off the page, while the empirical loss function would seem roughly constant.

significantly overpredicts the click-through rate of an ad, then further overpredictions are unlikely to matter because the ad is going to win the auction anyway. Similarly, if one significantly underpredicts the click-through rate of an ad, the further underpredictions are also unlikely to matter because this ad is not likely to be competitive in the auction. Thus an appropriate loss function for machine learning for advertising should not be especially

sensitive to large errors. For this reason, neither mean squared error nor the log likelihood loss function are appropriate loss functions in the context of machine learning for advertising.

Furthermore, the empirical loss function is relatively more sensitive to small errors in the predictions than either mean squared error or the log likelihood loss function. The insights in these few paragraphs are in fact more general than this and hold for other distributions besides the log-normal distribution. In particular, we have the following result:

Theorem 3. *If* $\lim_{A \to 0} g(A|b) = 0$ *and* $\lim_{A \to b} g(A|b) = 0$*, then the derivative of the empirical loss function in Theorem 1 with respect to the predicted click-through rate* q *becomes arbitrarily small in the limit as* $q \to 0$ *or* $q \to 1$*. However, the magnitudes of the derivative of the mean squared error and log likelihood loss functions with respect to* q *are increasing in the distance from* q *to* p*.*[6]

Theorem 3 underscores how different the empirical loss function is from the standard mean squared error and log likelihood loss functions. Under the mean squared error and log likelihood loss functions, the loss function changes the most as a result of small changes in the predicted click-through rate when the predicted click-through rate is as far away from the ad's actual click-through rate as possible. By contrast, as long as the distribution of highest competing eCPM's is such that ads are unlikely to have eCPM's that are either arbitrarily close to zero or arbitrarily large, then the empirical loss function will change by an arbitrarily small amount as a result of small changes in the predicted click-through rate when the predicted click-through rate is as far away from the ad's actual click-through rate as possible. Since the distribution of highest competing eCPM's will typically be such that the highest eCPM is very unlikely to be arbitrarily close to zero or arbitrarily large, this result indicates that the empirical loss function will typically exhibit the exact opposite behavior for predicted click-through rates that differ substantially from the true click-through rates than the standard mean squared error and log likelihood loss functions.

[6]Bax *et al.* (2012) note that in auctions for online advertising, typical click-through rates for ads are on the order of $\frac{1}{100}$ (for search ads) or $\frac{1}{1000}$ (for display ads) so the typical eCPM's for competing ads will be no greater than $\frac{b}{100}$. Thus an eCPM bid of b will typically be at least 100 times larger than the typical eCPM's of the competing ads in the auction, and $\lim_{A \to b} g(A|b)$ will almost certainly be very close to zero in real life.

Theorem 3 indicates that the empirical loss function can have dramatically different properties from the mean squared error and log likelihood loss functions. However, since these standard loss functions do have some desirable properties, we do still wish to verify whether the loss function we have proposed will satisfy these desirable properties. In particular, it is known that both the mean squared error and the log likelihood loss functions are well-calibrated in the sense that these loss functions will be minimized by predicting a click-through rate that is equal to the true expected value of the actual click-through rate of the ad. It is natural to wonder whether this is the case for the empirical loss function as well. It turns out that this loss function is also well-calibrated, as the following theorem illustrates:

Theorem 4. *Suppose that the true click-through rate of the ad is an unknown random variable drawn from the cumulative distribution function $F(\cdot)$. Then the magnitude of the expected empirical loss function in Theorem 1 is minimized by choosing a predicted click-through rate equal to the mean of $F(\cdot)$.*

Theorem 4 illustrates that the loss function considered in Theorem 1 has the desired property of being well-calibrated. We now turn to questions related to how other properties of this loss function compare to the standard loss functions.

One notable difference between the loss function we have proposed and standard loss functions regards the dependence of these loss functions on the bids made by the advertisers. While the standard mean squared error and log likelihood loss functions are independent of the bids made by any of the advertisers, the loss function considered in Theorem 1 is not. This loss function depends crucially on the bids submitted by the advertisers, and the extent to which the loss functions penalizes overpredictions or underpredictions can depend on the size of the bid made by the advertiser. In particular, we obtain the following result:

Theorem 5. *Suppose that $g(\cdot|b)$ is independent of b and single peaked at some eCPM bid $\hat{A}$. Then if $bp > \hat{A}$, the empirical loss function in Theorem 1 imposes stricter penalties for making small underpredictions than for making small overpredictions. But if $bp < \hat{A}$,*

then the empirical loss function in Theorem 1 imposes stricter penalties for making small overpredictions than for making small underpredictions.

Theorem 5 indicates that the loss function we are considering is highly sensitive to the bids that are made. If the CPC bidder submits a CPC bid that is large enough that the bidder's true eCPM is greater than the mode of the highest competing eCPM bid, then the loss function will be relatively more sensitive to underpredictions than to overpredictions, and a misspecified model may wind up being biased towards making overpredictions. By contrast, if the CPC bidder submits a small CPC bid, then the machine learning system will be relatively more sensitive to overpredictions than to underpredictions, and a misspecified model may be more likely to exhibit the opposite biases.

While there is nothing inherently wrong with allowing the predicted click-through rates to vary with the bids, in some applications this might not be desirable because it may lead to incentive compatibility issues arising from an advertiser attempting to manipulate his predicted click-through rate by changing his bid. If the machine learning system is being used to predict click-through rates for ads over a whole sequence of auctions and the bids of the advertisers vary from auction to auction, different advertisers who submit different bids in different auctions may receive different predicted click-through rates even if the underlying features of the ads are the same.

A system designer who wishes to ensure that an advertiser's predicted click-through rate never depends on the advertiser's bid may thus wish to design an alternative loss function that never depends on the particular bids made in any given auction. This system would need to employ a loss function that reflects the economic efficiency loss that is suffered when the predicted click-through rate of the ad differs from the actual click-through rate for a typical bidder. In this case, it is appropriate to use the following loss function:

Theorem 6. *Suppose that one wishes to maximize total economic efficiency while using a loss function that is independent of the bid placed by any particular advertiser. Then if the CPC bids of the bidders in the sequence of auctions are distributed according to the cumulative distribution function $H(\cdot)$, the appropriate loss function for a machine learning system that*

reflects the efficiency loss from predicting that an ad has a click-through rate of q when this ad actually has a click-through rate of p is $\int_0^\infty \int_{bp}^{bq} (bp - A) g(A|b)\, dA\, dH(b)$.

Throughout the analysis so far we have restricted attention to situations in which the loss function depends directly on the actual click-through rate of the ad p. However, in most situations we will not know the true click-through rates of the ads, and it will instead be necessary to define the loss function exclusively in terms of the individual clicks received by the ads. It is natural to ask whether one can easily express an alternative loss function different from that in Theorem 1 that instead depends only on whether an advertiser received a click. We show how one can do this in the following theorem:

Theorem 7. *Suppose that one wishes to maximize total economic efficiency while using a loss function that does not depend on the click-through rates of the advertiser. Then the appropriate loss function for a machine learning system that reflects the efficiency loss from predicting that an ad has a click-through rate of q is* $\int_{bc}^{bq} (bc - A) g(A|b)\, dA$, *where c is a dummy variable that equals* 1 *if the CPC ad received a click and* 0 *otherwise.*

While this loss function differs from the loss function in Theorem 1 in that it no longer depends on some potentially unobserved probability that an advertiser receives a click, it is worth noting that this loss function will still have the desirable property mentioned in Theorem 4 that the magnitude of the expected value of the loss function will be minimized by predicting a click-through rate equal to the actual click-through rate of the ad. For instance, if the true click-through rate of the ad is p, then one can simply apply Theorem 4 to the special case in which the distribution $F(\cdot)$ of actual click-through rates of the ad is a distribution that assumes the value 1 with probability p and 0 with probability $1 - p$, and it immediately follows from this theorem that the magnitude of the expected value of the loss function in Theorem 7 will be minimized by predicting a click-through rate equal to the actual click-through rate of the ad. Another consequence of this theorem is that the expected value of this empirical loss function when an ad's actual probability of a click is p and the predicted click-through rate of the ad is q is $-p \int_{bq}^{b} (b - A)\, g(A|b)\, dA - (1-p) \int_0^{bq} A\, g(A|b)\, dA$.

Thus far we have concerned ourselves solely with designing a loss function that is appropriate when one wishes to maximize expected economic efficiency in the auction. While this is a reasonable goal, one might naturally wonder whether it would also be appropriate to consider using loss functions that consider a weighted average of efficiency and revenue since many systems may care about both of these metrics. Unfortunately, there are significant problems with using a loss function that is optimal in a scenario when one wishes to maximize a weighted average of efficiency and revenue. This is illustrated in the following theorem:

Theorem 8. *The loss function for a machine learning system that maximizes a weighted average of economic efficiency and revenue may result in predictions that are not calibrated in the sense that the magnitude of the expected value of the loss function may not be minimized by predicting a click-through rate equal to the actual click-through rate of the ad.*

This result indicates that there are potential problems with using a loss function that is designed to reflect revenue losses in addition to efficiency losses. Since it is quite important to ensure that the loss function for machine learning for advertising is well-calibrated, Theorem 8 indicates that it is not appropriate to design a loss function that also reflects the revenue loss resulting from misestimates of an ad's click-through rate in addition to the efficiency loss. Using a loss function that reflects revenue losses would result in poorly calibrated predictions, so it is better to simply use a loss function that reflects the efficiency loss from misestimates. In fact, it is further the case that if one seeks to maximize a weighted average of efficiency and revenue subject to the constraint that the predicted click-through rates must be well-calibrated, then one must maximize economic efficiency.[7]

[7]There may be cases in which one wishes to act as if the click-through rate of an ad is higher than one's best estimate of the ad's click-through rate so that one can learn more about the true click-through rate of the ad and use this to more efficiently rank the ads in future auctions (Li *et al.* (2010)). And there may also be cases in which one wishes to act as if the click-through rate of an ad is lower than one's best estimate of the ad's click-through rate if one is highly uncertain about an ad's actual click-through rate (Bax *et al.* (2012) and Lahaie and McAfee (2011)). But while an auctioneer may wish to apply an after-the-fact adjustment to account for these possibilities, the goal of the machine learning system should still be to make unbiased predictions that can be used by the mechanism designer in the most appropriate fashion. Thus it is important for the expected value of the loss function to be optimized by making unbiased predictions.

We can further say something about the circumstances under which using a loss function that is designed to maximize revenue would be optimized by making underpredictions or by making overpredictions. This is done in the following theorem:

Theorem 9. *Suppose that one wishes to maximize revenue and $G(\cdot|b)$ is a cumulative distribution function that has bounded support and is independent of b. Then the appropriate loss function for a machine learning system is optimized by making underpredictions of click-through rates for CPC ads with large bids and overpredictions of click-through rates for CPC ads with small bids.*

To understand the intuition behind this result, note that if a bidder makes a small bid, then it is much more likely that this bidder will be second-pricing the highest competing eCPM bidder than it is that this bidder will win the auction, so there is more to be gained by raising this bidder's predicted click-through rate and increasing the price that the highest competing eCPM bidder will have to pay in the auction. Similarly, if a bidder makes a large bid, then it is more likely that this bidder will win the auction than it is that this bidder will be second-pricing the highest competing eCPM bid, so there is more to be gained by lowering this bidder's predicted click-through rate and increasing the price per click that this bidder will have to pay in the auction. This gives the result in Theorem 9.

4. Value Functions

In this section we briefly digress to introduce a general notion of value functions that reflect the expected value from predicting that an ad has a click-through rate of q. Note that any reasonable value function must be linear in the probability that an ad actually receives a click because if one's expected utility in cases where the ad would receive a click is u_c and one's expected utility in cases where the ad would not receive a click is u_n, then one's total expected utility when p equals the probability that an ad receives a click is $pu_c + (1-p)u_n$.

In general we can decompose a value function into two components. The first component of the value function, which we denote $u(q)$, reflects the expected value that one obtains if

the ad would have received a click when shown as a result of predicting that the click-through rate of the ad is q. It will typically be sensible for $u(q)$ to be increasing in q because an ad is more likely to reap the benefits of the fact that it would receive a click if shown in cases where the ad has a relatively higher predicted click-through rate.

The second component, which we denote $c(q)$, reflects the costs that one pays by predicting that the click-through rate of an ad is q independent of whether the ad actually would have received a click. This can reflect, for example, the expected opportunity cost that one forgoes from the other ads that one will not show as a result of predicting that the click-through rate of an ad is q. Thus an appropriate value function must be of the form $V = pu(q) - c(q)$ for appropriately chosen functions $u(q)$ and $c(q)$.

In order for the value function to be sensible, the cost $c(q)$ must be related to the utility $u(q)$ in such a way that the value function would be maximized by predicting that an ad's click-through rate is equal to the ad's actual click-through rate. Throughout we refer to a value function as being *unbiased* if this property is satisfied.[8] Below we note how the cost $c(q)$ must be related to the utility $u(q)$ in order for this property to be satisfied:

Theorem 10. *Any unbiased value function must be of the form $V = pu(q) - c(q)$, where $c(q)$ satisfies $c(q) = \int_0^q yu'(y)\,dy$.*

This approach of using value functions to represent the value associated with certain predicted click-through rates turns out to encompass minimizing the magnitude of any of the loss functions considered in this paper so far including mean squared error, the log likelihood loss function, and the empirical loss function considered in Theorem 1. In particular, by appropriately choosing the utility function $u(q)$, one can construct a value function V such that maximizing this value function will be equivalent to minimizing the magnitudes of the loss functions that we have considered so far in this paper. This is illustrated in the following theorem:

[8]This is related to the notion of proper scoring rules introduced by Reid and Williamson (2010).

Theorem 11. *When $u(q) = q$, maximizing the value function is equivalent to minimizing mean squared error. When $u(q) = \log(\frac{q}{1-q})$, maximizing the value function is equivalent to minimizing the magnitude of the log likelihood loss function. Finally, when $u(q) = bG(bq|b)$, where $G(\cdot|b)$ represents the distribution of highest competing eCPM bids, then maximizing the value function is equivalent to minimizing the magnitude of the empirical loss function in Theorem 1.*

5. Loss Functions for Position Auctions

Thus far we have focused on the question of the appropriate choice of loss function for standard second price auctions in which there is only a single advertising opportunity available at each auction. While this is certainly an important case in the online advertising world, there are also many cases in which there are several advertising opportunities available on the same page and a single auction is held to determine which advertisers will have their ads shown in each of the various positions on the page. We address the question of the appropriate choice of loss function for these so-called position auctions in this section.[9]

In position auctions there are a total of s positions on a webpage where advertisements can be shown. Each position k has some click-through rate x_k that reflects the relative number of clicks that an advertiser should expect to receive if the advertiser has his ad shown in position k. Throughout we assume that x_k is non-increasing in k so that positions that are higher on the page have higher click-through rates, and we also normalize the parameters so that $x_1 = 1$. The values x_k are assumed to be known.

A machine learning system predicts the click-through rate for an advertiser that reflects the probability that this advertiser's ad would receive a click if his ad was shown in the top position. Thus if p denotes the probability that this advertiser's ad would receive a click if

[9]Position auctions have been extensively analyzed in the literature. See, for example, Aggarwal *et al.* (2006; 2008), Athey and Ellison (2011), Edelman and Ostrovsky (2007), Edelman *et al.* (2007), Edelman and Schwarz (2010), Even-Dar *et al.* (2008), Gonen and Vassilvitskii (2008), Ostrovsky and Schwarz (2009), and Varian (2007).

the ad were in the top position, the probability this advertiser's ad will receive a click if the ad is in position k is $x_k p$.

Advertisers submit bids for clicks and are then ranked on the basis of the product of their bids and their predicted click-through rates, *i.e.* the advertisers are ranked on the basis of their eCPM bids. Throughout we assume for simplicity that advertisers are priced according to Vickrey-Clarke-Groves (VCG) pricing so that advertisers will have an incentive to bid truthfully. VCG pricing has been used by various Internet companies in setting prices for clicks in position auctions for display advertising including Facebook and Google.

As with standard second price auctions that we have considered for the bulk of the paper so far, here we consider the problem of the appropriate choice of loss function for a machine learning system that must predict a click-through rate for an advertiser who has submitted a CPC bid of b into such a position auction. The machine learning system predicts a click-through rate of the ad that is equal to q, and we let p denote the actual click-through rate of the ad, which may differ from q.

We also assume that in each auction, the s highest competing eCPM bids that this advertiser faces are a random draw from the cumulative distribution function $G(v_1, \ldots, v_s|b)$ with corresponding probability density function $g(v_1, \ldots, v_s|b)$, where we let v_k denote the k^{th}-highest bid submitted by some other advertiser. We consider the question of the appropriate loss function when the highest competing bids that the advertiser faces are random draws from this distribution.

