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

PATRICK HUMMEL* R. PRESTON MCAFEE**

FEBRUARY 27, 2017

ABSTRACT. We characterize the optimal loss functions for predicted click-through rates in auctions for online advertising. While standard loss functions such as mean squared error or log likelihood severely penalize large mispredictions while imposing little penalty on smaller mistakes, a loss function reflecting the true economic loss from mispredictions imposes significant penalties for small mispredictions and only slightly larger penalties on large mispredictions. We illustrate that when the model is misspecified, using such a loss function can improve economic efficiency, but the efficiency gain is likely to be small.

JEL Classification: C10; C13; C53; C55; C61; D44

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, mean squared error (MSE) and log likelihood (LL) 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 (pCTRs) in Internet advertisement auctions that reflect the true consequences of estimation errors.

*Corresponding author. Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA; *E-mail address*: phummel@google.com; *Phone*: (408) 865-1187; *Fax*: (408) 856-1187

**Microsoft Corp., One Microsoft Way, Redmond, WA 98052, USA.

Auctions for Internet advertising are used in auctions for ads on search engine result pages, like those of Bing, Google, and Yahoo!, as well as display advertising auctions run by Google (AdX and AdSense), Facebook (FBX), AppNexus, OpenX and others for a wide range of web publishers. In most of these auctions, an advertiser only pays if a user clicks on the advertiser's ad, which is known as cost-per-click (CPC) pricing, in contrast to cost-per-impression (CPM) advertising. In order to identify the most valuable ad, it is necessary to forecast the probability that an individual will click on the ad. For example, if one ad has a CPC bid of \$2, and another ad has a CPC bid of \$4, then 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.

Under CPC pricing, the probability of a click needs to be estimated. This is typically done by using machine learning to fit models involving hundreds of millions or even billions of categorical (0/1) variables.[1] These variables, which are commonly called features, reflect aspects of the ad, the page on which the ad will be shown, information about the user, and interactions betweeen these terms that may influence click-through rates (CTRs). Moreover, a history involving billions of data points per day on users visiting pages with advertisements is available as training data (McMahan *et al.* 2013). The standard technique for using this information to estimate click probabilities is to fit a model using logistic regression (Hilbe 2009). We propose to change this by tailoring the objective function to match the economic losses that result from misestimates in the cases where the estimates are employed.

The main reason to use a correctly specified loss function is to improve performance under misspecification. When the model giving the dependence of probabilities on the covariates is misspecified, even in the limit of an arbitrarily large amount of training 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. Since the models used in online auctions are likely to be misspecified,

[1]For instance, McMahan (2013) describes techniques developed by Google to predict the click-through rates of ads using models "with billions of coefficients" and Yahoo! notes that optimizing its display advertising inventory requires solving problems "involving tens to hundreds of millions of variables and hundreds of thousands to millions of constraints" (Yahoo! Labs 2013).

choosing the loss function to match the actual economic losses from misestimates may 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 mispredict the CTR of an ad, the auction will sometimes run an ad that is not best, thereby incurring a loss in efficiency. The efficiency loss is determined by whether alternative ads are close, and as a result, the distribution of alternatives plays a key role in the analysis. 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 \$0.02. 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 pCTRs is used. Thus, in contrast to LL or MSE, the empirical loss function only imposes slightly larger penalties on mispredictions beyond a certain level of inaccuracy.

There are two significant properties of both MSE and LL. Both are calibrated in the sense that they are minimized by predicting a CTR equal to the actual CTR, and they are convex, which ensures that an iterative process like gradient descent can be used to find the minimum loss. The empirical loss function is also calibrated but it is not convex. Because of the scale of the estimation problem, it would be 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 empirical loss function can improve economic performance under misspecified models. 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. We illustrate that using the empirical loss function rather than LL can indeed improve economic efficiency even when one has an

arbitrarily large amount of training data through some simulations on misspecified models. However, our simulations indicate that the efficiency gain is likely to be small.

The most closely related work to our paper is Chapelle (2015), which was completed after our paper was first circulated. In Chapelle (2015), the author computes the empirical loss function given in our paper using traffic logs from Criteo. Chapelle (2015) notes that prediction losses computed using the empirical loss function better correlate with economic outcomes than the prediction losses computed using MSE.

Our paper also relates to several other distinct strands of literature. First, our paper relates to the extensive literature in economics, machine learning, and statistics on loss functions (Arrow 1959; Bartlett *et al.* 2006; Denuit and Dhaene 2001; Dmochowski *et al.* 2010; Elliott *et al.* 2005; Elliott and Lieli 2013; Manski 2004; Patton and Timmerman 2007; Reid and Williamson 2010; 2011; Shalit and Yitzhaki 2002; Skalak *et al.* 2007; Steinwart 2007; Weiss 1996; Zhang 2004). Most of these papers focus on general theoretical questions related to loss functions, and none considers the best loss function for online advertising auctions.

Some existing work has shown that there can be value to choosing an economically-motivated loss function in other settings. For instance, Basu and Markov (2004) illustrates that using linear loss functions instead of quadratic loss can better account for the earnings forecasts made by analysts. Boudt and Croux (2004) shows how downweighting outliers in multivariate GARCH models better captures the volatility associated with major stock market downturns and Shalit and Yitzhaki (2002) shows how introducing risk aversion to the estimation procedure can reduce the sensitivity of estimators of beta coefficients for major firms to extreme observations. Elliott *et al.* (2005) notes that allowing for asymmetric loss functions can better account for IMF and OECD forecasts of budget deficits and Patton and Timmerman (2007) similarly notes that asymmetric loss functions better account for the Federal Reserve's forecasts of output growth. And Lieli and White (2010) uses the framework in Elliott and Lieli (2013) to show how a lender can achieve greater profits by using an estimation method based on the lender's objectives in making credit-approval decisions.[2]

[2]In addition, Cohen *et al.* (2003) empirically measures the loss functions for suppliers of semiconductor equipment by considering cancellation, holding, and delay costs.

Relative to this previous literature, our work differs in several ways. We consider forecasts of probabilities, whereas these previous papers all consider point estimates of outcomes. We also derive a convex approximation to the economically relevant loss function so that using our loss function will be computationally feasible, whereas this consideration does not arise in any of these previous papers. In addition, we apply our framework to a high-dimensional problem that would not be feasible in many of these previous frameworks, such as the Elliott and Lieli (2013) framework. Finally, our paper is the first to consider the question of the best loss function to use in the context of online auctions for advertising.