Note that in this environment, if the machine learning system predicts a click-through rate of q for the advertiser such that $bq \geq v_1$, then the total expected amount of welfare that is generated is $x_1 bp + x_2 v_1 + \ldots + x_s v_{s-1}$. If the machine learning system predicts a click-through rate of q for the advertiser such that $v_1 > bq \geq v_2$, then the total expected amount of welfare that is generated is $x_1 v_1 + x_2 bp + x_3 v_2 + \ldots + x_s v_{s-1}$. And in general if the machine learning system predicts a click-through rate of q for the advertiser such that $v_{k-1} > bq \geq v_k$ for some $k \geq s$, then the total expected amount of welfare that is generated

is $x_k bp + \sum_{j=1}^{k-1} x_j v_j + \sum_{j=k+1}^{s} x_j v_{j-1}$. Finally if the machine learning system predicts a click-through rate of q for the advertiser such that $bq < v_s$, then the total expected amount of welfare that is generated is $\sum_{j=1}^{s} x_j v_j$.

Now let $Pr(v_{k-1} > bq \geq v_k | q)$ denote the probability that $v_{k-1} > bq \geq v_k$ will be satisfied given a predicted click-through rate of q when the highest competing eCPM bids are drawn from the cumulative distribution function $G(v_1, \ldots, v_s | b)$. Here we abuse notation a bit by defining v_0 to be infinity and v_{s+1} to be 0 so that $Pr(v_0 > bq \geq v_1 | q)$ denotes the probability that $bq \geq v_1$ will be satisfied and $Pr(v_s > bq \geq v_{s+1} | q)$ denotes the probability that $bq < v_s$ will be satisfied given a predicted click-through rate of q when the highest competing eCPM bids are drawn from the cumulative distribution function $G(v_1, \ldots, v_s | b)$.

Given the probabilities defined in the previous paragraph, and the results in the paragraph before this, it follows that the expected welfare arising from predicting that the click-through rate of the CPC ad is q is $\sum_{k=1}^{s+1} [x_k bp + \sum_{j=1}^{k-1} x_j v_j + \sum_{j=k+1}^{s} x_j v_{j-1}] Pr(v_{k-1} > bq \geq v_k | q)$, where we abuse notation by defining x_{s+1} to be equal to zero. From this, we obtain the following result about the appropriate value function for position auctions:

Theorem 12. *The value function that maximizes total economic efficiency for a machine learning system from predicting that an ad has a click-through rate of q when this ad actually has a click-through rate of p is of the form $V = pu(q) - c(q)$, where $u(q) = \sum_{k=1}^{s} b x_k Pr(v_{k-1} > bq \geq v_k | q)$, and $c(q) = \int_0^q y u'(y)\, dy$.*

As with standard second price auctions, it may sometimes be necessary to use a value function or a loss function for position auctions that depends solely on whether a click was observed and not on the actual probability of a click, p, since these probabilities may not be known to the machine learning system. It is thus desirable to come up with some alternative value function or loss function that is still appropriate when one wishes to maximize efficiency in a position auction, but depends solely on whether a click was observed. We illustrate how this can be done in Theorem 13:

Theorem 13. *Suppose that one wishes to maximize economic efficiency while using a value function that does not depend on the click-through rates of the advertiser. Then the appropriate value function for a machine learning system that reflects the value from predicting that an ad has a click-through rate of q is of the form $V = cu(q) - \int_0^q yu'(y)\ dy$, where $u(q) = \sum_{k=1}^{s} bx_k Pr(v_{k-1} > bq \geq v_k|q)$, and c is a dummy variable that equals 1 if the CPC ad received a click and 0 otherwise.*

Finally, in order to transform this value function into a loss function, we must modify the value function in Theorem 13 in a way such that the maximum value the loss function ever assumes is zero while still ensuring that changes in the parameters p and q have the same effect on the loss function as they did on the value function in Theorem 13. The appropriate such loss function is given below:

Theorem 14. *Suppose that one wishes to maximize economic efficiency while using a loss function that does not depend on the click-through rates of the advertiser. Then the appropriate loss function for a machine learning system that reflects the efficiency loss from predicting that an ad has a click-through rate of q is $L = \int_c^q (c - y)u'(y)\ dy$, where $u(q) = \sum_{k=1}^{s} bx_k Pr(v_{k-1} > bq \geq v_k|q)$, and c is a dummy variable that equals 1 if the CPC ad received a click and 0 otherwise.*

6. Concave Value Functions

The analysis we have done so far indicates that the empirical loss function differs significantly from standard loss functions such as the log likelihood loss function. Nonetheless, there may still be a significant disadvantage to using the empirical loss function. Computationally it is much easier to calculate the coefficients that maximize a concave function loss function, and unlike the log likelihood loss function, there is no guarantee that the empirical loss function will be concave in its coefficients. This is illustrated in the following theorem:

Theorem 15. *The empirical loss function in Theorem 1 is not a concave function in q if the highest competing bid that an advertiser faces is drawn from a lognormal distribution.*

While the empirical loss function need not be concave in q, one can still construct loss functions that are preferable to standard loss functions such as mean squared error and the log likelihood loss function even if computational constraints mean that one must use a concave loss function. Practically this is achieved by using a loss function whose shape is equal to the empirical loss function for values of q where the empirical loss function is already concave in q but is then linear for values of q near zero and one where the empirical loss function ceases to be concave. This is illustrated in the following theorem, where we assume that all the bids are one for expositional simplicity:

Theorem 16. *Suppose that one wishes to maximize economic efficiency while using a concave loss function and the competing bid that an advertiser faces is drawn from a distribution such that $(p-q)g(q)$ is increasing in q for values of q near 0 and 1. Then the best concave loss function $L(q,p)$ will have derivative $\frac{\partial L(q,p)}{\partial q}$ that is constant in q for values of q near 0 and 1 and equal to the derivative of the empirical loss function for values of q near p.*

Thus while the empirical loss function may no longer be feasible when one wishes to use a concave loss function, one can still come up with an alternative loss function that is preferable to standard loss functions such as mean squared error and the log likelihood loss function. The solution involves coming up with a concave loss function that is as close as possible to the empirical loss function, while still satisfying the constraint that the loss function is concave in q. This is depicted in Figure 2, where this figure depicts the values of the derivative $\frac{\partial L(q,p)}{\partial q}$ for the empirical loss function in black and the best concave loss function in red.

Theorem 16 addresses the question of how one can construct an optimal loss function that is as close to the empirical loss function as possible while still satisfying the constraint that the loss function must be concave in q. While this is an important question, in some applications concavity of the loss function in q alone is not sufficient for the loss function

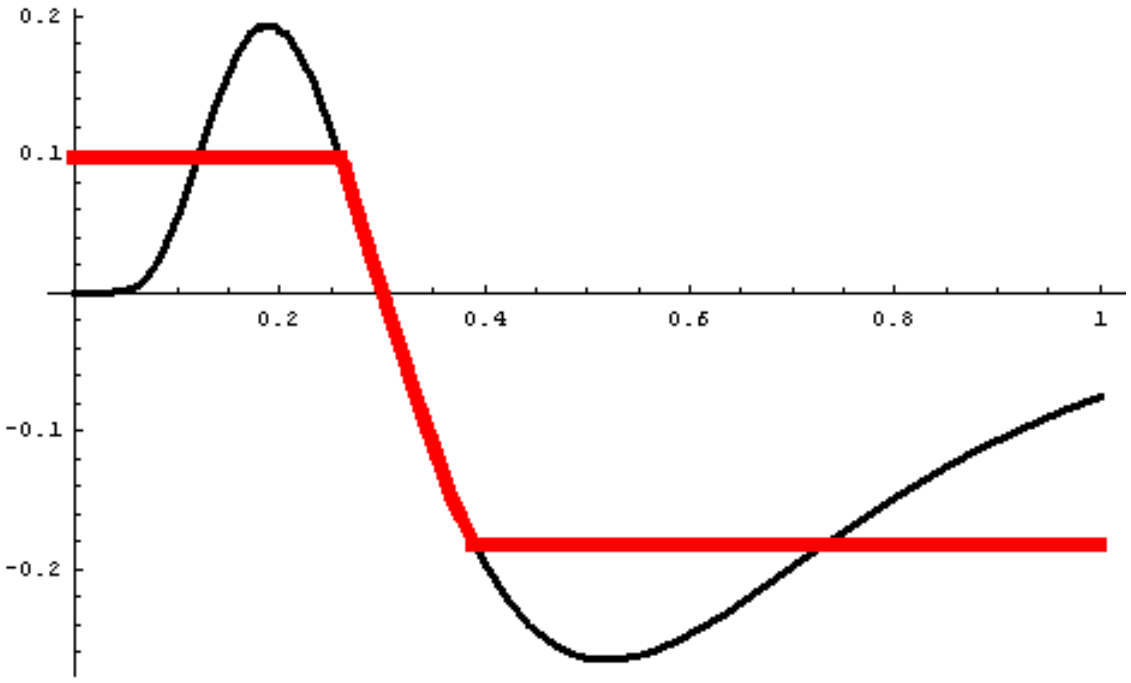

FIGURE 2. If the black curve represents the shape of the derivative of the empirical loss function with respect to the predicted click-through rate, then the red curve represents the shape of the derivative of the best concave loss function with respect to the predicted click-through rate.

minimization to be computationally feasible. Often one wishes to fit a model where the predicted click-through rate is a logistic function of the features of the ad or a model where q is of the form $q = \frac{1}{1+e^{-\sum_{i=1}^{n} \beta_i x_i}}$, where each x_i is an observable feature of the ad, and each β_i is a coefficient on the feature that the model is trying to estimate.

If the model is of this form, then in order for it to be computationally feasible for a system to find the values of the coefficients that minimize the magnitude of the loss function, it is no longer sufficient for the loss function to be concave in the predicted click-through rate q. Instead it must be the case that the loss function is concave in the coefficients $\vec{\beta} = (\beta_1, \ldots, \beta_n)$. We illustrate the form that the optimal loss function takes when the loss function must be concave in the coefficients $\vec{\beta} = (\beta_1, \ldots, \beta_n)$ in Theorem 17:

Theorem 17. *Suppose that one wishes to maximize economic efficiency while using a loss function that is concave in the coefficients* $\vec{\beta} = (\beta_1, \ldots, \beta_n)$ *and fitting a model of the form* $q = \frac{1}{1+e^{-\sum_{i=1}^{n} \beta_i x_i}}$. *Then the best concave loss function* $L(q, p)$ *will have derivative* $\frac{\partial L(q,p)}{\partial q} =$

$\frac{c}{q} + \frac{c}{1-q}$ *for some constant* c *for values of* q *near zero and one (where the constant* c *may be different for values of* q *near zero than it is for values of* q *near one) and equal to the derivative of the empirical loss function for values of* q *near* p*.*

Theorem 17 indicates that the shape of the derivative $\frac{\partial L(q,p)}{\partial q}$ of the best loss function that is concave in the coefficients $\vec{\beta} = (\beta_1, \ldots, \beta_n)$ when we are fitting a model of the form $q = \frac{1}{1+e^{-\sum_{i=1}^{n} \beta_i x_i}}$ is somewhat similar to the shape of the derivative of the log likelihood loss function for values of q near zero since $\frac{\partial L(q,p)}{\partial q} = \frac{p}{q} - \frac{1-p}{1-q}$ for the log likelihood loss function, so these derivatives both vary with $\frac{1}{q}$ for values of q near zero. Nonetheless, the magnitude of this derivative will typically be smaller for the loss function in Theorem 17 since the constant c will typically be lower than p for values of q near zero. Thus the optimal loss function that is concave in its coefficients will continue to impose relatively smaller penalties on large mispredictions than the log likelihood loss function.

In analyzing the shape of the optimal concave loss function so far, we have again assumed that we can allow the loss function to depend on the actual probability of a click p. Although this may be feasible in some circumstances, it is again important to consider what happens in the case where the loss functions may only depend on whether a click was observed, as this is frequently the case in practice. We address this below.

To do this, let $L_c(q)$ denote the loss that is recorded if an ad receives a click and we predicted a click-through rate of q, and let $L_n(q)$ denote the loss that is recorded if an ad does not receive a click and we predicted a click-through rate of q. We know from the reasoning in the proof of Theorem 17 that $L_c(q)$ and $L_n(q)$ will be concave in the coefficients $\vec{\beta} = (\beta_1, \ldots, \beta_n)$ if and only if $q(1-q)L_c''(q) + (1-2q)L_c'(q) \leq 0$ and $q(1-q)L_n''(q) + (1-2q)L_n'(q) \leq 0$. And we also know that if these loss functions are well-calibrated, then it must be the case that the expectations of these loss functions are minimized by making a prediction q that is equal to the true probability of a click p, meaning the derivatives of the expectations of these loss functions must be equal to zero when $q = p$. Thus it must also be the case that $qL_c'(q) + (1-q)L_n'(q) = 0$ for all q. By building on these insights, we derive the form that well-calibrated concave loss functions must take in the theorem below:

Theorem 18. *Suppose that one is fitting a model of the form* $q = \frac{1}{1+e^{-\sum_{i=1}^{n} \beta_i x_i}}$ *and one wishes to use a well-calibrated loss function that is concave in the coefficients* $\vec{\beta} = (\beta_1, \ldots, \beta_n)$*. In that case the set of feasible loss functions* $L_c(q)$ *and* $L_n(q)$ *for the losses that are incurred when one records a click or does not record a click are those satisfying* $L_c'(q) = \frac{h(q)}{q}$ *and* $L_n'(q) = -\frac{h(q)}{1-q}$ *for some non-negative function* $h(q)$ *satisfying* $-\frac{h(q)}{q} \leq h'(q) \leq \frac{h(q)}{1-q}$*.*

Note that the log likelihood loss function is equivalent to the special case of the loss functions in Theorem 18 where $h(q) = 1$ for all q. However, one is free to use other functions $h(q)$ that satisfy $-\frac{h(q)}{q} \leq h'(q) \leq \frac{h(q)}{1-q}$ and may better capture the actual economic loss that one incurs as a result of making inaccurate predictions.

Generally if p represents the actual click-through rate of an ad while q represents the predicted click-through rate of an ad, then the derivative of the expected loss that one incurs as a result of predicting a click-through rate of q with respect to q is $pL_c'(q) + (1-p)L_n'(q) = \frac{ph(q)}{q} - \frac{(1-p)h(q)}{1-q}$. Thus if one predicts a click-through rate q that is some fraction α of the true click-through rate p, then this derivative will be equal to $\frac{h(\alpha p)}{\alpha} - \frac{(1-p)h(\alpha p)}{1-\alpha p} = \frac{[1-\alpha p-\alpha(1-p)]h(\alpha p)}{\alpha(1-\alpha p)} = \frac{(1-\alpha)h(\alpha p)}{\alpha(1-\alpha p)}$ when $q = \alpha p$.

For the empirical loss function, the derivative of the loss function with respect to q will be relatively larger for values of q where it is relatively more likely that small changes in q will have an effect on the outcome of the auction, and it is relatively more likely that small changes in q will have an effect on the outcome of the auction for values of q near the peak of the density corresponding to the distribution of competing bids. This suggests that it would be best to use a loss function of the form in Theorem 18 where $h(q)$ is relatively larger for values of q near the peak of the density corresponding to the distribution of competing bids and relatively smaller for values of q far away from this peak. We derive the form of the optimal choice of this function $h(q)$ in Theorem 19 below:

Theorem 19. *Suppose that one is fitting a model of the form* $q = \frac{1}{1+e^{-\sum_{i=1}^{n} \beta_i x_i}}$ *and one wishes to use a well-calibrated loss function that is concave in the coefficients* $\vec{\beta} = (\beta_1, \ldots, \beta_n)$*. Also suppose that the competing bid that an advertiser faces is drawn from a distribution such*

that $g(q)$ is increasing in q for values of q near 0 and decreasing in q for values of q near 1. Then the optimal choice of the function $h(q)$ for the loss functions $L_c(q)$ and $L_n(q)$ satisfying $L_c'(q) = \frac{h(q)}{q}$ and $L_n'(q) = -\frac{h(q)}{1-q}$ is such that $h(q) = \frac{c_1}{q}$ for some constant c_1 for values of q near 1, $h(q) = \frac{c_0}{1-q}$ for some constant c_0 for values of q near 0, and $h(q) = q(1-q)g(q)$ for values of q where the derivative of $q(1-q)g(q)$ with respect to q is close to 0.

Theorem 19 verifies that it is indeed desirable to choose a function $h(q)$ for these loss functions that is relatively larger for values of q near the peak of the density corresponding to the distribution of competing bids and relatively smaller for values of q far away from this peak. In particular, if $g(q)$ denotes the density corresponding to the distribution of competing bids, then it is optimal to choose a function $h(q)$ that is as close to $q(1-q)g(q)$ as possible. For values of q where $h(q) = q(1-q)g(q)$ automatically satisfies the constraints $-\frac{h(q)}{q} \leq h'(q) \leq \frac{h(q)}{1-q}$ that are needed in Theorem 18, one can simply set $h(q) = q(1-q)g(q)$. Otherwise, one will want to use a function $h(q)$ that is as close to $q(1-q)g(q)$ as possible while still satisfying the constraints $-\frac{h(q)}{q} \leq h'(q) \leq \frac{h(q)}{1-q}$. This entails using a function $h(q)$ that varies with q at the rate $\frac{1}{q}$ for large q and $\frac{1}{1-q}$ for small q.

7. Misspecified Models

While there are clearly significant differences between the empirical loss function considered in Theorem 1 and standard loss functions such as mean squared error and the log likelihood loss function, it may not be clear from the results presented so far whether these differences can actually have a substantive effect on the parameters that are ultimately chosen by these models. Both the empirical loss function considered in Theorem 1 as well as these other loss functions have the property that it would be best to predict a click-through rate equal to the actual click-through rate of the ad, so it might seem that little would be lost by using a loss function that does not reflect the true economic losses that result from making an inaccurate prediction.

While the choice of loss function may very well have little effect if a model is specified perfectly, if a model is misspecified, then the model will sometimes fail to predict click-through rates equal to the actual click-through rate of the ad even if the model has an arbitrarily large amount of training data. The choice of loss function can then have a significant effect on the final parameters that are estimated for the model, even in the limit of an arbitrarily large amount of training data. This section illustrates that the choice of loss function can then have a significant effect on welfare, even with an arbitrarily large amount of training data.

To see this, we first consider a simple example that is based on the value function formulation considered in Section 4. Suppose that the decision maker has a value function $V = pu(q) - c(q)$, where $u(q)$ satisfies $u(q) = \frac{q^a}{a}$ for some a and $c(q) = \int_0^q yu'(y)\,dy = \frac{q^{a+1}}{a+1}$. Also suppose that any given ad has some observable feature x and each ad also has an unobserved click-through rate p that is a deterministic function of x, $p^*(x)$. However, we instead fit a model that gives a predicted click-through rate q that satisfies $q = \phi x$. Then the following result holds:

Theorem 20. *Suppose that the decision maker has a value function $V = pu(q) - c(q)$, where $u(q) = \frac{q^a}{a}$ and $c(q) = \frac{q^{a+1}}{a+1}$. Also suppose that each ad has an unobserved click-through rate p that is a deterministic function of some observable feature x, $p^*(x)$, and we fit a model that gives a predicted click-through rate q that satisfies $q = \phi x$. Then the value of ϕ that is estimated in the limit when there is an infinite amount of training data depends on the parameter a in the value function. In particular, this value of ϕ satisfies $\phi = \frac{E[p^*(x)x^a]}{E[x^{a+1}]}$.*

Theorem 20 indicates that the choice of loss function can have a significant effect on the parameters that are estimated in a misspecified model, even in the limit when there is an arbitrarily large amount of training data. However, it is less clear from this result whether using a loss function that does not accurately reflect the true loss resulting from making inaccurate predictions could have a significant effect on the utility of the decision maker. We address this next.

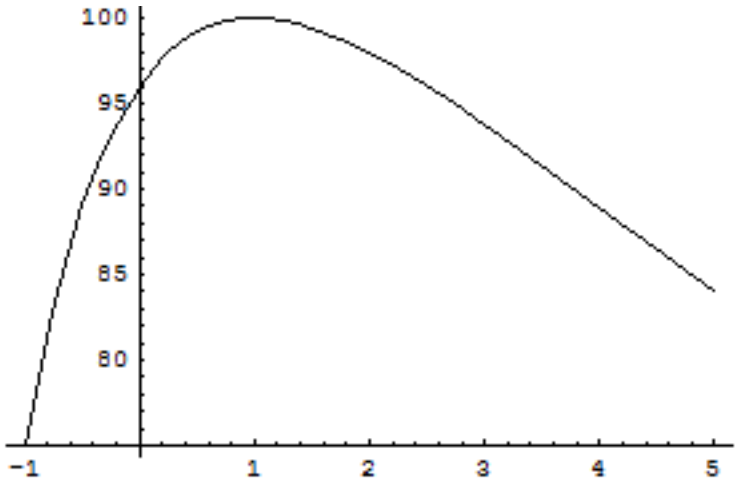

FIGURE 3. Percentage of the maximum welfare that is achieved as a function of a when using a value function with the parameter a.

To address, suppose that the decision maker's true value function is of the form $V = pu(q) - c(q)$, where $u(q)$ satisfies $u(q) = \frac{q^r}{r}$ and $c(q) = \int_0^q su'(s)\, ds = \frac{q^{r+1}}{r+1}$. Also suppose that ads have observable features x that are independent random draws from the uniform distribution on $[0, 1]$ and that the true click-through rate of any given ad, $p^*(x)$, satisfies $p^*(x) = x^\gamma$. However, we instead fit a model that gives a predicted click-through rate q that satisfies $q = \phi x$, and rather than maximize the actual value function, we maximize a different value function of the form $V = pu(q) - c(q)$, where $u(q)$ satisfies $u(q) = \frac{q^a}{a}$ and $c(q) = \int_0^q yu'(y)\, dy = \frac{q^{a+1}}{a+1}$.

Fitting a model that maximizes this different value function rather than the decision maker's true value function can have a significant effect on the decision maker's welfare. To see this, first consider an example in which $\gamma = 4$ and $r = 1$ and consider how the decision maker's actual welfare varies when the decision maker maximizes a value function of the form $V = pu(q) - c(q)$, where $u(q)$ satisfies $u(q) = \frac{q^a}{a}$ and $c(q) = \int_0^q yu'(y)\, dy = \frac{q^{a+1}}{a+1}$ for different values of a rather than maximizing the decision maker's actual value function. In Figure 3 we show how this welfare varies as a percentage of the maximum welfare that the decision maker could achieve when the decision maker maximizes the true value function.

Figure 3 illustrates that maximizing a value function that differs only slightly from the true value function can result in a significant loss in welfare. Moreover, it is generally the

case that this choice of loss function has a significant effect on the decision maker's actual value from using this misspecified model. This is illustrated in the following theorem:

Theorem 21. *Suppose that the decision maker has a value function* $V = pu(q) - c(q)$*, where* $u(q) = \frac{q^r}{r}$ *and* $c(q) = \frac{q^{r+1}}{r+1}$*. Also suppose that the ads have observable features* x *that are independent random draws from the uniform distribution on* $[0, 1]$ *and actual click-through rate* $p^*(x)$ *satisfying* $p^*(x) = x^\gamma$*. Suppose we fit a model that gives a predicted click-through rate* q *that satisfies* $q = \phi x$ *and the parameter* ϕ *is chosen to maximize the value function* $V = pu(q) - c(q)$*, where* $u(q) = \frac{q^a}{a}$ *and* $c(q) = \frac{q^{a+1}}{a+1}$*. Then the decision maker's actual expected value is maximized when* $a = r$*, and the decision maker's expected value from the resulting misspecified model is decreasing in the distance from* a *to* r*.*

We now turn to an example that is more directly connected to the auction environment that we have focused on for the bulk of the paper.

Suppose we are in an environment in which there are two ads in every auction. One ad is an ad that bids on a CPM basis and has a bid in every auction that is a random draw from the beta distribution with parameters α and β. The other ad is an ad that bids on a CPC basis. In every auction, the ad that bids on a CPC basis places a CPC bid of 1, and the ad also has some unknown click-through rate p that is a random draw from the beta distribution with parameters α and β. However, as before, we do not observe the true click-through rate of the ad but we instead only observe some feature x that is correlated with the true click-through rate of the ad. In particular, we assume throughout that the true click-through rate of any ad with feature x is $p^*(x) = x^\gamma$ for some constant γ.

While the true click-through rate of the ad is $p^*(x) = x^\gamma$, we instead fit a misspecified model that gives a predicted click-through rate q that satisfies $q = \phi x$ for some parameter ϕ. In fitting this model, we are able to use an infinite amount of training data on CPC ads with actual click-through rates that are drawn from the true population of estimated click-through rates of these ads. We compare the results that would be obtained by estimating this parameter under three possible loss functions that could be used to estimate this parameter. As is common in the literature, we will start by considering both mean squared error as well

as the log likelihood loss functions. We will then compare the results for these loss functions to the results for the empirical loss function in Theorem 1 that we have considered for much of the paper.

First we illustrate how the values of the parameter ϕ that will be estimated differ as a result of using the different loss functions.

Theorem 22. *Suppose we have an infinite amount of training data on ads with actual click-through rates drawn from the beta distribution with parameters α and β and observable features x such that the true click-through rate of the ad, $p^*(x)$, satisfies $p^*(x) = x^\gamma$. Then if we fit a model that gives a predicted click-through rate q satisfying $q = \phi x$ for some parameter ϕ, the value of the parameter ϕ that will be estimated depends on the choice of loss function. In particular, we have the following:*

(1). If the loss function is mean squared error, then the parameter ϕ that is estimated satisfies $E[\phi x^2 - x^{\gamma+1}] = 0$.

(2). If the loss function is the log likelihood loss function, then the parameter ϕ that is estimated is the $\phi \in (0,1)$ that satisfies $E[\frac{x^\gamma - \phi x}{1-\phi x}] = 0$ if such a ϕ exists and $\phi = 1$ otherwise.

(3). If the loss function is the empirical loss function, then the parameter ϕ that is estimated satisfies $E[x \cdot (x^\gamma - \phi x) \cdot g(\phi x)] = 0$.

Theorem 22 indicates that the parameter ϕ that will be estimated under the different loss functions will be solutions to equations that differ significantly. Furthermore, the values of ϕ that result from the solutions to these three equations may differ significantly. To illustrate this, we consider two possible distributions for the distribution of click-through rates of the CPC bidder. The first distribution we consider is the beta distribution with parameters $\alpha = 2$ and $\beta = 6$ and the second distribution we consider is the beta distribution with parameters $\alpha = 2$ and $\beta = 18$. We report the values of the parameter ϕ that will be estimated for the three loss functions under these distributions for various values of $\gamma < 1$ in Table 1:

Table 1 indicates that for values of $\gamma < 1$, the choice of the loss function can lead to dramatically different parameters in the model that fits predicted click-through rates q to a

Conditions	Mean squared error	Log likelihood loss	Empirical loss
$\alpha = 2, \beta = 6, \gamma = 0.25$	5.42	1.00	51.49
$\alpha = 2, \beta = 6, \gamma = 0.5$	2.20	1.00	3.55
$\alpha = 2, \beta = 6, \gamma = 0.75$	1.36	1.00	1.51
$\alpha = 2, \beta = 18, \gamma = 0.25$	34.82	1.00	822.26
$\alpha = 2, \beta = 18, \gamma = 0.5$	4.60	1.00	8.98
$\alpha = 2, \beta = 18, \gamma = 0.75$	1.76	1.00	2.07

TABLE 1. Estimated values for the parameter ϕ under various loss functions.

function of the form $q = \phi x$ for the observable features x. The parameters estimated for the empirical loss function can be dramatically greater than the parameters estimated for the mean squared error loss function, and these parameters can in turn be significantly greater than those estimated by the log likelihood loss function. In fact, while the parameters for the empirical loss function become increasingly large for values of γ that are significantly less than 1, the corresponding parameters for the log likelihood loss function never exceed 1.

The reason for this is that if one ever chooses a value of $\phi > 1$, then the $\log(1 - \phi x)$ term that appears in the log likelihood loss function will diverge to $-\infty$ in the limit as $x \to \frac{1}{\phi}$, so the loss estimated by this loss function will be arbitrarily large. Thus the largest value of ϕ that will ever be chosen under the log likelihood loss function is $\phi = 1$.

By contrast, under the empirical loss function, there are no such arbitrary penalties for using a parameter ϕ that is greater than 1. Instead, the empirical loss function seeks to choose a value of ϕ so that the misspecified model $q = \phi x$ will match the actual model of click-through rates, $p = x^\gamma$, as well as possible in the region of values of x where this is most likely to have an effect on the auction. Since the predicted click-through rate is more likely to have an effect on which ad wins the auction for the beta distributions considered in Table 1 in cases where x is relatively small, and the slope of how the actual click-through rate varies with x is quite large for small x, this implies that it will be better to choose a very large value of ϕ under the empirical loss function. In fact, the optimal such ϕ becomes larger and larger as γ decreases further from 1 for all the simulations considered in Table 1.

Table 1 indicates that the choice of loss function can have a significant effect on the parameters that are estimated in a misspecified model, but it is not immediately obvious

from this table whether these different parameters will actually have a significant effect on the resulting revenue and efficiency in the auction. We address this next.

To investigate this, we perform the following exercise. We conduct 10,000 simulations of the auction in which we perform the following for each simulation: For each simulation, we randomly draw a CPM bid for the CPM bidder from the beta distribution with parameters α and β and we also independently draw the actual click-through rate of the CPC bidder, p, from the same beta distribution. We then compute the publicly observed feature x by setting $x = p^{1/\gamma}$ and compute predicted click-through rates $q = \phi x$ using the parameters ϕ that were computed in Table 1 for the various loss functions. The CPM bidder wins the auction if this bidder's CPM bid is greater than q and the CPC bidder wins the auction otherwise.

If the CPM bidder wins the auction, then total social welfare or efficiency from running the auction is equal to the CPM bidder's bid and total revenue is equal to q. If the CPC bidder wins the auction, then total efficiency and revenue depends on whether this bidder's ad receives a click. However, expected efficiency will be equal to the probability that this bidder receives a click and expected revenue will be equal to $\frac{pA}{q}$, where p represents the probability that this bidder receives a click, A represents the CPM bidder's bid, and q represents the CPC bidder's predicted click-through rate.

For each simulation of the auction, we note both expected efficiency and expected revenue that would result from the particular randomly drawn CPM bids and actual click-through rates given the analysis in the previous two paragraphs. We then sum the efficiencies and revenues that resulted in each of the 10,000 simulations of the auctions and compute the percentage differences between efficiency and revenue under the different loss functions. The results are summarized in Tables 2-3, where standard errors in the estimated percentage differences are in parentheses.

Tables 2-3 indicate that using the empirical loss function rather than the log likelihood loss function or the mean squared error loss function can have a dramatic effects on both economic efficiency and revenue. The effects are especially pronounced for values of γ that

Conditions	Percentage increase in revenue	Percentage increase in efficiency
$\alpha = 2, \beta = 6, \gamma = 0.25$	54.30% (1.81%)	10.04% (0.25%)
$\alpha = 2, \beta = 6, \gamma = 0.5$	4.35% (0.53%)	1.77% (0.08%)
$\alpha = 2, \beta = 6, \gamma = 0.75$	0.20% (0.04%)	0.16% (0.01%)
$\alpha = 2, \beta = 18, \gamma = 0.25$	83.96% (3.02%)	15.96% (0.35%)
$\alpha = 2, \beta = 18, \gamma = 0.5$	5.16% (0.76%)	2.66% (0.10%)
$\alpha = 2, \beta = 18, \gamma = 0.75$	0.76% (0.19%)	0.24% (0.02%)

TABLE 2. Percentage increase in revenue and efficiency from using the empirical loss function rather than mean squared error with standard errors in parentheses. The results are all statistically significant at the $p < .001$ level.

Conditions	Percentage increase in revenue	Percentage increase in efficiency
$\alpha = 2, \beta = 6, \gamma = 0.25$	395.64% (5.20%)	23.43% (0.44%)
$\alpha = 2, \beta = 6, \gamma = 0.5$	46.08% (1.29%)	11.04% (0.22%)
$\alpha = 2, \beta = 6, \gamma = 0.75$	6.73% (0.44%)	2.24% (0.06%)
$\alpha = 2, \beta = 18, \gamma = 0.25$	5352.90% (60.14%)	30.48% (0.58%)
$\alpha = 2, \beta = 18, \gamma = 0.5$	159.36% (2.71%)	23.27% (0.40%)
$\alpha = 2, \beta = 18, \gamma = 0.75$	17.98% (0.78%)	5.21% (0.12%)

TABLE 3. Percentage increase in revenue and efficiency from using the empirical loss function rather than the log likelihood loss function with standard errors in parentheses. The results are all statistically significant at the $p < .001$ level.

are significantly lower than 1, where using the empirical loss function can lead to an overwhelming improvement in both efficiency and revenue compared to the other loss functions. Thus the choice of loss function does not just have a significant effect on the parameters in the model. The parameter differences lead to significant differences in underlying economic performance as well.