2. The Model

In each auction, a machine learning system predicts a CTR for an advertiser who has submitted a CPC bid into an online ad auction where advertisers are ranked by expected cost-per-impression (eCPM). Thus if this CPC bidder submits a CPC bid of b and the machine learning system predicts the ad's CTR is q, then this advertiser's eCPM bid is bq. The CTRs are all for a particular context, and thus may vary from auction to auction.

While the machine learning system predicts a CTR of q, the actual CTR may differ from this. We let p denote the actual CTR of the CPC ad in question. This probability may depend on both observed features of the auction x, such as features of the publisher or the user, and unobserved features ϵ, so we write $p = \tilde{p}(x, \epsilon)$ for some function $\tilde{p}(\cdot)$. We also assume that the highest competing eCPM bid is a random draw from the distribution $G(\cdot|b, x)$ with corresponding probability density function $g(\cdot|b, x)$.

The dependence of $G(\cdot|b, x)$ on b and x indicates that the highest competing eCPM bid may be correlated with b and x, which further implies that the highest competing eCPM bid may be correlated with p. However, we do assume that $G(\cdot|b, x)$ is independent of p ($G(\cdot|b, x, p) = G(\cdot|b, x)$) so that for any fixed features of the auction x, also knowing the CTR for the CPC advertiser p would not enable one to better predict the highest competing eCPM bid A. This effectively implies that the unobserved features ϵ also do not influence the competing bids because they are universally unobserved.

Throughout we let A denote the highest competing eCPM bid that the advertiser faces. The value of the highest competing bid will not generally be known at the time we predict the CTR of the CPC bidder.[3] However, one can estimate the distribution of highest competing bids by analyzing past auctions, and then use this distribution $G(\cdot)$ to construct the loss function. Formulating the highest competing bid as an eCPM bid allows the highest competing bidder to be bidding on either a CPM or CPC basis, which is important since CPC bidders often compete against CPM bidders in online advertising auctions.

3. Preliminaries

We first address the question of the appropriate choice of loss function when one wishes to maximize economic efficiency, or the true eCPM value of the ad that is displayed. Given our estimate of the pCTR 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 surplus is bp, where p denotes the actual CTR of the CPC ad, but in the second case the total surplus is A. Thus the expected surplus generated when the machine learning system predicts a CTR of q is $bpPr(bq \geq A) + E[A|bq < A]Pr(bq < A)$.[4]

To derive an appropriate loss function, we must compare the expected surplus generated by using a pCTR of q with the expected surplus under perfect prediction. We refer to this as the *empirical loss function* because it reflects the true empirical loss that results from misestimates of the pCTRs. We derive the appropriate such loss function in Theorem 1:

Theorem 1. *Suppose that one wishes to maximize economic efficiency. Then the correct loss function from using a pCTR of q when the actual CTR is p is $\int_{bp}^{bq}(bp - A)g(A|b, x)\, dA$.*

[3]For example, when AdSense bids on behalf of Google's advertisers for advertising opportunities on an ad exchange (AdX), Google's advertisers bid on a CPC basis, but the AdX auction is conducted on a CPM basis and Google must submit CPM bids on behalf of its CPC advertisers without knowing what the competing bids from other ad networks will be. In order to do this, Google must predict a CTR for the CPC bidder and use that to convert the CPC bidder's bid to a CPM bid for the AdX auction.

[4]If the highest competing ad is a CPC bidder with an uncertain CTR, then the actual efficiency resulting from showing such an ad may differ from A. However, as long as the pCTRs are unbiased on average, the expected efficiency that would arise would be A, and all the analysis in our paper would continue to hold.

Proof. The expected surplus that results from using a pCTR of q when the CTR of the ad is p is $bpPr(bq \geq A) + E[A|bq < A]Pr(bq < A) = \int_{bq}^{\infty} A\ g(A|b,x)\ dA + \int_{0}^{bq} bp\ g(A|b,x)\ dA$. Thus the expected surplus that results from correctly predicting that the CTR of an ad is p is $\int_{bp}^{\infty} A\ g(A|b,x)\ dA + \int_{0}^{bp} bp\ g(A|b,x)\ dA$. From this it follows that the efficiency loss from using a pCTR of q when the CTR of the ad is p is $\int_{bq}^{\infty} A\ g(A|b,x)\ dA + \int_{0}^{bq} bp\ g(A|b,x)\ dA - [\int_{bp}^{\infty} A\ g(A|b,x)\ dA + \int_{0}^{bp} bp\ g(A|b,x)\ dA] = \int_{bp}^{bq}(bp - A)g(A|b,x)\ dA$. □

In most machine learning systems it is standard to use loss functions such as MSE or LL. Given the result in Theorem 1, it is natural to ask whether these standard loss functions are compatible with the empirical loss function. Our next result illustrates that there are some distributions of competing eCPM bids such that the empirical loss function given in Theorem 1 will be compatible with MSE.

Example 1. *Suppose that the highest competing eCPM bid is drawn from a uniform distribution that is independent of b. Then the empirical loss function is equivalent to MSE because* $\int_{bp}^{bq}(bp - A)g(A|b,x)\ dA$ *is proportional to* $\int_{bp}^{bq}(bp - A)\ dA = bp(bq - bp) - \frac{(bq)^2-(bp)^2}{2} = -\frac{(bq)^2-2(bq)(bp)+(bp)^2}{2} = -\frac{b^2}{2}(q-p)^2$*, which is equivalent to MSE.*

While minimizing the empirical loss function is equivalent to minimizing MSE when the highest competing eCPM bid is drawn from a uniform distribution, this is not the case for other distributions. Empirically the uniform distribution is a poor representation of the distribution of competing eCPM bids, as Lahaie and McAfee (2011), Ostrovsky and Schwarz (2016), and Sun *et al.* (2014) have noted that these distributions are better modeled by a log-normal distribution in sponsored search auctions on Yahoo! and Baidu.

Under more realistic distributions of the highest competing eCPM bid, the empirical loss function will no longer be equivalent to either MSE or LL. This can be readily seen in Figure 1, where we plot MSE (in a solid black line), LL (in long red dotted lines), and the empirical loss function (in short blue dotted lines) that results from using a pCTR of q when the actual CTR of the ad is $p = 0.019$ and the highest competing eCPM bid is drawn from a log-normal distribution with parameters μ and σ^2.

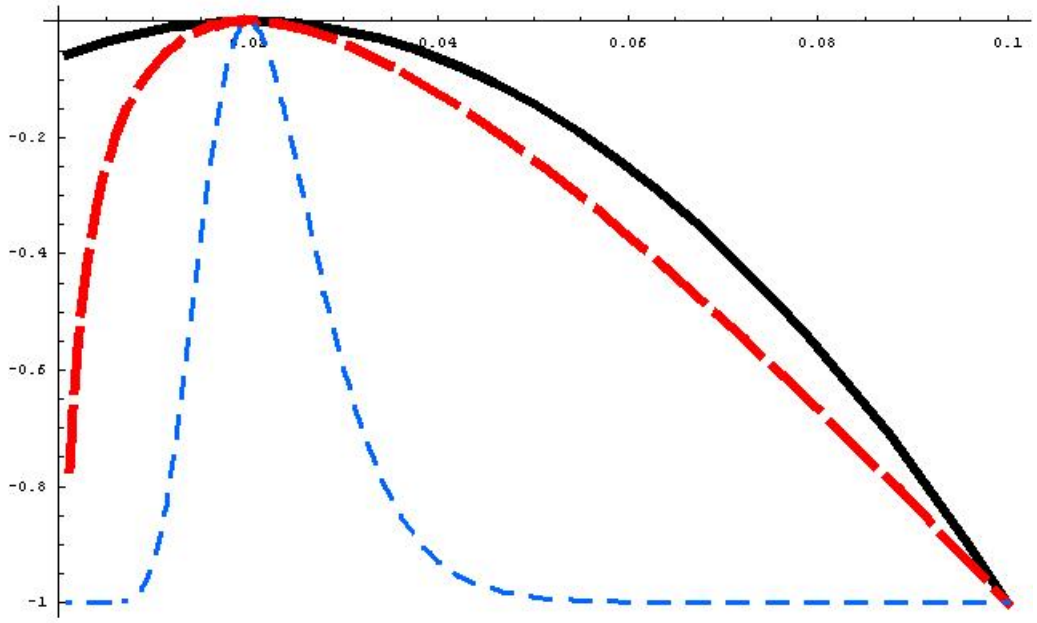

FIGURE 1. In this figure, the blue curve (short dotted lines) represents the empirical loss function, the red curve (long dotted lines) represents LL, and the black curve (solid line) represents MSE. The magnitudes of all three loss functions are minimized by using a pCTR equal to the true CTR. However, the empirical loss function imposes similar penalties on mispredictions that are off by more than some fixed amount, whereas LL and MSE impose significantly greater penalties for large mispredictions than for small mispredictions.

This figure indicates that the empirical loss function differs dramatically from MSE and LL when the highest competing eCPM bid is drawn from a log-normal distribution. Both MSE and LL impose significantly larger penalties for large mispredictions than they do for predictions that are only somewhat inaccurate. However, the empirical loss function imposes nearly identical losses for pCTRs that are off by more than a certain amount.

The reason for this is as follows. If a pCTR is far off the mark, additional errors are unlikely to affect the outcome of the auction. If one significantly overpredicts the CTR 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 CTR of an ad, then further underpredictions are also unlikely to matter because this ad is not likely to be competitive in the auction. Thus the loss function should be relatively insensitive to large errors. This insight also holds for other distributions besides the log-normal distribution:

Theorem 2. *If* $\lim_{A\to 0} g(A|b,x) = 0$ *and* $\lim_{A\to b} g(A|b,x) = 0$*, then the derivative of the empirical loss function with respect to the pCTR* q *becomes arbitrarily small in the limit as* $q \to 0$ *or* $q \to 1$*. However, the magnitudes of the derivative of MSE and LL with respect to* q *are increasing in the distance from* q *to* p*.*[5]

Theorem 2 underscores how different the empirical loss function is from MSE and LL. Under MSE and LL, the loss function changes the most when the pCTR is as far away from the ad's actual CTR as possible. By contrast, as long as the competing ads are unlikely to have eCPMs that are either arbitrarily close to zero or arbitrarily large, the empirical loss function will barely vary with the pCTR when the pCTR is so far from the actual CTR. Thus the empirical loss function will typically have the exact opposite behavior from MSE and LL for pCTRs that differ substantially from the true CTRs.

Theorem 2 indicates that the empirical loss function can differ significantly from MSE and LL, but these standard loss functions do have some desirable properties. In particular, suppose the auctioneer is uncertain about the true CTR of the ad p. In this case, the auctioneer may nonetheless have a sense of the likely CTR of this ad, and may believe that this true CTR is a random draw from some distribution $F(\cdot)$. In this setting, both MSE and LL are well-calibrated in the sense that the expected values of these loss functions will be minimized by using a pCTR equal to the expected CTR, $E_F[p]$. We thus wish to verify that this is also the case for the empirical loss function. This is done in Theorem 3:

Theorem 3. *Suppose the true CTR of the ad is unknown. Then the magnitude of the expected empirical loss function is minimized by a pCTR equal to the expected CTR.*

We now turn to other questions about how the properties of the empirical loss function compare to standard loss functions. One notable difference between the empirical loss function and standard loss functions regards the dependence of these loss functions on the

[5]Bax *et al.* (2012) notes that in auctions for online advertising, typical CTRs for ads are on the order of $\frac{1}{100}$ (for search ads) or $\frac{1}{1000}$ (for display ads) so the typical eCPMs 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 eCPMs of the competing ads, and $\lim_{A\to b} g(A|b,x)$ will almost certainly be very close to zero in real auctions.

advertisers' bids. While the MSE and LL loss functions are independent of the advertisers' bids, the empirical loss function is not. Instead, the extent to which the loss function penalizes overpredictions or underpredictions can depend on the size of an advertiser's bid:

Theorem 4. *Suppose that $g(\cdot|b,x)$ is independent of b and single-peaked at some eCPM bid $\hat{A}$. Then if $bp > (<)\hat{A}$, the empirical loss function imposes stricter penalties for small underpredictions (overpredictions) than for small overpredictions (underpredictions).*[6]

While there is nothing inherently wrong with allowing the pCTRs to vary with the bids, in some applications this might not be desirable because an advertiser may attempt to manipulate its pCTR by changing its bid. A system designer who wishes to ensure that an advertiser's pCTR never depends on its bid may thus wish to design an alternative loss function that never depends on the particular bids made in any given auction. In this case, an immediate corollary of Theorem 1 is that it is appropriate to use the following loss function:

Corollary 1. *Suppose that one wishes to maximize economic efficiency while using a loss function that is independent of an advertiser's bid. Then the correct loss function from using a pCTR of q when the actual CTR is p is $E[\int_{bp}^{bq}(bp-A)g(A|b,x)\,dA]$, where the expectation is taken over the distribution of the CPC bidders' bids.*

Throughout the analysis so far we have restricted attention to situations in which the loss function depends directly on the actual CTR of the ad p. However, in most situations we will not know the true CTRs of the ads, and it will instead be necessary to define the loss function exclusively in terms of clicks. We thus illustrate how one can extend the loss function in Theorem 1 to only depend on whether an advertiser received a click:

Theorem 5. *Suppose that one wishes to maximize economic efficiency while using a loss function that does not depend on the advertisers' CTRs. Then the correct loss function from*

[6]Nonetheless, the expectation of the empirical loss function is still minimized by using a pCTR equal to the expected CTR, as noted in Theorem 3. This shows that assumptions of symmetric loss and single-peaked probability distributions that are used in Granger (1969) are not needed to ensure that making a prediction equal to the conditional mean is optimal, as the loss function need not be symmetric in this setting.

using a pCTR of q *is* $\int_{bc}^{bq}(bc - A)g(A|b,x)\,dA$*, where* c *is a dummy variable that equals* 1 *if the CPC ad received a click and* 0 *otherwise.*

This loss function will still have the desirable property mentioned in Theorem 3 that the magnitude of the expected value of the loss function will be minimized by a pCTR equal to the actual CTR. For instance, if the true CTR of the ad is p, then one can apply Theorem 3 to the special case in which the distribution $F(\cdot)$ of actual CTRs 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 5 will be minimized by a pCTR equal to the actual CTR.

Thus far we have concerned ourselves with designing a loss function to maximize economic efficiency. While this is a reasonable goal, one might naturally wonder what happens if we use a loss function that optimizes a weighted average of efficiency and revenue since many systems may care about both of these metrics. Unfortunately, there are significant problems with using such a loss function:

Theorem 6. *Suppose that one wishes to maximize a weighted average of economic efficiency and revenue. Then the correct loss function 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 a pCTR equal to the actual CTR.*

Since it is quite important to ensure that the loss function is well-calibrated, Theorem 6 indicates that it is not appropriate to choose a loss function that optimizes revenue in addition to efficiency. Using such a loss function would result in poorly calibrated predictions, so it is better to simply use a loss function that reflects the efficiency loss from misestimates. We can further say something about when using a loss function that maximizes revenue would be optimized by making underpredictions or overpredictions:

Theorem 7. *Suppose that one wishes to maximize revenue and* $G(\cdot|b,x)$ *has bounded support and is independent of* b*. Then the correct loss function is optimized by making underpredictions (overpredictions) of CTRs for CPC ads with large (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 revenue will be optimized by raising this bidder's pCTR to increase the highest competing eCPM bidder's costs. Similarly, if a bidder makes a large bid, then this bidder is likely to win the auction, so one can increase revenue by lowering this bidder's pCTR and increasing the bidder's click cost.

We end this section with one last remark on our results. Throughout this section we have analyzed loss functions under standard second-price auctions for a single advertising opportunity. However, many online ad auctions are position auctions in which several advertising opportunities on the page are auctioned off at the same time. It is thus natural to ask whether the possibility of position auctions would have a significant effect on the results.

The appropriate loss functions turn out to be virtually unaffected by the possibility of position auctions. We show in a working paper (Hummel and McAfee 2015) that the optimal loss functions for position auctions only differ in that $g(A|b,x)$ is replaced by a weighted sum of densities corresponding to the distributions of the k^{th}-highest competing bids, where the weights correspond to the relative CTRs of the various positions. Thus we lose little by restricting attention to auctions for a single advertising opportunity.

4. Concave Value Functions

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

Example 2. *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 log-normal 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 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 8. *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 $g(\cdot)$ such that $(p-q)g(q)$ is increasing in q for values of q near 0 and 1.*[7] *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. 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. Such a concave approximation to the empirical loss function would take no more time to optimize than LL.

Theorem 8 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 minimization to be computationally feasible. Often one wishes to fit a model where the pCTR is a logistic function or a model where q is of the form $q = \frac{1}{1+e^{-\sum_{i=1}^{m}\beta_i x_i}}$, where each x_i is an observed feature, and each β_i is a coefficient on the feature that the model is trying to estimate.

[7]This condition on $g(\cdot)$ is automatically satisfied if $\lim_{q\to 0} g(q) = 0$, $g(q)$ is increasing in q for values of q near 0, and $g(q)$ is decreasing in q at a rate faster than $\frac{1}{q}$ for values of q near 1, as would surely be the case in practice. In practical applications, $g(\cdot)$ is typically modeled as a log-normal distribution with a mode near a typical CTR, and such a distribution would satisfy these conditions.

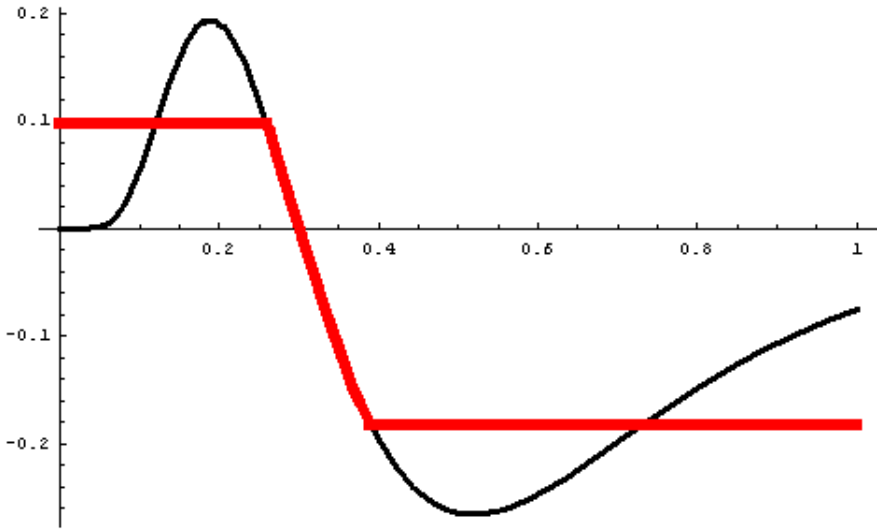

FIGURE 2. If the black curve represents the shape of the derivative of the empirical loss function with respect to the pCTR, then the red curve represents the shape of the derivative of the best concave loss function with respect to the pCTR.

If the model is of this form, then in order for it to be computationally feasible to find the values of the coefficients that minimize the loss function, it is no longer sufficient for the loss function to be concave in the pCTR. Instead the loss function must be concave in the coefficients $\vec{\beta} = (\beta_1, \ldots, \beta_m)$. We illustrate the form that the optimal loss function takes when the loss function must be concave in the coefficients $\vec{\beta} = (\beta_1, \ldots, \beta_m)$ in Theorem 9:

Theorem 9. *Suppose that one wishes to maximize economic efficiency while using a loss function that is concave in the coefficients* $\vec{\beta} = (\beta_1, \ldots, \beta_m)$ *and fitting a model of the form* $q = \frac{1}{1+e^{-\sum_{i=1}^{m} \beta_i x_i}}$. *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.