While the above tables indicate that it it possible for the choice of loss function to have a dramatic effect on both the estimated parameters as well as the resulting revenue and efficiency from the auction in a misspecified model, the choice of loss function need not always have a significant effect on the parameters or on revenue and efficiency, even in a severely misspecified model. Tables 4-6 below report the results of similar simulations for the case where $\gamma > 1$.

Conditions	Mean squared error	Log likelihood loss	Empirical loss
$\alpha = 2, \beta = 6, \gamma = 1.25$	0.810	0.794	0.781
$\alpha = 2, \beta = 6, \gamma = 2$	0.562	0.539	0.531
$\alpha = 2, \beta = 6, \gamma = 4$	0.389	0.375	0.373
$\alpha = 2, \beta = 18, \gamma = 1.25$	0.686	0.658	0.647
$\alpha = 2, \beta = 18, \gamma = 2$	0.365	0.338	0.334
$\alpha = 2, \beta = 18, \gamma = 4$	0.200	0.1874	0.1867

TABLE 4. Estimated values for the parameter ϕ under various loss functions.

Conditions	Percentage increase in revenue	Percentage increase in efficiency
$\alpha = 2, \beta = 6, \gamma = 1.25$	-0.02% (0.05%)	0.01% (0.01%)
$\alpha = 2, \beta = 6, \gamma = 2$	-0.002% (0.08%)	0.104%** (0.034%)
$\alpha = 2, \beta = 6, \gamma = 4$	-0.25%** (0.09%)	0.07% (0.05%)
$\alpha = 2, \beta = 18, \gamma = 1.25$	0.01% (0.07%)	0.07%*** (0.01%)
$\alpha = 2, \beta = 18, \gamma = 2$	0.07% (0.12%)	0.16%*** (0.04%)
$\alpha = 2, \beta = 18, \gamma = 4$	-0.41%** (0.14%)	0.07% (0.07%)

TABLE 5. Percentage increase in revenue and efficiency from using the empirical loss function rather than mean squared error with standard errors in parentheses. Here ** denotes $p < .01$ and *** denotes $p < .001$.

As seen in these tables, the choice of loss function has very little effect on either the resulting parameters in the model, revenue, or efficiency, when estimating a misspecified model when $\gamma > 1$. Table 4 indicates that the resulting parameters that are estimated in the model in which predicted click-through rates satisfy $q = \phi x$ are very similar under all

Conditions	Percentage increase in revenue	Percentage increase in efficiency
$\alpha = 2, \beta = 6, \gamma = 1.25$	-0.02% (0.02%)	0.004% (0.006%)
$\alpha = 2, \beta = 6, \gamma = 2$	-0.112%** (0.038%)	-0.03% (0.02%)
$\alpha = 2, \beta = 6, \gamma = 4$	-0.040% (0.023%)	-0.001% (0.01%)
$\alpha = 2, \beta = 18, \gamma = 1.25$	-0.03% (0.02%)	0.010% (0.006%)
$\alpha = 2, \beta = 18, \gamma = 2$	-0.02% (0.03%)	-0.018% (0.014%)
$\alpha = 2, \beta = 18, \gamma = 4$	-0.04%* (0.02%)	0.16%*** (0.04%)

TABLE 6. Percentage increase in revenue and efficiency from using the empirical loss function rather than the log likelihood loss function with standard errors in parentheses. Here * denotes $p < .05$, ** denotes $p < .01$, and *** denotes $p < .001$.

three loss functions considered when $\gamma > 1$, and Tables 5-6 indicate that the difference in revenue and efficiency for the auctions is virtually identical as well. These results hold even when the model is severely misspecified in the sense that γ is significantly greater than 1, the exponent on x that we are attempting to use in the model.

This insight that the choice of loss function has little effect on the auction even when the model is severely misspecified in the sense that γ is significantly greater than 1 is one that continues to hold even in the limit as the model becomes arbitrarily misspecified. In fact, in the limit as γ becomes arbitrarily large, the choice of loss function has no effect on either the resulting parameters in the model, revenue, or efficiency, even though the actual dependence of the click-through rates on the features is arbitrarily different from the that given by the model we are trying to fit. In particular, we have the following result:

Theorem 23. *In the limit as $\gamma \to \infty$, the parameters ϕ that are estimated for the three loss functions considered in Theorem 22 all approach $\phi = E[p]$.*

To understand the intuition behind Theorem 23, note that in the limit as $\gamma \to \infty$, the observable features x of the ads will all become arbitrarily close to 1 regardless of the actual click-through rates of the ads. Thus fitting a model that predicts click-through rates of

the form $q = \phi x$ reduces to fitting a model that predicts very similar click-through rates for almost all the ads. In this case, regardless of the choice of loss function, the optimal parameters will be such that the predicted click-through rates of the ads are all arbitrarily close to the average click-through rates of the ads, so a parameter ϕ arbitrarily close to $E[p]$ will be optimal for all the loss functions.

Finally, we turn to a setting motivated by the analysis of concave value functions in Section 6. As discussed in Section 6, in practical applications one typically fits a model of the form $q = \frac{1}{1+e^{-\sum_{i=1}^{n} \beta_i x_i}}$ while using a well-calibrated loss function that is concave in the coefficients $\vec{\beta} = (\beta_1, \ldots, \beta_n)$ and depends only on whether an advertiser received a click. We now address whether performance improvements can be obtained from using a concave approximation to the empirical loss function when the model that one is trying to fit is misspecified.

As before, we suppose we are in an environment in which there are two ads in every auction. One ad is an ad that bids on a CPM basis and has a bid in every auction that is a random draw from the beta distribution with parameters α and β. The other ad is an ad that bids on a CPC basis. In every auction, the ad that bids on a CPC basis places a CPC bid of 1, and the ad also has some unknown click-through rate p that is a random draw from the beta distribution with parameters α and β. However, as before, we do not observe the true click-through rate of the ad but we instead only observe some feature x that is correlated with the true click-through rate of the ad. In particular, we assume throughout that the true click-through rate of any ad with feature x is $p^*(x) = x^\gamma$ for some constant γ.

While the true click-through rate of the ad is $p^*(x) = x^\gamma$, we instead fit a misspecified model that gives a predicted click-through rate q that satisfies $q = \frac{1}{1+e^{-\phi_1 - \phi_2 x}}$ for some parameters ϕ_1 and ϕ_2. In fitting this model, we are able to use an infinite amount of training data on CPC ads with actual click-through rates that are drawn from the true population of estimated click-through rates of these ads.

We compare the results that would be obtained by estimating these parameters under three possible loss functions that could be used to estimate these parameters. As before we consider the log likelihood loss function and the empirical loss function, but we now also consider what

would happen if we used a concave approximation to the empirical loss function based on our analysis of the best concave approximation to the empirical loss function considered in Theorem 19. In particular, we consider the parameters that would result from using loss functions $L_c(q)$ and $L_n(q)$ corresponding to the losses that are recorded if an ad receives a click or does not receive a click respectively that satisfy $L_c'(q) = \frac{h(q)}{q}$ and $L_n'(q) = -\frac{h(q)}{1-q}$, where $h(q)$ is a function satisfying $h(q) = \frac{c_1}{q}$ for values of $q \leq q_L$, $h(q) = q(1-q)g(q)$ for values of $q \in (q_L, q_H)$, and $h(q) = \frac{c_2}{1-q}$ for values of $q \geq q_H$, where $g(q)$ denotes the probability density function corresponding to the beta distribution with parameters α and β, c_1 and c_2 are appropriately chosen constants, and (q_L, q_H) denotes the interval in which the function $h(q) = q(1-q)g(q)$ satisfies the constraint $-\frac{h(q)}{q} \leq h'(q) \leq \frac{h(q)}{1-q}$ given as a necessary and sufficient condition for concavity in Theorem 18.

First we present results noting how the parameters that will be selected vary with the choice of loss function.

Theorem 24. *Suppose we have an infinite amount of training data on ads with actual click-through rates drawn from the beta distribution with parameters α and β and observable features x such that the true click-through rate of the ad, $p^*(x)$, satisfies $p^*(x) = x^\gamma$. Then if we fit a model that gives a predicted click-through rate q satisfying $q = \frac{1}{1+e^{-\phi_1-\phi_2 x}}$ for some parameters ϕ_1 and ϕ_2, the values of the parameters that will be estimated depend on the choice of loss function. In particular, we have the following:*

(1). If the loss function is the log likelihood loss function, then the parameters ϕ_1 and ϕ_2 that are estimated will be solutions to the set of equations $E[\frac{x^\gamma}{1+e^{\phi_1+\phi_2 x}} - \frac{1-x^\gamma}{1+e^{-\phi_1-\phi_2 x}}] = 0$ and $E[\frac{x^{\gamma+1}}{1+e^{\phi_1+\phi_2 x}} - \frac{x(1-x^\gamma)}{1+e^{-\phi_1-\phi_2 x}}] = 0$.

(2). If the loss function is the empirical loss function, then the parameters ϕ_1 and ϕ_2 that are estimated will be solutions to the set of equations $E[\frac{x^\gamma e^{\alpha(\phi_1+\phi_2 x)}}{(1+e^{\phi_1+\phi_2 x})^{\alpha+\beta+1}} - \frac{(1-x^\gamma)e^{(\alpha+1)(\phi_1+\phi_2 x)}}{(1+e^{-\phi_1-\phi_2 x})^{\alpha+\beta+1}}] = 0$ and $E[\frac{x^{\gamma+1}e^{\alpha(\phi_1+\phi_2 x)}}{(1+e^{\phi_1+\phi_2 x})^{\alpha+\beta+1}} - \frac{x(1-x^\gamma)e^{(\alpha+1)(\phi_1+\phi_2 x)}}{(1+e^{-\phi_1-\phi_2 x})^{\alpha+\beta+1}}] = 0$.

(3). If the loss function is the concave approximation to the empirical loss function given above, then the parameters ϕ_1 and ϕ_2 that are estimated will be solutions to the set of equations $n_L E[x^\gamma(1+e^{\phi_1+\phi_2 x})-1|x \leq x_L]Pr(x \leq x_L) + E[\frac{x^\gamma e^{\alpha(\phi_1+\phi_2 x)}}{(1+e^{\phi_1+\phi_2 x})^{\alpha+\beta+1}} - \frac{(1-x^\gamma)e^{(\alpha+1)(\phi_1+\phi_2 x)}}{(1+e^{-\phi_1-\phi_2 x})^{\alpha+\beta+1}}|x \in$

$(x_L, x_H)]Pr(x \in (x_L, x_H)) + n_H E[x^\gamma(1 + e^{-\phi_1-\phi_2 x}) - 1|x \geq x_H]Pr(x \geq x_H) = 0$ *and* $n_L E[x^{\gamma+1}(1 + e^{\phi_1+\phi_2 x}) - x|x \leq x_L]Pr(x \leq x_L) + E[\frac{x^{\gamma+1}e^{\alpha(\phi_1+\phi_2 x)}}{(1+e^{\phi_1+\phi_2 x})^{\alpha+\beta+1}} - \frac{x(1-x^\gamma)e^{(\alpha+1)(\phi_1+\phi_2 x)}}{(1+e^{-\phi_1-\phi_2 x})^{\alpha+\beta+1}}|x \in (x_L, x_H)]Pr(x \in (x_L, x_H)) + n_H E[x^{\gamma+1}(1 + e^{-\phi_1-\phi_2 x}) - x|x \geq x_H]Pr(x \geq x_H) = 0$ *where* $x_L \equiv \frac{\log(\frac{q_L}{1-q_L})-\phi_1}{\phi_2}$, $x_H \equiv \frac{\log(\frac{q_H}{1-q_H})-\phi_1}{\phi_2}$, $n_L \equiv q_L^\alpha(1-q_L)^{\beta+1}$, *and* $n_H \equiv q_H^{\alpha+1}(1-q_H)^\beta$.

Having noted how the parameters will vary with the underlying loss functions, we now report the parameters and the changes in economic efficiency that would result from the use of the various loss functions considered. These results are reported in Tables 7 and 8.

Conditions	Log likelihood loss	Concave empirical loss	Empirical loss
$\alpha = 2, \beta = 2, \gamma = 1$	(-2.36, 4.73)	(-2.34, 4.67)	(-2.25, 4.50)
$\alpha = 2, \beta = 10, \gamma = 1$	(-2.85, 6.57)	(-2.93, 7.07)	(-2.97, 7.58)
$\alpha = 2, \beta = 10, \gamma = 2$	(-4.14, 6.04)	(-4.18, 6.16)	(-4.12, 6.07)
$\alpha = 2, \beta = 10, \gamma = 0.5$	(-2.16, 11.99)	(-2.25, 14.73)	(-2.45, 22.94)
$\alpha = 2, \beta = 10, \gamma = 0.25$	(-1.80, 41.90)	(-1.88, 67.99)	(-2.05, 183.57)
$\alpha = 2, \beta = 20, \gamma = 0.5$	(-2.72, 27.77)	(-2.85, 38.60)	(-3.21, 88.88)
$\alpha = 2, \beta = 20, \gamma = 0.25$	(-2.41, 203.19)	(-2.51, 439.66)	(-2.88, 2439.5)
$\alpha = 2, \beta = 40, \gamma = 0.5$	(-3.34, 77.97)	(-3.49, 121.45)	(-3.54, 240.62)
$\alpha = 2, \beta = 40, \gamma = 0.25$	(-3.03, 783.82)	(-3.17, 4129.6)	(-3.18, 7701.2)
$\alpha = 2, \beta = 100, \gamma = 0.5$	(-4.35, 507.40)	(-4.38, 649.83)	(-4.34, 732.21)
$\alpha = 1, \beta = 20, \gamma = 0.5$	(-3.31, 49.58)	(-3.51, 86.99)	(-3.89, 228.93)

TABLE 7. Estimated values for the parameter ϕ under various loss functions.

As with the previous simulations, the results in Tables 7 and 8 illustrate that using a concave approximation to the empirical loss function will sometimes, although not always, lead to a significant improvement in performance compared to using the log likelihood loss function. For some of the simulations considered in these tables, using a concave approximation to the empirical loss function results in dramatically different parameters as well as efficiency gains greater than a full percentage point—a tremendous improvement in social welfare considering the scale of most online advertising systems. But other simulations considered in these tables, using a concave approximation to the empirical loss function has little effect on either the parameters of the model or the overall efficiency of the system.

While using the actual empirical loss function always results in some statistically significant efficiency gain relative to using the concave approximation to the empirical loss function,

Conditions	Initial efficiency increase	Additional efficiency increase
$\alpha = 2, \beta = 2, \gamma = 1$	0.0021% (0.0004%)	0.0045% (0.0007%)
$\alpha = 2, \beta = 10, \gamma = 1$	0.055% (0.003%)	0.033% (0.003%)
$\alpha = 2, \beta = 10, \gamma = 2$	0.0045% (0.0007%)	0.0034% (0.0007%)
$\alpha = 2, \beta = 10, \gamma = 0.5$	0.309% (0.009%)	0.320% (0.013%)
$\alpha = 2, \beta = 10, \gamma = 0.25$	0.805% (0.019%)	1.217% (0.023%)
$\alpha = 2, \beta = 20, \gamma = 0.5$	0.602% (0.014%)	0.411% (0.021%)
$\alpha = 2, \beta = 20, \gamma = 0.25$	1.209% (0.027%)	1.872% (0.018%)
$\alpha = 2, \beta = 40, \gamma = 0.5$	0.920% (0.018%)	0.172% (0.019%)
$\alpha = 2, \beta = 40, \gamma = 0.25$	2.331% (0.042%)	0.937% (0.021%)
$\alpha = 2, \beta = 100, \gamma = 0.5$	0.592% (0.014%)	0.203% (0.012%)
$\alpha = 1, \beta = 20, \gamma = 0.5$	1.333% (0.026%)	1.060% (0.028%)

TABLE 8. Initial percentage increase in efficiency from using the concave approximation empirical loss function rather than the log likelihood loss function as well as the further percentage increase in efficiency from using the empirical loss function rather than the concave approximation to the empirical loss function. Standard errors are in parentheses. The results are all statistically significant at the $p < .001$ level.

there appears to be no general relationship as to the relative fraction of the economic efficiency gains that are lost by using the empirical loss function rather than the concave approximation to the empirical loss function. For some of the parameters considered in Tables 7 and 8, using a concave approximation to the empirical loss function rather than the actual empirical loss function enables one to capture the vast majority of the efficiency gains that could be achieved by using a different loss function than the log likelihood loss function. For other values of the parameters, using a concave approximation to the empirical loss function only enables one to capture less than half of the efficiency gains that could be achieved by using the empirical loss function. Thus it is difficult to say whether the need

to impose concavity is likely to result in significant efficiency losses relative to what could otherwise be achieved if it were feasible to optimize non-concave loss functions in practice.

8. Conclusion

This paper has considered the question of the choice of the optimal loss functions for predicted click-through rates in auctions for online advertising. We have shown that a loss function reflecting the true empirical loss that one suffers as a result of making inaccurate predictions would impose significant penalties for small mispredictions while imposing only slightly larger penalties on large mispredictions. This is in stark contrast to standard loss functions such as mean squared error and the log likelihood loss function. Moreover, using the empirical loss function rather than these standard loss functions has the potential to dramatically improve performance in the auction if the model that one is trying to fit is misspecified.

Our analysis has also delivered a number of other insights. We have illustrated that the empirical loss function may depend on the bids of the advertisers in such a way that underpredictions of click-through rates are more severely penalized than overpredictions of click-through rates for advertisers with large bids, while overpredictions of click-through rates are more severely penalized than underpredictions of click-through rates for advertisers with small bids. We have also shown that similar optimal loss functions to those derived in our main setting with an auction for a single advertising opportunity can be derived for position auctions in which there is an auction for several advertising opportunities on the page at the same time.

Finally, we have considered the question of the optimal loss function one can use for predicted click-through rates when one is restricted to using a concave loss function for reasons of computational tractability. When one is restricted to using a concave loss function, it may no longer be feasible to use the true empirical loss function. However, we have shown how one can still improve on standard loss functions such as the log likelihood loss function by instead adopting a loss function that is equal to the true empirical loss function in regions

where the empirical loss function is concave, while coming as close to the empirical loss function as possible without violating concavity in regions where the empirical loss function is not concave.

We close with two suggestions for future research. Throughout this paper we have focused on a situation in which a machine learning system predicts a point estimate for an ad's click-through rate but does not give a conditional distribution for the ad's true click-through rate given the uncertainty the system has in its prediction. This corresponds well to practice, since most machine learning systems typically only predict a point estimate for an ad's click-through rate rather than a conditional distribution. However, it is interesting to think about the question of the loss function that one would use in online auctions if a machine learning system could indeed predict a distribution of an ad's possible click-through rates rather than just a point estimate. We expect that similar types of analyses could illuminate the appropriate choice of loss function in this framework, and leave a full investigation of this to future work.

Finally, we note that the loss functions considered in this paper represent loss functions that could be implemented in practice in serving ads for online auctions. Using the loss functions in this paper rather than standard loss functions such as log likelihood has the potential to improve the welfare of both advertisers and publishers participating in auctions for online advertising. We leave the possibility of implementing these loss functions in practice to future work.

Acknowledgments

We thank Glenn Ellison, Evan Ettinger, Daniel Golovin, David Grether, Pierre Grinspan, Chris Harris, Randall Lewis, Charles Manski, Brendan McMahan, Robert Porter, Steve Scott, Tal Shaked, Hal Varian, and Martin Zinkevich for helpful comments and discussions.

References

Aggarwal, Gagan, Jon Feldman, S. Muthukrishnan, and Martin Pál. 2008. "Sponsored Search Auctions with Markovian Users." *Proceedings of the 4th International Workshop on Internet and Network Economics* (WINE) 4: 621-628.

Aggarwal, Gagan, Ashish Goel, and Rajeev Motwani. 2006. "Truthful Auctions for Pricing Search Keywords." *Proceedings of the 7th ACM Conference on Electronic Commerce* (EC) 7: 1-7.

Altun, Yasemin, Mark Johnson, and Thomas Hofmann. 2003. "Investigating Loss Functions and Optimization Methods for Discriminative Learning of Label Sequences." *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing* (EMNLP) 145-152.

Arrow, Kenneth J. 1959. "Decision Theory and the Choice of a Level of Significance for the t-Test." In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, ed. by Ingram Olkin, pp. 70-78.

Athey, Susan and Glenn Ellison. 2011. "Position Auctions with Consumer Search." *Quarterly Journal of Economics* 126(3): 1211-1270.

Bartlett, Peter L., Michael I. Jordan, and Jon D. McAuliffe. 2006. "Convexity, Classification, and Risk Bounds." *Journal of the American Statistical Association* 101(473): 138-156.

Bax, Eric, Anand Kuratti, R. Preston McAfee, and Julian Romero. 2012. "Comparing Predicted Prices in Auctions for Online Advertising." *International Journal of Industrial Organization* 30: 80-88.

Beck, Marissa and Paul Milgrom. 2012. "Auctions, Adverse Selection, and Internet Display Advertising." Stanford University Typescript.

Edelman, Benjamin and Michael Ostrovsky. 2007. "Strategic Bidder Behavior in Sponsored Search Auctions." *Decision Support Systems.* 43(1): 192-198.

Edelman, Benjamin, Michael Ostrovsky, and Michael Schwarz. 2007. "Internet Advertising and the Generalized Second Price Auction: Selling Billions of Dollars Worth of Keywords." *American Economic Review.* 97(1): 242-259.

Edelman, Benjamin and Michael Schwarz. 2010. "Optimal Auction Design and Equilibrium Selection in Sponsored Search Auctions." *American Economic Review Papers and Proceedings.* 100(2): 597-602.

Even-Dar Eyal, Jon Feldman, Yishay Mansour, and S. Muthukrishnan. 2008. "Position Auctions with Bidder-Specific Minimum Prices." *Proceedings of the 4th International Workshop on Internet and Network Economics* (WINE) 4: 577-584.

Gonen, Rica and Sergei Vassilvitskii. 2008. "Sponsored Search Auctions with Reserve Prices: Going Beyond Seperability." *Proceedings of the 4th International Workshop on Internet and Network Economics* (WINE) 4: 597-608.

Hilbe, Joseph M. 2009. *Logistic Regression Models.* Chapman & Hall: London.

Lahaie, Sébastian and R. Preston McAfee. 2011. "Efficient Ranking in Sponsored Search." *Proceedings of the 7th International Workshop on Internet and Network Economics* (WINE) 7: 254-265.

Lahaie, Sébastian and David M. Pennock. 2007. "Revenue Analysis of a Family of Ranking Rules for Keyword Auctions." *Proceedings of the 8th ACM Conference on Electronic Commerce* (EC) 8: 50-56.

Li, Sai-Ming, Mohammad Mahdian, and R. Preston McAfee. 2010. "Value of Learning in Sponsored Search Auctions." *Proceedings of the 6th International Workshop on Internet and Network Economics* (WINE) 6: 294-305.

Manski, Charles F. 2004. "Statistical Treatment Rules for Heterogeneous Populations." *Econometrica* 72(4): 1221-1246.

Manski, Charles F. 2006. "Search Profiling with Partial Knowledge of Deterrence." *Economic Journal* 116: 385-401.

Manski, Charles F. 2009. "The 2009 Lawrence R. Klein Lecture: Diversified Treatment Under Ambiguity." *International Economic Review* 50(4): 1013-1041.

Maronna, Ricardo, R. Douglas Martin, and Victor Yohai. 2006. *Robust Statistics - Theory and Methods.* John Wiley & Sons, Ltd.: West Sussex.

McAfee, R. Preston, Kishore Papineni, and Sergei Vassilvitskii. 2013. "Maximally Representative Allocations for Guaranteed Delivery Advertising Campaigns." *Review of Economic Design* 17: 83-94.

McMahan, Brendan H., Gary Holt, D. Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, Sharat Chikkerur, Dan Liu, Martin Wattenberg, Arnar Mar Hrafnkelsson, Tom Boulos, and Jeremy Kubica. 2013. "Ad Click Prediction: A View from the Trenches." *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD) 19.

Moyé, Lemuel A. 2006. *Statistical Reasoning in Medicine: The Intuitive P-Value Primer.* Springer: New York.

Ostrovsky, Michael and Michael Schwarz. 2009. "Reserve Prices in Internet Advertising Auctions: A Field Experiment." Stanford Graduate School of Business Typescript.

Reid, Mark D. and Robert C. Williamson. 2010. "Composite Binary Losses." *Journal of Machine Learning Research* 11(Sep): 2387-2422.

Reid, Mark D. and Robert C. Williamson. 2011. "Information, Divergence, and Risk for Binary Experiments." *Journal of Machine Learning Research* 12(Mar): 731-817 (2011).

Sawa, Takamitsu. 1978. "Information Criteria for Discriminating Among Alternative Regression Models." *Econometrica* 46(6): 1273-1291.

Skalak, David B., Alexandru Niculescu-Mizil, and Rich Caruna. 2007. "Classifier Loss Under Metric Uncertainty." *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing* (EMNLP) 310-322.

Steinwart, Ingo. 2007. "How to Compare Different Loss Functions and Their Risks." *Constructive Approximation* 26(2): 225-287.

Varian, Hal R. 1974. "A Bayesian Approach to Real Estate Assessment." In *Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage*, ed. Stephen E. Fienberg and Arnold Zellner, pp. 195-208. Amsterdam: North-Holland.

Varian, Hal R. 2007. "Position Auctions." *International Journal of Industrial Organization.* 25(6): 1163-1178.

White, Halbert. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica* 50(1): 1-25.

Vuong, Quang H. 1989. "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica* 57(2): 307-333.

Zhang, Tong. 2004. "Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization." *Annals of Mathematical Statistics* 32(1): 56-134.

Online Appendix: Proofs of Main Results

Proof of Theorem 1: We have seen that the expected social welfare that results from predicting a click-through rate of q when the actual click-through rate of the ad is p is $bpPr(bq \geq A) + APr(bq < A) = \int_{bq}^{\infty} A\ g(A|b)\ dA + \int_{0}^{bq} bp\ g(A|b)\ dA$. Thus the expected social welfare that results from correctly predicting that the click-through rate of an ad is p when the actual click-through rate of the ad is p is $\int_{bp}^{\infty} A\ g(A|b)\ dA + \int_{0}^{bp} bp\ g(A|b)\ dA$. From this it follows that the loss in efficiency that results from predicting that the click-through rate of an ad is q when the actual click-through rate of the ad is p is $\int_{bq}^{\infty} A\ g(A|b)\ dA + \int_{0}^{bq} bp\ g(A|b)\ dA - [\int_{bp}^{\infty} A\ g(A|b)\ dA + \int_{0}^{bp} bp\ g(A|b)\ dA] = \int_{bp}^{bq}(bp - A)g(A|b)\ dA$. The result then follows. □

Proof of Theorem 2: If $g(A|b)$ is constant throughout its support, then $\int_{bp}^{bq}(bp-A)g(A|b)\ dA$ is proportional to $\int_{bp}^{bq}(bp-A)\ dA = bp(bq-bp) - \frac{(bq)^2-(bp)^2}{2} = -\frac{(bq)^2-2(bq)(bp)+(bp)^2}{2} = -\frac{b^2}{2}(q-p)^2$. Since minimizing this function is equivalent to minimizing mean squared error, the result then follows. □

Proof of Theorem 3: The derivative of the empirical loss function in Theorem 1 with respect to q is $b^2(p-q)g(bq|b)$. If $\lim_{A\to 0} g(A|b) = 0$ and $\lim_{A\to b} g(A|b) = 0$, then this derivative goes to zero in the limit as $q \to 0$ or $q \to 1$. Now the derivative of $(q-p)^2$ with respect to q is $2(q-p)$, which is increasing in the distance from q to p. And the derivative of $p\log(q) + (1-p)\log(1-q)$ with respect to q is $\frac{p}{q} - \frac{1-p}{1-q}$. This derivative is decreasing in q and equal to zero when $q = p$, so the magnitude of this derivative is increasing in the distance from q to p both for values of $q < p$ and for values of $q > p$. Thus the magnitudes of the derivative of the mean squared error and log likelihood loss functions with respect to q are increasing in the distance from q to p, but the derivative of the loss function in Theorem 1 with respect to the predicted click-through rate q becomes arbitrarily small in the limit as $q \to 0$ or $q \to 1$. □

Proof of Theorem 4: If $g(\cdot) \equiv g(\cdot|b)$ is single peaked at some eCPM bid $\hat{A}$ and $bp > \hat{A}$, then $g(bp-\epsilon) > g(bp+\epsilon)$ for all small values of $\epsilon > 0$, so $|(bp-A)g(A)|$ is greater when $A = b(p-\epsilon)$ than when $A = b(p+\epsilon)$ for all small values of $\epsilon > 0$. From this it follows that $|\int_{bp}^{bq}(bp-A)g(A)\,dA|$ is larger when $q = p-\epsilon$ than when $q = p+\epsilon$ for all small values of $\epsilon > 0$, and the loss function in Theorem 1 imposes stricter penalties for making small underpredictions than for making small overpredictions if $bp > \hat{A}$.

But if $bp < \hat{A}$, then $g(bp-\epsilon) < g(bp+\epsilon)$ for all small values of $\epsilon > 0$, so $|(bp-A)g(A)|$ is greater when $A = b(p+\epsilon)$ than when $A = b(p-\epsilon)$ for all small values of $\epsilon > 0$. From this it follows that $|\int_{bp}^{bq}(bp-A)g(A)\,dA|$ is larger when $q = p+\epsilon$ than when $q = p-\epsilon$ for all small values of $\epsilon > 0$. Thus the loss function in Theorem 1 imposes stricter penalties for making small overpredictions than for making small underpredictions if $bp < \hat{A}$, and the result then follows. □

Proof of Theorem 5: If $g(\cdot) \equiv g(\cdot|b)$ is single peaked at some eCPM bid $\hat{A}$ and $bp > \hat{A}$, then $g(bp-\epsilon) > g(bp+\epsilon)$ for all small values of $\epsilon > 0$, so $|(bp-A)g(A)|$ is greater when $A = b(p-\epsilon)$ than when $A = b(p+\epsilon)$ for all small values of $\epsilon > 0$. From this it follows that $|\int_{bp}^{bq}(bp-A)g(A)\,dA|$ is larger when $q = p-\epsilon$ than when $q = p+\epsilon$ for all small values of $\epsilon > 0$, and the loss function in Theorem 1 imposes stricter penalties for making small underpredictions than for making small overpredictions if $bp > \hat{A}$.

But if $bp < \hat{A}$, then $g(bp-\epsilon) < g(bp+\epsilon)$ for all small values of $\epsilon > 0$, so $|(bp-A)g(A)|$ is greater when $A = b(p+\epsilon)$ than when $A = b(p-\epsilon)$ for all small values of $\epsilon > 0$. From this it follows that $|\int_{bp}^{bq}(bp-A)g(A)\,dA|$ is larger when $q = p+\epsilon$ than when $q = p-\epsilon$ for all small values of $\epsilon > 0$. Thus the loss function in Theorem 1 imposes stricter penalties for making small overpredictions than for making small underpredictions if $bp < \hat{A}$, and the result then follows. □

Proof of Theorem 6: We have seen in Theorem 1 that if the CPC bidder in the auction has a CPC bid of b, then the appropriate loss function for a machine learning system that reflects the efficiency loss from predicting that an ad has a click-through rate of q when this

ad actually has a click-through rate of p is $\int_{bp}^{bq}(bp - A)g(A|b)\ dA$. From this it follows that if the CPC bidder in the auction has a CPC bid that is a random draw from the cumulative distribution function $H(\cdot)$, then the expected efficiency loss that results from predicting that this ad has a click-through rate of q when this ad actually has a click-through rate of p is $\int_0^{\infty}\int_{bp}^{bq}(bp - A)g(A|b)\ dA\ dH(b)$. The result then follows. $\square$

Proof of Theorem 7: The social welfare that results from predicting a click-through rate of q is $bcPr(bq \geq A) + APr(bq < A) = \int_{bq}^{\infty} A\ g(A|b)\ dA + \int_0^{bq} bc\ g(A|b)\ dA$. And the social welfare that results from correctly predicting whether an ad will receive a click in any given auction is $\int_{bc}^{\infty} A\ g(A|b)\ dA + \int_0^{bc} bc\ g(A|b)\ dA$. From this it follows that the loss in efficiency that results from predicting that the click-through rate of an ad is q rather than correctly predicting whether the ad will receive a click is $\int_{bq}^{\infty} A\ g(A|b)\ dA + \int_0^{bq} bc\ g(A|b)\ dA - [\int_{bc}^{\infty} A\ g(A|b)\ dA + \int_0^{bc} bc\ g(A|b)\ dA] = \int_{bc}^{bq}(bc - A)g(A|b)\ dA$. The result then follows. $\square$

Proof of Theorem 8: To prove this result, it suffices to show that if one wishes to maximize expected revenue, then the appropriate loss function for a machine learning system that reflects the expected revenue loss from predicting that an ad has a click-through rate of q when this ad actually has a click-through rate of p may result in predictions that are not calibrated. Suppose the only two ads in the system are the CPC bidder and the highest competing eCPM bid, which is a CPM bid. In this case, if $bq \geq A$, then the CPC bidder wins the auction and makes an expected payment of $\frac{pA}{q}$. But if $bq < A$, then the competing bidder wins the auction and makes a payment of bq.