The derivative of the loss function in Theorem 9 is somewhat similar to the derivative of LL for values of q near zero since $\frac{\partial L(q,p)}{\partial q} = \frac{p}{q} - \frac{1-p}{1-q}$ for LL, so these derivatives both vary with $\frac{1}{q}$ for values of q near zero. Nonetheless, the magnitude of this derivative may be smaller for

the loss function in Theorem 9 since c may be lower than p for values of q near zero, so this loss function will impose relatively smaller penalties on large mispredictions than LL.

Finally, we analyze the case where the loss functions only depend on whether a click was observed. To do this, let $L_c(q)$ denote the loss if an ad receives a click, and let $L_n(q)$ denote the loss if an ad does not receive a click and the pCTR is q. We first derive the conditions needed to ensure the loss function is concave in the coefficients $\vec{\beta} = (\beta_1, \ldots, \beta_m)$ and calibrated in that the expected loss is minimized by a pCTR equal to the CTR:

Lemma 1. *Suppose that one is fitting a model of the form* $q = \frac{1}{1+e^{-\sum_{i=1}^m \beta_i x_i}}$ *and one wishes to use a calibrated loss function that is concave in the coefficients* $\vec{\beta} = (\beta_1, \ldots, \beta_m)$. *In that case the set of feasible loss functions* $L_c(q)$ *and* $L_n(q)$ *for the losses that are incurred under either a click or no 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* $-\frac{h(q)}{q} \leq h'(q) \leq \frac{h(q)}{1-q}$.

For the empirical loss function, the derivative of the loss function with respect to q will be relatively larger 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 Lemma 1 where $h(q)$ is relatively larger for values of q near the peak of this density. We derive the form of the optimal choice of this function $h(q)$ in Theorem 10 below:

Theorem 10. *Suppose that one is fitting a model of the form* $q = \frac{1}{1+e^{-\sum_{i=1}^m \beta_i x_i}}$ *and one wishes to use a well-calibrated loss function that is concave in the coefficients* $\vec{\beta} = (\beta_1, \ldots, \beta_m)$. *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 10 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 Lemma 1, 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.

5. Simulations Using Misspecified Models

While there are significant differences between the empirical loss function in Theorem 1 and standard loss functions such as LL, it is not clear from the results presented so far whether these differences would actually have a substantive effect on online ad auctions. In this section we investigate whether using the correct loss function will have a significant effect on efficiency in simulations that are designed to closely parallel a real-world auction environment when the model for pCTRs is misspecified.

Throughout we consider an auction in which a CPC bidder with a CPC bid of 1 is competing against a field of CPM bidders for a single advertising opportunity. This type of setting is frequently encountered in online auctions. For example, when AdSense bids on behalf of Google's advertisers for display advertising opportunities on an ad exchange, Google's advertisers bid nearly exclusively on a CPC basis, whereas the advertisers from other ad networks all bid on a CPM basis. Similarly, in auctions for advertising opportunities on YouTube, there is typically a mix of CPC and CPM bidders.

In real-world auctions, an ad's CTR depends on the context in which it is shown. Advertiser characteristics, publisher characteristics, and user characteristics can all influence the CTR of an ad. Moreover, the presence of multiple such characteristics can further influence the probability that an ad receives a click. In a typical machine learning system, these characteristics are modeled as features x_i that assume the value 1 if the feature is present and 0 otherwise. Moreover, there will be additional features $x_{i,j} = x_i x_j$ that capture these

interaction effects arising from multiple characteristics being present at the same time that may also influence the CTRs of the ads.

We consider a situation in which there are 1000 possible contexts, each of which is equally likely to occur, and 211 possible features that may influence CTRs. In each of these contexts, there is one feature x_0 that is always present (*i.e.* $x_0 = 1$ in all contexts). There are also 20 features $x_1, \ldots, x_{20}$ which are each present in a given context with probability $\frac{1}{5}$ (*i.e.* $x_i = 1$ with probability $\frac{1}{5}$ and $x_i = 0$ with probability $\frac{4}{5}$ in each context independent of the other features). Finally, the 190 remaining features represent the various possible interaction terms $x_{i,j}$ that are present if and only if both x_i and x_j are present (*i.e.* $x_{i,j} = x_i x_j$ for all possible distinct pairs of positive integers i and j less than or equal to 20).

Throughout we consider a situation in which the logit of the pCTR is a linear function of the values of the features. That is, if q_c denotes the pCTR for the CPC ad in context c, then there are some coefficients β_k on the various features x_k such that $\text{logit}(q_c) = \sum_k \beta_k x_k^c$, where x_k^c denotes the value that the feature x_k assumes in context c. This parallels the functional forms used by many machine learning systems for pCTRs.

We consider two possible models for the true CTRs of the ads. In the first model, the logit of the true CTR is also a linear function of the values of the features. Under this model, the only source of misspecification will come from including the wrong features in the model for pCTRs. In the second model, a scaled probit of the actual CTR is a linear function of the values of the features. That is, if p_c denotes the true probability that the CPC ad will receive a click in context c, then there are some coefficients β_k such that $\text{probit}(p_c) = \sqrt{\frac{\pi}{8}} \sum_k \beta_k x_k^c$, where x_k^c denotes the value that the feature x_k assumes in context c. In this model, misspecification errors may come from both functional form misspecification as well as from including the wrong features in the model for pCTRs.

The true coefficients on the features are set as follows. To capture the fact that many features that are present in the real world are features that will not actually have any effect on the CTRs of the ads, we assume throughout that the true values of some of the coefficients

are zero. In particular, we assume that a random set of 110 of the 190 features capturing the interaction terms have true coefficients equal to 0.

The true coefficients on the remaining features are set in such a way that the distribution of the true CTRs of the ads in the various possible contexts will closely parallel those given in existing empirical work such as Lahaie and McAfee (2011). For the setting in which the logit of the true CTR is a linear function of the values of the features, we assume that the true value of the constant coefficient satisfies $\beta_0 = -6$ and the true values of the other non-zero coefficients are independent random draws from the normal distribution with mean 0 and standard deviation 1. Here the constant coefficient is set to be quite negative to capture the fact that the average CTRs in online auctions tend to be on the order of $\frac{1}{100}$ or even $\frac{1}{1000}$ (Bax *et al.* 2012), while the random draws of the other coefficients ensure that the variance is comparable to that of distributions of CTRs encountered in existing empirical work.