Thus the expected revenue from predicting that an ad has a click-through rate of q when this ad actually has a click-through rate of p is $\frac{p}{q}\int_0^{bq} A\ g(A|b)\ dA + bq(1 - G(bq|b))$, and the appropriate loss function for a machine learning system that wishes to maximize expected revenue is equal to $\frac{p}{q}\int_0^{bq} A\ g(A|b)\ dA + bq(1 - G(bq|b)) - [\int_0^{bp} A\ g(A|b)\ dA + bp(1 - G(bp|b))]$. Differentiating this loss function with respect to q gives $-\frac{p}{q^2}\int_0^{bq} A\ g(A|b)\ dA + b^2p\ g(bp|b) + b(1 - G(bq|b)) - b^2q\ g(bq|b)$, and in the case where $G(\cdot|b)$ represents the uniform distribution

on $[0,1]$ for all b, this derivative reduces to $-\frac{p}{q^2}\int_0^{bq} A\ dA + b^2p + b(1-bq) - b^2q = \frac{b^2p}{2} - 2b^2q + b$, meaning the derivative is zero when $q = \frac{p}{4} + \frac{1}{2b}$.

But this example indicates that it is possible for the magnitude of the loss function to be minimized at a prediction q that is not equal to p. From this it follows that the appropriate loss function for a machine learning system that reflects the weighted efficiency and revenue loss from predicting that an ad has a click-through rate of q when this ad actually has a click-through rate of p may result in predictions that are not calibrated in the sense that the the magnitude of the expected value of the loss function may not be minimized when $q = p$. $\square$

Proof of Theorem 9: We know from the proof of Theorem 8 that the derivative of the loss function with respect to q is $-\frac{p}{q^2}\int_0^{bq} A\ g(A)\ dA + b^2p\ g(bp) + b(1 - G(bq)) - b^2q\ g(bq)$ when $G(\cdot) \equiv G(\cdot|b)$ and $g(\cdot) \equiv g(\cdot|b)$, meaning this derivative is $-\frac{1}{p}\int_0^{bp} A\ g(A)\ dA + b^2p\ g(bp) + b(1 - G(bp)) - b^2p\ g(bp) = b(1 - G(bp)) - \frac{1}{p}\int_0^{bp} A\ g(A)\ dA$ when $q = p$. Note that in the limit as $b \to 0$, the term $-\frac{1}{p}\int_0^{bp} A\ g(A)\ dA$ is $O(b^2)$ and the term $b(1 - G(bp))$ is $\Theta(b)$, so this derivative is positive for small values of b. From this it follows that the loss functions is optimized by making overpredictions of click-through rates for CPC ads with small bids.

Similarly, in the limit as bp approaches the upper bound of the support of $G(\cdot)$, $b(1-G(bp))$ approaches 0 and $\frac{1}{p}\int_0^{bp} A\ g(A)\ dA$ approaches $\frac{E[A]}{b}$. Thus in the limit as bp approaches the upper bound of the support of $G(\cdot)$, the derivative $b(1 - G(bp)) - \frac{1}{p}\int_0^{bp} A\ g(A)\ dA$ becomes negative. From this it follows that the loss function is optimized by making underpredictions of click-through rates for CPC ads with large bids, and the result then follows. $\square$

Proof of Theorem 10: In order for the value function to be unbiased, it must be the case that the value function is maximized for any p when $q = p$, so the derivative of the value function with respect to q must be zero when evaluated at $q = p$. From this it follows that it must be the case that $pu'(p) - c'(p) = 0$, so $c'(q) = qu'(q)$. By integrating both sides, it then follows that $c(q) = \int_0^q yu'(y)\ dy$. $\square$

Proof of Theorem 11: When $u(q) = q$, we have $u'(q) = 1$ and $c(q) = \int_0^q y\ dy = \frac{q^2}{2}$. Thus $V = pu(q) - c(q) = pq - \frac{q^2}{2}$, and maximizing V is equivalent to maximizing $-\frac{p^2}{2} + pq - \frac{q^2}{2} = -\frac{1}{2}(p-q)^2$. From this it is apparent that maximizing this value function is equivalent to minimizing mean squared error.

When $u(q) = \log(\frac{q}{1-q}) = \log(q) - \log(1-q)$, we have $u'(q) = \frac{1}{q} + \frac{1}{1-q}$ and $c(q) = \int_0^q yu'(y)\ dy = \int_0^q (1 + \frac{y}{1-y})\ dy = \int_0^q \frac{1}{1-y}\ dy = -\log(1-q)$. Thus $V = pu(q) - c(q) = p(\log(q) - \log(1-q)) + \log(1-q) = p\log(q) + (1-p)\log(1-q)$, and maximizing V is equivalent to minimizing the magnitude of the log likelihood loss function.

Finally, when $u(q) = bG(bq|b)$, we have $u'(q) = b^2 g(bq|b)$, and $c(q) = b^2 \int_0^q y\ g(by|b)\ dy$. Thus $V = bpG(bq|b) - b^2 \int_0^q y\ g(by|b)\ dy = bp - \int_{bq}^{\infty} bp\ g(A|b)\ dA - \int_0^{bq} A\ g(A|b)\ dA$, and maximizing V is equivalent to minimizing $\int_{bq}^{\infty} bp\ g(A|b)\ dA + \int_0^{bq} A\ g(A|b)\ dA$, which is in turn equivalent to minimizing $\int_{bq}^{\infty} bp\ g(A|b)\ dA + \int_0^{bq} A\ g(A|b)\ dA - [\int_{bp}^{\infty} bp\ g(A|b)\ dA + \int_0^{bp} A\ g(A|b)\ dA] = -\int_{bp}^{bq} (bp - A) g(A|b)\ dA$. From this it follows that maximizing V is equivalent to minimizing the magnitude of the empirical loss function in Theorem 1. □

Proof of Theorem 12: We have seen in the previous section that any unbiased value function must be of the form $V = pu(q) - c(q)$, where $c(q) = \int_0^q yu'(y)\ dy$. Furthermore, we have seen that the total expected value that arises from predicting that the click-through rate of the CPC ad is q when the actual click-through rate of this ad is p is $\sum_{k=1}^{s+1} [x_k bp + \sum_{j=1}^{k-1} x_j v_j + \sum_{j=k+1}^{s} x_j v_{j-1}] Pr(v_{k-1} > bq \geq v_k | q)$. This function varies linearly with p, and the coefficient on the term that gives the slope of how this value function varies with p is $\sum_{k=1}^{s+1} bx_k Pr(v_{k-1} > bq \geq v_k | q) = \sum_{k=1}^{s} bx_k Pr(v_{k-1} > bq \geq v_k | q)$, where the equality follows from the fact that $x_{s+1} = 0$. From this it follows that the $u(q)$ term in the value function of the form $V = pu(q) - c(q)$ must be $u(q) = \sum_{k=1}^{s} bx_k Pr(v_{k-1} > bq \geq v_k | q)$. The result then follows. □

Proof of Theorem 13: The appropriate value function if one wishes to maximize efficiency while using a value function that does not depend on the click-through rates of the advertiser is a value function that has the same expected value as the value function in Theorem 12

but does not depend on the unknown probability of a click p. Since the expected value of the value function $V = cu(q) - \int_0^q yu'(y)\ dy$ for any given predicted click-through rate q is $pu(q) - \int_0^q yu'(y)\ dy$, where p denotes the actual click-through rate of the ad, it follows that the value function given in Theorem 13 is the appropriate value function in this case. $\square$

Proof of Theorem 14: In order for L to be the appropriate loss function for a machine learning system that reflects the efficiency loss from predicting that an ad has a click-through rate of q, it must be the case that maximizing L is equivalent to maximizing the value function in Theorem 13 and $L \leq 0$ always holds. First note that $L \leq 0$ indeed always holds because for any $q \in [0, 1]$, $\int_c^q (c - y)u'(y)\ dy \leq 0$ regardless of whether $c = 1$ or $c = 0$.

Also note that $L = \int_c^q (c - y)u'(y)\ dy = \int_c^q cu'(y)\ dy - \int_c^q yu'(y)\ dy = c(u(q) - u(c)) - \int_c^q yu'(y)\ dy$. Thus maximizing L is equivalent to maximizing $c(u(q) - u(c)) - \int_c^q yu'(y)\ dy + cu(c) - \int_0^c yu'(y)\ dy = cu(q) - \int_0^q yu'(y)\ dy$, meaning that maximizing L is equivalent to maximizing the value function in Theorem 13. From this it follows that if one wishes to maximize efficiency while using a loss function that does not depend on the click-through rates of the advertiser, then the appropriate loss function for a machine learning system that reflects the efficiency loss from predicting that an ad has a click-through rate of q is $L = \int_c^q (c - y)u'(y)\ dy$. $\square$

Proof of Theorem 15: Differentiating the loss function in Theorem 1 with respect to q gives $b^2(p - q)g(bq|b)$, so the second derivative of the loss function in Theorem 1 with respect to q is $b^2[b(p - q)g'(bq|b) - g(bq|b)]$. Thus the empirical loss function is a concave function if and only if $b(p - q)g'(bq|b) - g(bq|b) \leq 0$, which holds if and only if $b(p - q) \leq \frac{g(bq|b)}{g'(bq|b)}$. But this expression will generally fail to hold for lognormal distributions in the limit as $q \to 0$. From this it follows that the empirical loss function in Theorem 1 is not a concave function in q if the highest competing bid that an advertiser faces is drawn from a lognormal distribution. $\square$

Proof of Theorem 16: If one uses the loss function $L(q,p)$, then this loss function will result in some distribution of predicted click-through rates given the actual click-through rates which we can model by the cumulative distribution function $F(q|p)$. The machine learning system will select distribution $F(q|p)$ amongst the set of feasible distributions that minimizes the magnitude of the expected loss. Now if $H(p)$ denotes the distribution of actual values of p in the population, this means that the machine learning system selects the distribution $F(q|p)$ amongst the set of feasible distributions that maximizes $\int_0^1 \int_0^1 L(q,p)\ dF(q|p)\ dH(p) = \int_0^1 \int_0^1 \frac{\partial L(q,p)}{\partial q}(1 - F(q|p))\ dq\ dH(p)$.

Now if one minimizes the magnitude of the true empirical loss function, then $\frac{\partial L(q,p)}{\partial q} = (p-q)g(q)$, and the machine learning system selects the distribution $F(q|p)$ amongst the set of feasible distributions that maximizes $\int_0^1 \int_0^1 (p-q)g(q)(1 - F(q|p))\ dq\ dH(p)$. Thus if one wishes to use a loss function that maximizes efficiency subject to the constraint that the loss function must be concave, then one should use a loss function $L(q,p)$ such that $L(q,p)$ is concave in q while $\frac{\partial L(q,p)}{\partial q}$ is as close as possible to $(p-q)g(q)$.

Now for values of q that are close to p, $(p-q)g(q)$ is necessarily decreasing in q, so the empirical loss function is concave in q for values of q near p. Thus the best concave loss function $L(q,p)$ will simply have derivative $\frac{\partial L(q,p)}{\partial q}$ that is equal to the derivative of the empirical loss function for values of q near p. And for values of q that are close to zero or one, $(p-q)g(q)$ is increasing in q, so the best concave loss function will instead have derivative $\frac{\partial L(q,p)}{\partial q}$ that is as close to $(p-q)g(q)$ as possible while still being nonincreasing in q, meaning $\frac{\partial L(q,p)}{\partial q}$ will be constant in q. The result then follows. $\square$

Proof of Theorem 17: Note that if we are fitting a model of the form model where q is of the form $q = \frac{1}{1+e^{-\sum_{i=1}^n \beta_i x_i}}$, then the loss function will be concave in the coefficients $\vec{\beta} = (\beta_1, \ldots, \beta_n)$ if and only if the loss function is concave in β when we are fitting a model of the form $q = \frac{1}{1+Ce^{-\beta x}}$ for all constants C. Thus in deriving the optimal loss function subject to the constraint that the loss function is concave in the coefficients $\vec{\beta} = (\beta_1, \ldots, \beta_n)$, it

suffices to derive the optimal loss function subject to the constraint that the loss function is concave in β when $q = \frac{1}{1+Ce^{-\beta x}}$.

Now when $q = \frac{1}{1+Ce^{-\beta x}}$, we have $\frac{\partial q}{\partial \beta} = \frac{Cxe^{-\beta x}}{(1+Ce^{-\beta x})^2}$, and $\frac{\partial^2 q}{\partial \beta^2} = \frac{-Cx^2e^{-\beta x}(1+Ce^{\beta x})^2+2C^2x^2(1+Ce^{\beta x})e^{-2\beta x}}{(1+Ce^{-\beta x})^4} = \frac{-Cx^2e^{-\beta x}(1+Ce^{\beta x})(1+Ce^{\beta x}-2Ce^{\beta x})}{(1+Ce^{-\beta x})^4} = \frac{Cx^2e^{-\beta x}(C^2e^{-2\beta x}-1)}{(1+Ce^{-\beta x})^4}$. From this it follows that $\frac{\partial^2 L(q,p)}{\partial \beta^2} = \frac{\partial^2 L(q,p)}{\partial q^2}(\frac{\partial q}{\partial \beta})^2 + \frac{\partial L(q,p)}{\partial q}\frac{\partial^2 q}{\partial \beta^2} = \frac{\partial^2 L(q,p)}{\partial q^2}\frac{C^2x^2e^{-2\beta x}}{(1+Ce^{-\beta x})^4} + \frac{\partial L(q,p)}{\partial q}\frac{Cx^2e^{-\beta x}(C^2e^{-2\beta x}-1)}{(1+Ce^{-\beta x})^4}$. This in turn implies that $\frac{\partial^2 L(q,p)}{\partial \beta^2} \leq 0$ if and only if $\frac{\partial^2 L(q,p)}{\partial q^2}Ce^{-\beta x} + \frac{\partial L(q,p)}{\partial q}(C^2e^{-2\beta x}-1) \leq 0$. Now since $q = \frac{1}{1+Ce^{-\beta x}}$, it follows that $Ce^{-\beta x} = \frac{1-q}{q}$ and $C^2e^{-2\beta x}-1 = \frac{1-2q}{q^2}$. Thus $\frac{\partial^2 L(q,p)}{\partial \beta^2} \leq 0$ if and only if $\frac{\partial^2 L(q,p)}{\partial q^2}\frac{1-q}{q} + \frac{\partial L(q,p)}{\partial q}\frac{1-2q}{q^2} \leq 0$, which in turn holds if and only if $\frac{\partial^2 L(q,p)}{\partial q^2}q(1-q) + \frac{\partial L(q,p)}{\partial q}(1-2q) \leq 0$.

Now we know by the reasoning in the proof of Theorem 16 that if one wishes to use a loss function that maximizes efficiency subject to the constraint that the loss function must be concave in the coefficients, then one should use a loss function $L(q,p)$ such that $L(q,p)$ is concave in its coefficients while $\frac{\partial L(q,p)}{\partial q}$ is as close as possible to $(p-q)g(q)$. From the results in the previous paragraph, it follows that this is equivalent to using a loss function $L(q,p)$ such that $\frac{\partial^2 L(q,p)}{\partial q^2}q(1-q) + \frac{\partial L(q,p)}{\partial q}(1-2q) \leq 0$ and $\frac{\partial L(q,p)}{\partial q}$ is as close as possible to $(p-q)g(q)$.

Now for values of q that are close to p, if $\frac{\partial L(q,p)}{\partial q} = (p-q)g(q)$, then $L(q,p)$ necessarily satisfies the constraint $\frac{\partial^2 L(q,p)}{\partial q^2}q(1-q) + \frac{\partial L(q,p)}{\partial q}(1-2q) \leq 0$, so the empirical loss function is concave in its coefficients for values of q near p. Thus the best loss function $L(q,p)$ that is concave in its coefficients will simply have derivative $\frac{\partial L(q,p)}{\partial q}$ that is equal to the derivative of the empirical loss function for values of q near p. And for values of q that are close to zero or one, if $\frac{\partial L(q,p)}{\partial q} = (p-q)g(q)$, then $L(q,p)$ will not satisfy the constraint $\frac{\partial^2 L(q,p)}{\partial q^2}q(1-q) + \frac{\partial L(q,p)}{\partial q}(1-2q) \leq 0$, so the best loss function $L(q,p)$ that is concave in its coefficients will instead have derivative that is as close to $(p-q)g(q)$ as possible while still satisfying the constraint $\frac{\partial^2 L(q,p)}{\partial q^2}q(1-q) + \frac{\partial L(q,p)}{\partial q}(1-2q) \leq 0$.