For the setting in which the probit of the true CTR is a linear function of the values of the features, we instead assume that the true value of the constant coefficient satisfies $\beta_0 = -4.5$ and the true values of the other non-zero coefficients are independent random draws from the normal distribution with mean 0 and standard deviation $\frac{1}{2}$. The differences in the coefficients reflect the need to choose different coefficients in order to match empirical evidence on ad CTRs when the probit of the true CTR is a linear function of the values of the features rather than a logit. For instance, if $p = 0.0024$, then $\text{probit}(p)/\sqrt{\frac{\pi}{8}} = -4.5$ while $\text{logit}(p) = -6.0$, and if $p = 0.000085$, then $\text{probit}(p)/\sqrt{\frac{\pi}{8}} = -6.0$ while $\text{logit}(p) = -9.4$. To ensure that the actual CTRs of the least frequently clicked ads are not unrealistically small for the probit model, it is necessary to make the true value of the constant coefficient larger and induce less variance in the other coefficients.

We seek to illustrate whether fitting the coefficients with the empirical loss function rather than LL would improve performance when there is an arbitrarily large amount of training data if the model is misspecified. We address this question in three settings. First we consider a setting in which the only source of misspecification arises from including the wrong features in the model for pCTRs. Next we consider a setting in which the only

source of misspecification is the functional form misspecification arising from the fact that the probit of the true CTR is a linear function of the values of the features rather than the logit. Finally, we consider a setting in which both sources of misspecification are present.

In settings where one source of misspecification arises from including the wrong features in the model, we endogenously generate a misspecified model that could realistically arise given standard procedures used by machine learning systems. In particular, we first consider a situation in which we have a finite amount of training data that could be used to fit the model in the sense that, for each of the 1000 possible contexts, we observe 100 independent impressions that were either clicked or not clicked. Given these $100,000$ impressions of training evidence, we then fit a model to optimize LL with L_1-regularization. That is, we choose the coefficients on the features β_k to maximize

$$\sum_c [\tilde{p}_c \log(q_c) + (1 - \tilde{p}_c) \log(1 - q_c)] - \lambda \sum_k |\beta_k|, \tag{1}$$

where $\tilde{p}_c$ denotes the empirically observed frequency with which an ad in context c was clicked, q_c denotes the pCTR for an ad in context c given the coefficients β_k, $\lambda \geq 0$ denotes a regularization term, and β_k denotes the value of the coefficient on feature k. This technique of using L_1-regularization to select the features that will be used in the model is one that is commonly used in machine learning (Koh *et al.* 2007; Ng 2004; Wainwright *et al.* 2006) and is also used by Google in its ad click prediction algorithms (McMahan *et al.* 2013).

An effect of optimizing LL with L_1-regularization is that many of the coefficients β_k in the model will be set to zero. The $-\lambda \sum_k |\beta_k|$ term in equation (1) is optimized by setting all the coefficients to be exactly 0, so if the value of β_k that would be best for LL is only slightly different from zero, β_k will typically be set to zero as a result of this regularization. However, given the finite amount of training evidence, the features whose coefficients are set to zero will not necessarily be the same as the features where the true value of the coefficient is zero. Instead there will be some features whose true coefficients are zero where the fitted coefficients are non-zero and some features whose true coefficients are non-zero where the fitted coefficients are zero.

A result of this feature selection process is that we will generate a misspecified model in which the features included in the model are not the same as the features that should be in the model. Given this endogenously generated misspecified model, we then reestimated the coefficients in this model using both LL and the empirical loss function with an infinitely large training sample. We then investigated whether using the empirical loss function rather than LL improves economic efficiency.

In doing this, we must specify the distribution of competing CPM bids that will be used to calculate both economic surplus and the appropriate coefficients to optimize the empirical loss function. Because there is empirical evidence that the distribution of competing bids can be reasonably well modeled by a log-normal distribution (Lahaie and McAfee 2011; Ostrovsky and Schwarz 2016; Sun *et al.* 2014), throughout we consider a situation in which the highest competing CPM bid is drawn from a log-normal distribution with mean μ and variance σ^2. Moreover, to ensure that the highest competing CPM bid tends to be around the same order of magnitude as the typical true eCPM of the CPC bidder, we consider a scenario in which $\mu = -6$ and $\sigma^2 = 4$.

We begin by presenting the results for the case where the only source of model misspecification comes from including the wrong features in the model. Here we considered five different possible values of λ: 0.05, 0.09, 0.15, 0.25, and 0.35. For each of these possible values of λ, we conducted a dozen different simulations. In each simulation we randomly generated the correct values of the true coefficients in the model and then endogenously generated the misspecified model using the procedure described above. After fitting the coefficients of this misspecified model using both LL and the best concave approximation to the empirical loss function with an arbitrarily large training set, we then calculated the efficiencies resulting from using the model trained by these different objective functions.

Table 1 summarizes the results of our simulations. In analyzing our results, we first consider the set of simulations we conducted in which the misspecified models were generated using a value of λ satisfying $\lambda = 0.09$. Such a value of λ is one that seems to generate a realistic amount of misspecification. When $\lambda = 0.09$, the number of features that end up

λ	Missing Features	Misspecification Errors	Changes in Logit(pCTR)	Efficiency Increases
0.05	[16, 52] 34	[0.09, 0.43] 0.22	[0.04, 0.10] 0.06	[0.003%, 0.011%] 0.006%
0.09	[14, 37] 30	[0.18, 0.51] 0.33	[0.06, 0.15] 0.10	[0.004%, 0.021%] 0.013%
0.15	[29, 48] 38	[0.14, 0.58] 0.43	[0.06, 0.16] 0.12	[0.004%, 0.036%] 0.022%
0.25	[35, 74] 52	[0.40, 0.96] 0.67	[0.10, 0.30] 0.17	[0.021%, 0.067%] 0.048%
0.35	[51, 74] 61	[0.53, 1.08] 0.83	[0.12, 0.37] 0.20	[0.036%, 0.134%] 0.069%

TABLE 1. Ranges and mean values of our simulation results when the misspecification errors come entirely from including the wrong features in the model. The second column gives the number of features with non-zero coefficients that are not present in the misspecified model. The third column gives the average difference between the logit of the pCTR and the logit of the actual CTR in the various contexts in the model fit using LL. The fourth column gives the average difference between the logit of the pCTRs in the models fit using the different loss functions. And the final column gives the efficiency increase from using the empirical loss function.

being included in the model is close to the number of features that actually have an effect on the CTRs of the ads. However, the model is still misspecified in the sense that, on average, 30 of the 101 features that actually have an effect on the CTR of the ad are not included in the model, while a similar number of features whose true coefficients are zero are included in the model. Fitting such a model results in average errors in the pCTRs that range from anywhere between $20 - 50\%$, which seem in the realm of what is plausible.[8]

In each of the dozen simulations we conducted using a misspecified model generated with $\lambda = 0.09$, we obtained efficiency gains from using the empirical loss function. However, the efficiency gains were never greater than 0.021%. Logically it makes sense that these gains are small. Suppose, for example, that the average error in our pCTRs is roughly 30%. In this case, using a different loss function should, on average, change our pCTRs by only a fraction of 30%. And indeed in our simulations, we found that the average change in pCTRs as a result of using the empirical loss function when $\lambda = 0.09$ was only about 10%.

[8]Hummel and McAfee (2016a) note that the average errors in predicted clicks for new ads are likely to be in the $20 - 30\%$ range. See Appendix B for a discussion of how these error rates can be estimated.

Now even if we managed to improve the pCTRs by 10% on every single impression, such a change would likely only increase economic efficiency by a few tenths of a percentage point: Changing an ad's pCTR by 10% will only affect whether that ad wins the auction in about 10% of auctions. And conditional on this change affecting whether an ad wins the auction, the expected efficiency gain will only be about half the size of the change in pCTRs, or about 5%. Thus improving the pCTRs by 10% on every single impression would only result in an efficiency gain of about 10% of 5% or about half of a percentage point.

But in our setting, we are not improving the pCTRs by 10% on every single impression by using a different loss function. Instead we are improving the pCTRs by about 10% on impressions where this is relatively more likely to have an effect on which ad wins the auction, while making our pCTRs less accurate in other cases where this is relatively less likely to have an effect on which ad wins the auction. The benefits to this are likely to be an order of magnitude smaller than the benefits to making the pCTRs more accurate on every single impression, so it is no surprise that the efficiency gains from using the empirical loss function are only a few hundredths, rather than a few tenths, of a percentage point.

Doing more regularization by using larger values of λ in equation (1) generally results in models with a smaller number of features and a larger misspecification error.[9] In cases where we have greater misspecification error, we find that there is relatively more benefit to using the empirical loss function. Figure 3 plots the realized efficiency gain as a function of the average error in the pCTRs. From this figure, it is clear that there is greater benefit to using the empirical loss function when the model is more severely misspecified. However, this figure also indicates that even when a model is severely misspecified, the gains from using the empirical loss function are still not very large. Even though our simulations include cases

[9]These models also reduce the number of features in the model that should not be included in the model. However, since the models are being trained with an arbitrarily large amount of data, all the models considered in these simulations are effectively underfitted, and reducing the number of features in the model tends to reduce accuracy. In practice there may be benefits to using an underfitted model, even if using such a model comes at a slight cost in accuracy, since larger models require more computing resources and are thus more costly to deploy. Thus overfitting is also relatively less likely to be an issue in practice. See McMahan *et al.* (2013) for techniques Google employed to achieve memory savings in deploying machine learning models for predicting ad CTRs and Shamir (2015) for novel regularization techniques that can achieve significant reductions in model size at little cost in accuracy.

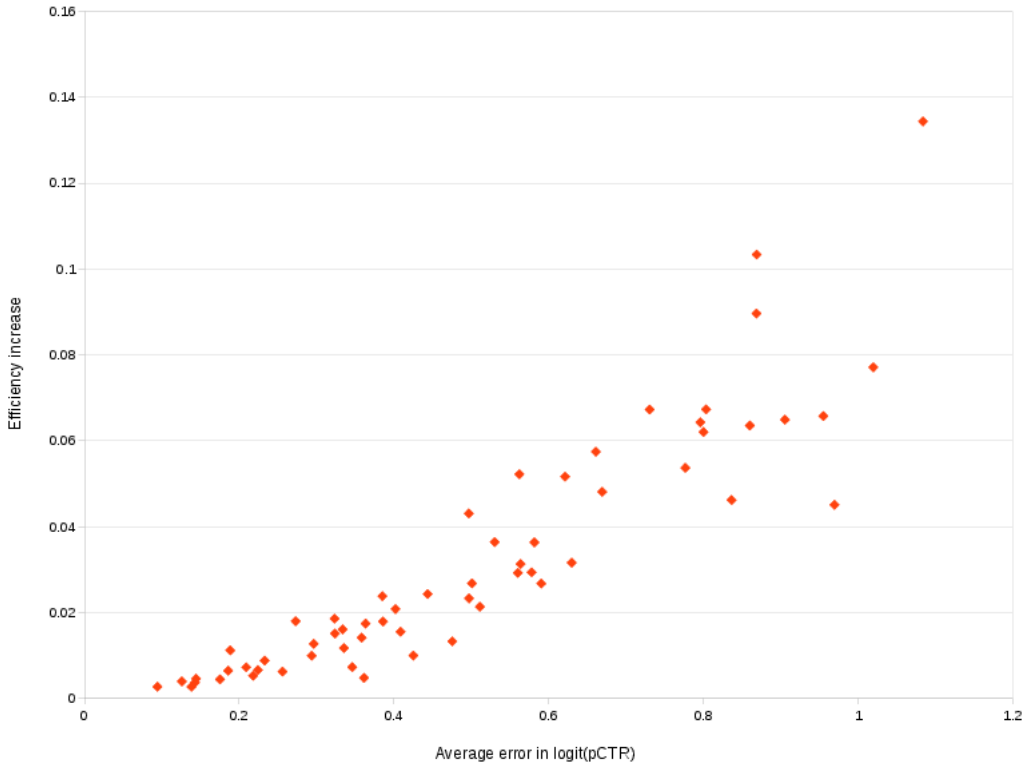

FIGURE 3. Percentage efficiency increase from using empirical loss function as a function of the average error in logit(pCTR).

in which the average error in the pCTRs is over 100%, the efficiency gains from using the empirical loss function were never greater than 0.14%.[10]

The above results were for the case in which the only source of model misspecification came from including the wrong features in the model of pCTRs, but we also found similar substantive conclusions for the case in which there was functional form misspecification. For our simulations in which the probit of the true CTR was a linear function of the values of the features, we considered several possible settings.

In the first setting, we assumed that the features in the model were the same as the features that actually had non-zero coefficients in the true function giving the actual CTRs of the ads.