Now the above objective is achieved by choosing a loss function $L(q,p)$ that satisfies $\frac{\partial^2 L(q,p)}{\partial q^2}q(1-q) + \frac{\partial L(q,p)}{\partial q}(1-2q) = 0$ for values of q near zero and one. This is equivalent to choosing a loss function $L(q,p)$ that satisfies $\frac{\partial}{\partial q}[\frac{\partial L(q,p)}{\partial q}q(1-q)] = 0$, meaning the loss function $L(q,p)$ satisfies $\frac{\partial L(q,p)}{\partial q}q(1-q) = c$ for some constant c for values of q near zero and one. Thus the best concave loss function $L(q,p)$ will simply have derivative $\frac{\partial L(q,p)}{\partial q} = \frac{c}{q(1-q)} = \frac{c}{q} + \frac{c}{1-q}$

for values of q near zero and one (where the constant c may be different for values of q near zero than it is for values of q near one). The result then follows. $\square$

Proof of Theorem 18: In order for a loss function to be well-calibrated it must be the case that $qL_c'(q)+(1-q)L_n'(q)=0$ for all q. Thus if we let $f_c(q) \equiv L_c'(q)$ and we let $f_n(q) \equiv L_n'(q)$, then it must be the case that $qf_c(q)+(1-q)f_n(q)=0$ for all q, meaning $f_c(q)=-\frac{1-q}{q}f_n(q)$ and $f_c'(q)=\frac{f_n(q)}{q^2}-\frac{1-q}{q}f_n'(q)$.

Now in order for the loss functions $L_c(q)$ and $L_n(q)$ to be concave in the coefficients $\vec{\beta}=(\beta_1,\ldots,\beta_n)$, we know from the reasoning in the proof of Theorem 17 that it must be the case that $q(1-q)L_c''(q)+(1-2q)L_c'(q)\leq 0$ and $q(1-q)L_n''(q)+(1-2q)L_n'(q)\leq 0$. Thus if $f_c(q)=L_c'(q)$ and $f_n(q)=L_n'(q)$, then it also must be the case that $q(1-q)f_c'(q)+(1-2q)f_c(q)\leq 0$ and $q(1-q)f_n'(q)+(1-2q)f_n(q)\leq 0$. And since $f_c(q)=-\frac{1-q}{q}f_n(q)$ and $f_c'(q)=\frac{f_n(q)}{q^2}-\frac{1-q}{q}f_n'(q)$, the first of these inequalities is equivalent to $q(1-q)[\frac{f_n(q)}{q^2}-\frac{1-q}{q}f_n'(q)]-\frac{(1-2q)(1-q)}{q}f_n(q)\leq 0$, which is in turn equivalent to $2f_n(q)-(1-q)f_n'(q)\leq 0$.

Now $q(1-q)f_n'(q)+(1-2q)f_n(q)\leq 0$ holds if and only if $q(1-q)f_n'(q)+(1-2q)f_n(q)=-a(q)$ for some nonnegative function $a(q)$. This in turn holds if and only if $\frac{d}{dq}[q(1-q)f_n(q)]=-a(q)$, which then holds if and only if $q(1-q)f_n(q)=-b(q)$ for some non-negative and non-decreasing function $b(q)$. From this it follows that $f_n(q)$ must be of the form $f_n(q)=-\frac{b(q)}{q(1-q)}$ for some non-negative and non-decreasing function $b(q)$.

Now if $f_n(q)=-\frac{b(q)}{q(1-q)}$, then $f_n'(q)=-\frac{q(1-q)b'(q)-(1-2q)b(q)}{(q(1-q))^2}=-\frac{b'(q)}{q(1-q)}+\frac{(1-2q)b(q)}{q^2(1-q)^2}$. From this it follows that $2f_n(q)-(1-q)f_n'(q)\leq 0$ holds if and only if $-\frac{2b(q)}{q(1-q)}+\frac{b'(q)}{1-q}-\frac{(1-2q)b(q)}{q^2(1-q)}\leq 0$, which in turn holds if and only if $-2qb(q)+q(1-q)b'(q)-(1-2q)b(q)\leq 0 \Leftrightarrow q(1-q)b'(q)-b(q)\leq 0$.

Now let $h(q)\equiv\frac{b(q)}{q}$ so that $b(q)=qh(q)$. In this case, $b'(q)=h(q)+qh'(q)$, so the condition $q(1-q)b'(q)-b(q)\leq 0$ reduces to $-qh(q)+q(1-q)h(q)+q^2(1-q)h'(q)\leq 0$, which is in turn equivalent to $-h(q)+(1-q)h'(q)\leq 0$ or $h'(q)\leq\frac{h(q)}{1-q}$. At the same time, since $b(q)$ is non-decreasing in q, we know that $b'(q)\geq 0$, so the condition that $b'(q)=h(q)+qh'(q)$ implies that $h(q)+qh'(q)\geq 0$, meaning $h'(q)\geq-\frac{h(q)}{q}$.

Now since $f_n(q) = -\frac{b(q)}{q(1-q)}$, $f_c(q) = -\frac{1-q}{q} f_n(q)$, and $h(q) = \frac{b(q)}{q}$, it follows that $f_n(q) = -\frac{h(q)}{1-q}$ and $f_c(q) = \frac{h(q)}{q}$. And we have seen that if $f_n(q) = -\frac{b(q)}{q(1-q)}$, $f_c(q) = -\frac{1-q}{q} f_n(q)$, and $h(q) = \frac{b(q)}{q}$, then the loss functions will be concave in the coefficients $\vec{\beta} = (\beta_1, \ldots, \beta_n)$ if and only if $h'(q) \geq -\frac{h(q)}{q}$ and $h'(q) \leq \frac{h(q)}{1-q}$. By combining all the above analysis, we see that the set of feasible loss functions $L_c(q)$ and $L_n(q)$ for the losses that are incurred when one records a click or does not record a click are those satisfying $L_c'(q) = \frac{h(q)}{q}$ and $L_n'(q) = -\frac{h(q)}{1-q}$ for some non-negative function $h(q)$ satisfying $-\frac{h(q)}{q} \leq h'(q) \leq \frac{h(q)}{1-q}$. $\square$

Proof of Theorem 19: We know from the reasoning in the proof of Theorem 16 that if $L(q,p)$ denotes the expected loss that one incurs as a result of predicting that the click-through rate of an ad is q when the actual click-through rate of the ad is p, and one wishes to use a loss function that maximizes efficiency subject to the constraints given in the statement of Theorem 19, then one should use a loss function such that $\frac{\partial L(q,p)}{\partial q}$ is as close as possible to $(p-q)g(q)$ while satisfying these constraints. Now if p represents the actual click-through rate of an ad while q represents the predicted click-through rate of an ad, then the derivative of the expected loss that one incurs as a result of predicting a click-through rate of q with respect to q, $\frac{\partial L(q,p)}{\partial q}$, is $pL_c'(q) + (1-p)L_n'(q) = \frac{ph(q)}{q} - \frac{(1-p)h(q)}{1-q}$. Thus if one predicts a click-through rate q that is some fraction α of the true click-through rate p, then this derivative will be equal to $\frac{h(\alpha p)}{\alpha} - \frac{(1-p)h(\alpha p)}{1-\alpha p} = \frac{[1-\alpha p-\alpha(1-p)]h(\alpha p)}{\alpha(1-\alpha p)} = \frac{(1-\alpha)h(\alpha p)}{\alpha(1-\alpha p)}$ when $q = \alpha p$.

Now when $q = \alpha p$, we know that $(p-q)g(q) = (1-\alpha)pg(\alpha p)$. Thus $\frac{(1-\alpha)h(\alpha p)}{\alpha(1-\alpha p)} = (1-\alpha)pg(\alpha p)$ whenever $h(\alpha p) = \alpha p(1-\alpha p)g(\alpha p)$, which holds whenever $h(q) = q(1-q)g(q)$. For values of q where the derivative of $q(1-q)g(q)$ with respect to q is close to zero, the condition $-\frac{h(q)}{q} \leq h'(q) \leq \frac{h(q)}{1-q}$ that is necessary and sufficient for the loss functions $L_c(q)$ and $L_n(q)$ to satisfy the desired properties will automatically hold when $h(q) = q(1-q)g(q)$, so it will be optimal to set $h(q) = q(1-q)g(q)$ for such values of q.

And for values of q that are near zero and one, the condition $-\frac{h(q)}{q} \leq h'(q) \leq \frac{h(q)}{1-q}$ that is necessary and sufficient for the loss functions $L_c(q)$ and $L_n(q)$ to satisfy the desired properties will not be satisfied. When $h(q) = q(1-q)g(q)$, we have $h'(q) = (1-2q)g(q) + q(1-q)g'(q)$,

so $-\frac{h(q)}{q} \leq h'(q) \leq \frac{h(q)}{1-q}$ holds if and only if $-(1-q)g(q) \leq (1-2q)g(q)+q(1-q)g'(q) \leq qg(q)$, which in turn holds if and only if $(3q-2)g(q) \leq q(1-q)g'(q) \leq (3q-1)g(q)$. For values of q near zero and one, this will not be satisfied since $g'(q) \geq 0$ for values of q near 0 and $g'(q) \leq 0$ for values of q near 1.

Thus for values of q near zero and one, the optimal choice of the function $h(q)$ for the loss functions $L_c(q)$ and $L_n(q)$ satisfying $L_c'(q) = \frac{h(q)}{q}$ and $L_n'(q) = -\frac{h(q)}{1-q}$ will be such that $h(q)$ is as close to $q(1-q)g(q)$ while still satisfying the conditions $-\frac{h(q)}{q} \leq h'(q) \leq \frac{h(q)}{1-q}$. For values of q near zero this entails using a function $h(q)$ that satisfies $h'(q) = \frac{h(q)}{1-q}$, meaning $h(q)$ will be of the form $h(q) = \frac{c_0}{1-q}$ for some constant c_0 for values of q near zero. And for values of q near one this entails using a function $h(q)$ that satisfies $h'(q) = -\frac{h(q)}{q}$, meaning $h(q)$ will be of the form $h(q) = \frac{c_1}{q}$ for some constant c_1 for values of q near one. The result then follows. $\square$

Proof of Theorem 20: Note that the expected value that the decision maker obtains by fitting a model of the form $q = \phi x$ is $E[p^*(x)u(\phi x) - \int_0^{\phi x} yu'(y)\ dy]$. Differentiating this expression with respect to ϕ gives $E[p^*(x)u'(\phi x)x - \phi x u'(\phi x)x] = E[u'(\phi x)x(p^*(x) - \phi x)]$. Thus the optimal value of ϕ that is estimated in the limit when there is an infinite amount of training data satisfies $E[u'(\phi x)x(p^*(x)-\phi x)] = 0$, which is equivalent to $E[(\phi x)^{a-1}x(p^*(x)-\phi x)] = 0$. This in turn implies that the optimal value of ϕ satisfies $\phi = \frac{E[p^*(x)x^a]}{E[x^{a+1}]}$. $\square$

Proof of Theorem 21: If we fit a model that gives a predicted click-through rate q that satisfies $q = \phi x$ and the parameter ϕ is chosen to maximize the value function $V = pu(q) - c(q)$, where $u(q) = \frac{q^a}{a}$ and $c(q) = \frac{q^{a+1}}{a+1}$, then we know from Theorem 20 that the parameter ϕ that is chosen will satisfy $\phi = \frac{E[p^*(x)x^a]}{E[x^{a+1}]}$. Thus if the decision maker has a value function $V = pu(q) - c(q)$, where $u(q) = \frac{q^r}{r}$ and $c(q) = \frac{q^{r+1}}{r+1}$, then the decision maker's actual expected value from using this misspecified model is $V = E[p^*(x)u(\phi x) - c(\phi x)] = E[p^*(x)\frac{(\phi x)^r}{r} - \frac{(\phi x)^{r+1}}{r+1}] = E[\frac{p^*(x)x^r}{r}(\frac{E[p^*(x)x^a]}{E[x^{a+1}]})^r] - E[\frac{x^{r+1}}{r+1}(\frac{E[p^*(x)x^a]}{E[x^{a+1}]})^{r+1}]$.

Now if x is a random draw from the uniform distribution on $[0,1]$ and the true click-through rate of any given ad, $p^*(x)$, satisfies $p^*(x) = x^\gamma$, then the above expression for the

decision maker's value function simplifies to $V = \frac{1}{r(r+\gamma+1)}(\frac{a+2}{a+\gamma+1})^r - \frac{1}{(r+1)(r+2)}(\frac{a+2}{a+\gamma+1})^{r+1} = (\frac{a+2}{a+\gamma+1})^r[\frac{1}{r(r+\gamma+1)} - \frac{a+2}{(r+1)(r+2)(a+\gamma+1)}]$. Differentiating this expression with respect to a gives $\frac{dV}{da} = -(\frac{a+2}{a+\gamma+1})^{r-1}\frac{(a-r)(\gamma-1)^2}{(r+2)(a+\gamma+1)^3(\gamma+r+1)}$, which is zero if and only if $a = r$, negative if $a > r$, and positive if $a < r$. From this it follows that the decision maker's expected value is maximized when $a = r$, and the decision maker's expected value is also decreasing in the distance from a to r for both values of $a < r$ and values of $a > r$. $\square$

Proof of Theorem 22: (1). If the loss function being minimized is mean squared error, then we seek to minimize $E[(q-p)^2]$. Since the true click-through rate of the ad, p, satisfies $p^*(x) = x^\gamma$ and we are fitting a model of the form $q = \phi x$ for some parameter ϕ, this in turn implies that we seek to minimize $E[(\phi x - x^\gamma)^2]$. Differentiating this expression with respect to ϕ then gives $E[2\phi x^2 - 2x^{\gamma+1}]$, meaning the derivative of the loss function with respect to ϕ is zero if and only if $E[\phi x^2 - x^{\gamma+1}] = 0$. From this it follows that if the loss function is mean squared error, then the parameter ϕ that is estimated satisfies $E[\phi x^2 - x^{\gamma+1}] = 0$.

(2). If the loss function being minimized is the log likelihood loss function, then we seek to maximize $E[p\log(q) + (1-p)\log(1-q)]$. Since the true click-through rate of the ad, p, satisfies $p^*(x) = x^\gamma$ and we are fitting a model of the form $q = \phi x$ for some parameter ϕ, this in turn implies that we seek to maximize $E[x^\gamma \log(\phi x) + (1-x^\gamma)\log(1-\phi x)]$. Differentiating this expression with respect to ϕ gives $E[\frac{x^\gamma}{\phi} - \frac{x(1-x^\gamma)}{1-\phi x}] = E[\frac{x^\gamma - \phi x^{\gamma+1} - \phi x + \phi x^{\gamma+1}}{\phi(1-\phi x)}] = E[\frac{x^\gamma - \phi x}{\phi(1-\phi x)}]$, meaning this derivative is zero if and only if $E[\frac{x^\gamma - \phi x}{1-\phi x}] = 0$. From this it follows that if there is some parameter $\phi \in (0,1)$ that satisfies $E[\frac{x^\gamma - \phi x}{1-\phi x}] = 0$, then this parameter will minimize the loss function.

If there is no such parameter ϕ, then it must be the case that $E[\frac{x^\gamma - \phi x}{1-\phi x}] > 0$ for all $\phi \in (0,1)$ since $E[\frac{x^\gamma - \phi x}{1-\phi x}] > 0$ for values of ϕ arbitrarily close to zero and $E[\frac{x^\gamma - \phi x}{1-\phi x}]$ is continuous in ϕ. But if $E[\frac{x^\gamma - \phi x}{1-\phi x}] > 0$ for all $\phi \in (0,1)$, then it must be the case that the function $E[x^\gamma \log(\phi x) + (1-x^\gamma)\log(1-\phi x)]$ is maximized for values of $\phi \in (0,1]$ when $\phi = 1$. Thus if there is no parameter $\phi \in (0,1)$ that satisfies $E[\frac{x^\gamma - \phi x}{1-\phi x}] = 0$, then $\phi = 1$.

(3). If the loss function being optimized is the empirical loss function in Theorem 1, then we seek to maximize $E[\int_p^q (p-A)g(A)\,dA]$, where $g(A)$ denotes the probability density function corresponding to the beta distribution with parameters α and β. Since the true click-through rate of the ad, p, satisfies $p^*(x) = x^\gamma$ and we are fitting a model of the form $q = \phi x$ for some parameter ϕ, this in turn implies that we seek to maximize $E[\int_{x^\gamma}^{\phi x} (x^\gamma - A)g(A)\,dA]$. Differentiating this expression with respect to ϕ then gives $E[x(x^\gamma - \phi x)g(\phi x)]$, meaning the derivative of the loss function with respect to ϕ is zero if and only if $E[x(x^\gamma - \phi x)g(\phi x)] = 0$. From this it follows that if the loss function is the empirical loss function in Theorem 1, then the parameter ϕ that is estimated satisfies $E[x(x^\gamma - \phi x)g(\phi x)] = 0$. $\square$

Proof of Theorem 23: If the loss function is mean squared error, then we know from Theorem 22 that the parameter ϕ that is estimated satisfies $E[\phi x^2 - x^{\gamma+1}] = 0$ or $E[x(x^\gamma - \phi x)] = 0$. Since the true click-through rate of the ad satisfies $p = x^\gamma$, it then follows that the parameter ϕ that is estimated also satisfies $E[p^{1/\gamma}(p - \phi p^{1/\gamma})] = 0$. Now in the limit as $\gamma \to \infty$, $p^{1/\gamma} \to 1$ for almost all values of $p \in [0,1]$, so the solution ϕ to the equation $E[p^{1/\gamma}(p - \phi p^{1/\gamma})] = 0$ approaches the solution to the equation $E[p - \phi] = 0$. Thus in the limit as $\gamma \to \infty$, the parameter ϕ that is estimated for the mean squared error loss function in Theorem 22 approaches $\phi = E[p]$.

If the loss function is the log likelihood loss function, then the parameter ϕ that is estimated is the $\phi \in (0,1)$ that satisfies $E[\frac{x^\gamma - \phi x}{1-\phi x}] = 0$. Since the true click-through rate of the ad satisfies $p = x^\gamma$, it then follows that the parameter ϕ that is estimated also satisfies $E[\frac{p - \phi p^{1/\gamma}}{1-\phi p^{1/\gamma}}] = 0$. Now in the limit as $\gamma \to \infty$, $p^{1/\gamma} \to 1$ for almost all values of $p \in [0,1]$, so the solution ϕ to the equation $E[\frac{p - \phi p^{1/\gamma}}{1-\phi p^{1/\gamma}}] = 0$ approaches the solution to the equation $E[\frac{p-\phi}{1-\phi}] = 0$, which is equivalent to $E[p - \phi] = 0$. Thus in the limit as $\gamma \to \infty$, the parameter ϕ that is estimated for the log likelihood loss function considered in Theorem 22 approaches $\phi = E[p]$.

Finally, if the loss function is the empirical loss function, then the parameter ϕ that is estimated satisfies $E[x(x^\gamma - \phi x)g(\phi x)] = 0$. Since $E[x(x^\gamma - \phi x)g(\phi x)]$ is proportional to

$E[x^\alpha(x^\gamma - \phi x)(1-\phi x)^{\beta-1}]$ for values of $\phi x \in [0,1]$, it then follows that the parameter ϕ that is estimated also satisfies $E[x^\alpha(x^\gamma - \phi x)(1-\phi x)^{\beta-1}] = 0$. Since the true click-through rate of the ad satisfies $p = x^\gamma$, it then follows that the parameter ϕ that is estimated also satisfies $E[p^{\alpha/\gamma}(p - \phi p^{1/\gamma})(1-\phi p^{1/\gamma})^{\beta-1}] = 0$. Now in the limit as $\gamma \to \infty$, $p^{1/\gamma} \to 1$ and $p^{\alpha/\gamma} \to 1$ for almost all values of $p \in [0,1]$, so the solution ϕ to the equation $E[p^{\alpha/\gamma}(p - \phi p^{1/\gamma})(1-\phi p^{1/\gamma})^{\beta-1}] = 0$ approaches the solution to the equation $E[(p-\phi)(1-\phi)^{\beta-1}] = 0$, which is equivalent to $E[p - \phi] = 0$. Thus in the limit as $\gamma \to \infty$, the parameter ϕ that is estimated for the empirical loss function considered in Theorem 22 approaches $\phi = E[p]$. $\square$

Proof of Theorem 24: (1). If the loss function being used is the log likelihood loss function, then we seek to maximize $E[p \log q + (1-p)\log(1-q)] = E[x^\gamma \log(\frac{1}{1+e^{-\phi_1-\phi_2 x}}) + (1 - x^\gamma)\log(\frac{1}{1+e^{\phi_1+\phi_2 x}})]$. Differentiating this loss function with respect to ϕ_1 gives $E[\frac{x^\gamma}{1+e^{\phi_1+\phi_2 x}} - \frac{1-x^\gamma}{1+e^{-\phi_1-\phi_2 x}}]$ and differentiating this loss function with respect to ϕ_2 gives $E[\frac{x^{\gamma+1}}{1+e^{\phi_1+\phi_2 x}} - \frac{x(1-x^\gamma)}{1+e^{-\phi_1-\phi_2 x}}]$. From this it follows that if the loss function is the log likelihood loss function, then the parameters ϕ_1 and ϕ_2 that are estimated will be solutions to the set of equations $E[\frac{x^\gamma}{1+e^{\phi_1+\phi_2 x}} - \frac{1-x^\gamma}{1+e^{-\phi_1-\phi_2 x}}] = 0$ and $E[\frac{x^{\gamma+1}}{1+e^{\phi_1+\phi_2 x}} - \frac{x(1-x^\gamma)}{1+e^{-\phi_1-\phi_2 x}}] = 0$.

(2). Since our proof of this part will make use of some preliminary result first derived in our proof of part (3), we first prove part (3) and then return to prove part (2).

(3). Note that for values of q where the function $h(q)$ being used in the loss functions $L_c'(q) = \frac{h(q)}{q}$ and $L_n'(q) = \frac{-h(q)}{1-q}$ satisfies $h(q) = \frac{c_1}{1-q}$ for some constant c_1, then $L_c'(q) = c_1(\log q - \log(1-q))$ and $L_n'(q) = -\frac{c_1}{1-q}$ (up to a constant), so the expected value of the loss function is $c_1 E[p\log(\frac{q}{1-q}) - \frac{1-p}{1-q}] = c_1 E[p\log(\frac{q}{1-q}) - \frac{1-p}{1-q}] = c_1 E[x^\gamma \log(e^{\phi_1+\phi_2 x}) - (1 - x^\gamma)(e^{\phi_1+\phi_2 x}+1)] = c_1 E[x^\gamma(\phi_1+\phi_2 x) - (1-x^\gamma)(e^{\phi_1+\phi_2 x}+1)]$ (up to a constant). Thus for values of q in this range, the derivative of the expected loss function with respect to ϕ_1 is $c_1 E[x^\gamma(1+e^{\phi_1+\phi_2 x}) - 1]$ and the derivative of the expected loss function with respect to ϕ_2 is $c_1 E[x^{\gamma+1}(1+e^{\phi_1+\phi_2 x}) - x]$.

Similarly, for values of q where the function $h(q)$ being used in the loss functions $L_c'(q) = \frac{h(q)}{q}$ and $L_n'(q) = \frac{-h(q)}{1-q}$ satisfies $h(q) = \frac{c_2}{q}$ for some constant c_2, then $L_c(q) = -\frac{c_2}{q}$ and

$L_n(q) = c_2 \log(\frac{1-q}{q})$ (up to a constant), so the expected value of the loss function is $c_2 E[-\frac{p}{q} + (1-p)\log(\frac{1-q}{q})] = c_2 E[-x^\gamma(1+e^{-\phi_1-\phi_2 x}) + (1-x^\gamma)(-\phi_1-\phi_2 x)]$ (up to a constant). Thus for values of q in this range, the derivative of the expected loss function with respect to ϕ_1 is $c_2 E[x^\gamma(1+e^{-\phi_1-\phi_2 x}) - 1]$ and the derivative of the expected loss function with respect to ϕ_2 is $c_2 E[x^{\gamma+1}(1+e^{-\phi_1-\phi_2 x}) - x]$.

And for values of q where the function $h(q)$ being used in the loss functions $L_c'(q) = \frac{h(q)}{q}$ and $L_n'(q) = \frac{-h(q)}{1-q}$ satisfies $h(q) = q(1-q)g(q)$, then the derivative of the expected loss function with respect to ϕ_1 is $E[pL_c'(q)\frac{dq}{d\phi_1} + (1-p)L_n'(q)\frac{dq}{d\phi_1}] = E[p(1-q)g(q)\frac{e^{\phi_1+\phi_2 x}}{(1+e^{\phi_1+\phi_2 x})^2} - (1-p)qg(q)\frac{e^{\phi_1+\phi_2 x}}{(1+e^{\phi_1+\phi_2 x})^2}]$, which is proportional to $E[pq^{\alpha-1}(1-q)^\beta \frac{e^{-\phi_1-\phi_2 x}}{(1+e^{-\phi_1-\phi_2 x})^2} - (1-p)q^\alpha(1-q)^{\beta-1}\frac{e^{-\phi_1-\phi_2 x}}{(1+e^{-\phi_1-\phi_2 x})^2}] = E[x^\gamma \frac{e^{-(\beta+1)(\phi_1+\phi_2 x)}}{(1+e^{-\phi_1-\phi_2 x})^{\alpha+\beta+1}} - (1-x^\gamma)\frac{e^{-\beta(\phi_1+\phi_2 x)}}{(1+e^{-\phi_1-\phi_2 x})^{\alpha+\beta+1}}] = E[x^\gamma \frac{e^{\alpha(\phi_1+\phi_2 x)}}{(1+e^{\phi_1+\phi_2 x})^{\alpha+\beta+1}} - (1-x^\gamma)\frac{e^{(\alpha+1)(\phi_1+\phi_2 x)}}{(1+e^{\phi_1+\phi_2 x})^{\alpha+\beta+1}}]$. A similar argument shows that the derivative of the expected loss function with respect to ϕ_2 for values of q in this region is proportional to $E[x^{\gamma+1}\frac{e^{\alpha(\phi_1+\phi_2 x)}}{(1+e^{\phi_1+\phi_2 x})^{\alpha+\beta+1}} - x(1-x^\gamma)\frac{e^{(\alpha+1)(\phi_1+\phi_2 x)}}{(1+e^{\phi_1+\phi_2 x})^{\alpha+\beta+1}}]$ (with the same proportionality constant).

Now let x_L denote the value of x at which $q_L = \frac{1}{1+e^{-\phi_1-\phi_2 x}}$. We then have $q_L = \frac{1}{1+e^{-\phi_1-\phi_2 x_L}}$, which holds if and only if $1 + e^{-\phi_1-\phi_2 x_L} = \frac{1}{q_L} \Leftrightarrow e^{-\phi_1-\phi_2 x_L} = \frac{1}{q_L} - 1 = \frac{1-q_L}{q_L} \Leftrightarrow e^{\phi_1+\phi_2 x_L} = \frac{q_L}{1-q_L} \Leftrightarrow \phi_1 + \phi_2 x_L = \log(\frac{q_L}{1-q_L}) \Leftrightarrow x_L = \frac{\log(\frac{q_L}{1-q_L})-\phi_1}{\phi_2}$. Similarly, if x_H denote the value of x at which $q_H = \frac{1}{1+e^{-\phi_1-\phi_2 x}}$, then $x_H = \frac{\log(\frac{q_H}{1-q_H})-\phi_1}{\phi_2}$. From this it follows that the function $h(q)$ being used in the loss functions $L_c'(q) = \frac{h(q)}{q}$ and $L_n'(q) = \frac{-h(q)}{1-q}$ will satisfy $h(q) = \frac{c_1}{1-q}$ for some constant c_1 when $x \leq x_L$, this function will satisfy $h(q) = \frac{c_2}{q}$ for some constant c_2 when $x \geq x_H$, and the function will satisfy $h(q) = q(1-q)g(q)$ when $x \in (x_L, x_H)$ for the values of x_L and x_H satisfying $x_L = \frac{\log(\frac{q_L}{1-q_L})-\phi_1}{\phi_2}$ and $x_H = \frac{\log(\frac{q_H}{1-q_H})-\phi_1}{\phi_2}$.

Now the constants c_1 and c_2 must be chosen in such a way that the derivatives of the loss functions $L_c'(q)$ and $L_n'(q)$ are continuous in q at q_L and q_H. In the third paragraph of the proof this part, we have noted what the derivative of the expected loss function with respect would be when the function $h(q)$ being used in the loss functions $L_c'(q) = \frac{h(q)}{q}$ and $L_n'(q) = \frac{-h(q)}{1-q}$ satisfies $h(q) = q^\alpha(1-q)^\beta$ (which is proportional to $h(q) = q(1-q)g(q)$). In order for the derivatives of the loss functions $L_c'(q)$ and $L_n'(q)$ to be continuous in q at q_L, it must then be the case that the constant c_1 satisfies $\frac{c_1}{1-q} = q^\alpha(1-q)^\beta$, meaning

$c_1 = q^{\alpha}(1-q)^{\beta+1}$. Similarly, in order for the loss functions $L_c'(q)$ and $L_n'(q)$ to be continuous in q at q_H, it must then be the case that the constant c_2 satisfies $\frac{c_2}{q} = q^{\alpha}(1-q)^{\beta}$, meaning $c_2 = q^{\alpha+1}(1-q)^{\beta}$.

But this then implies that if $n_L = q_L^{\alpha}(1-q_L)^{\beta+1}$ and $n_H = q_H^{\alpha+1}(1-q_H)^{\beta}$, then the derivative of the expected loss function with respect to ϕ_1 is equal to n_L times the expectation given in the first paragraph of the proof of this part given that $x \leq x_L$ times the probability that $x \leq x_L$ plus n_H times the expectation given in the second paragraph of the proof of this part given that $x \geq x_H$ times the probability that $x \geq x_H$ plus the expectation given in the third paragraph of the proof of this part given that $x \in (x_L, x_H)$ times the probability that $x \in (x_L, x_H)$. A similar result holds for the derivative of the expected loss function with respect to ϕ_2. Since the parameters will be estimated are parameters where both of these derivatives are equal to zero, it then follows that the parameters ϕ_1 and ϕ_2 that are estimated will be solutions to the set of equations given in the statement of the theorem.

(2). We now return to proving part (2) of this theorem. Note that the empirical loss function only differs from the concave approximation to the empirical loss function in that the function $h(q)$ being used in the loss functions $L_c'(q) = \frac{h(q)}{q}$ and $L_n'(q) = \frac{-h(q)}{1-q}$ satisfies $h(q) = q(1-q)g(q)$ for all q. Thus from the reasoning in the third paragraph of the proof of part (3) of this theorem, it follows that the derivative of expected loss function with respect to ϕ_1 is proportional to $E[x^{\gamma}\frac{e^{\alpha(\phi_1+\phi_2 x)}}{(1+e^{\phi_1+\phi_2 x})^{\alpha+\beta+1}} - (1-x^{\gamma})\frac{e^{(\alpha+1)(\phi_1+\phi_2 x)}}{(1+e^{\phi_1+\phi_2 x})^{\alpha+\beta+1}}]$ and the derivative of the expected loss function with respect to ϕ_2 is proportional to $E[x^{\gamma+1}\frac{e^{\alpha(\phi_1+\phi_2 x)}}{(1+e^{\phi_1+\phi_2 x})^{\alpha+\beta+1}} - x(1-x^{\gamma})\frac{e^{(\alpha+1)(\phi_1+\phi_2 x)}}{(1+e^{\phi_1+\phi_2 x})^{\alpha+\beta+1}}]$. Thus if the loss function is the empirical loss function, then the parameters ϕ_1 and ϕ_2 that are estimated will be solutions to the set of equations $E[\frac{x^{\gamma}e^{\alpha(\phi_1+\phi_2 x)}}{(1+e^{\phi_1+\phi_2 x})^{\alpha+\beta+1}} - \frac{(1-x^{\gamma})e^{(\alpha+1)(\phi_1+\phi_2 x)}}{(1+e^{-\phi_1-\phi_2 x})^{\alpha+\beta+1}}] = 0$ and $E[\frac{x^{\gamma+1}e^{\alpha(\phi_1+\phi_2 x)}}{(1+e^{\phi_1+\phi_2 x})^{\alpha+\beta+1}} - \frac{x(1-x^{\gamma})e^{(\alpha+1)(\phi_1+\phi_2 x)}}{(1+e^{-\phi_1-\phi_2 x})^{\alpha+\beta+1}}] = 0$. □