[10]It is also worth noting that while adopting the empirical loss function always resulted in an increase in efficiency, use of this alternative loss function gives no guarantee of increasing revenue. For example, for seven of the twelve simulations in Table 1 with $\lambda = 0.05$, revenue declined as a result of adopting the empirical loss function. The average revenue change in this case was -0.01% with a range of $[-0.40\%, 0.71\%]$. In fact, even making the predicted click-through rates more accurate in every single auction would not have a guarantee of increasing revenue (Fu *et al.* 2012; Hummel and McAfee 2016b).

In the second setting, we followed the same procedure given above to select features using L_1-regularization and thereby generate a misspecified model in which the features included in the model are not the same as the features that should be in the model. Here we considered four different possible values of the regularization parameter λ: 0.03, 0.06, 0.09, and 0.12. For each of these values of λ, we conducted a dozen different simulations using the same procedure described above. Smaller values of λ were selected because the fact that the true coefficients in the model were typically smaller meant that less regularization was needed to generate a given amount of model misspecification.

The results of these simulations are summarized in Table 2. Table 2 again reveals little gain in economic efficiency as a result of using the concave approximation to the empirical loss function rather than LL to fit the model of pCTRs. In fact, if anything, the gains in economic efficiency from using the concave approximation to the empirical loss function appear to be even smaller for these simulations. Although these simulations also include examples in which the average error in the pCTRs is over 100%, the efficiency gains from using the empirical loss function in these simulations were never greater than 0.05%. Interestingly, these simulations also noted considerably lower correlation between the size of the misspecification errors and the economic efficiency gains resulting from using an economically motivated loss function.

λ	Missing Features	Misspecification Errors	Changes in Logit(pCTR)	Efficiency Increases
N/A	N/A	[0.20, 0.63] 0.44	[0.02, 0.22] 0.11	[0.001%, 0.042%] 0.018%
0.03	[14, 30] 24	[0.39, 0.59] 0.50	[0.05, 0.15] 0.10	[0.005%, 0.024%] 0.014%
0.06	[32, 53] 39	[0.45, 0.79] 0.60	[0.06, 0.18] 0.11	[0.009%, 0.041%] 0.021%
0.09	[29, 50] 41	[0.56, 0.79] 0.66	[0.07, 0.22] 0.13	[0.007%, 0.046%] 0.027%
0.12	[34, 82] 54	[0.52, 1.59] 0.84	[0.05, 0.18] 0.13	[0.003%, 0.048%] 0.029%

TABLE 2. Ranges and mean values of our simulation results for simulations in which the model contained functional form misspecification. The first row reports simulation results when the functional form misspecification was the sole source of model misspecification and the remaining rows consider cases in which additional model misspecification arose from including the wrong features in the model for different values of λ. The meanings of the columns are the same as in Table 1.

6. Experimental Results

In this section we sought to experimentally verify whether making use of the empirical loss function would have a significant effect on efficiency in online ad auctions via experiments on Google's AdSense product. To do this, we trained two separate models of predicted click-through rates. The first model was trained to optimize LL, as is standard in machine learning. The second model was trained to optimize the empirical loss function by estimating the overall distribution of eCPM bids $G(\cdot)$ on AdSense and using this estimated distribution to calculate the empirical loss function. Throughout we made use of an approximation to the empirical loss function based on Corollary 1 that does not depend on an advertiser's bids to ensure that no advertiser could attempt to manipulate its pCTR by changing its bid. The amount of regularization in the model trained to optimize the empirical loss function was chosen in such a way to ensure that the models trained to optimize LL and the empirical loss function would have similar numbers of features.

To analyze whether making use of the empirical loss function had a significant effect, we considered two separate metrics. First we measured the average difference between the logit of the pCTRs of the models trained to optimize the different loss functions. Here we found that the average change in logit(pCTR) that resulted from using the model trained to optimize the empirical loss function was 0.12. This average change is broadly consistent with the typical simulation results in Table 2, and thus suggests that the typical efficiency gains may be similar to those in the simulations in Table 2.

In addition to measuring how much the pCTRs changed by using the empirical loss function, we then conducted an experiment in which we used our newly trained pCTR model to serve traffic on AdSense. We randomly split traffic on AdSense into a control group and an experimental group. In the control group, we made use of a standard pCTR model that was trained to optimize LL in deciding which ads to display. In the experimental group, we instead made use of the model that was trained to optimize the empirical loss function. We then compared revenue and efficiency that resulted from using a model trained to optimize the empirical loss function rather than LL. Since AdSense makes use of a VCG auction in

which bidders have an incentive to make bids equal to their true values, we assumed the advertisers' values per click were equal to their bids in computing efficiency.

In the experiment we measured a revenue change of -0.01% (with a 95% confidence interval of $[-0.18\%, 0.17\%]$) and an efficiency change of 0.03% (with a 95% confidence interval of $[-0.56\%, 0.61\%]$) as a result of using a model trained to optimize the empirical loss function. These results are consistent with the simulation results in Section 5 since the estimated efficiency gains in these simulations all fall within the confidence intervals given in this experiment. However, the experiment does not establish the exact amount of benefit to using the empirical loss function since improvements of the size given in the simulations in Section 5 are typically undetectable in such an experiment.[11]

7. Conclusion

This paper has considered the question of the choice of loss function for pCTRs in auctions for online advertising. We have shown that a loss function reflecting the true empirical loss from 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 MSE and LL.

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 CTRs are more severely penalized than overpredictions for advertisers with large bids, while overpredictions are more severely penalized for advertisers with small bids. We have also shown that the best loss function for maximizing economic efficiency will lead to calibrated predictions, while loss functions chosen to optimize revenue will not.

We have also considered the question of the optimal loss function when one is restricted to using a concave loss function for reasons of computational tractability. When one must use

[11]For instance, the average efficiency increases in each of the settings considered in Table 2 are less than 0.03%, while the half-widths in the confidence intervals in the above experiment are roughly $\pm 0.6\%$. To obtain confidence intervals with half-widths of $\pm 0.03\%$ would thus require reducing these confidence intervals by a factor of 20, which would in turn require over 400 times as much traffic. Since this far exceeds the amount of traffic we would be able to use for this experiment, we cannot detect improvements of the size in Table 2 in a live traffic experiment.

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 by 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.

Finally, we have addressed the question of whether using the empirical loss function rather than standard loss functions would improve performance in online auctions if the pCTR model is misspecified. In the simulations we conducted, we were consistently able to improve economic efficiency by using the empirical loss function, but the size of the improvement was small. Our most realistic simulations never revealed efficiency gains greater than a few hundredths of a percentage point, and even with a large amount of model misspecification, we never obtained gains significantly greater than a tenth of a percentage point.

Acknowledgments

We thank Glenn Ellison, David Grether, Charles Manski, Robert Porter, numerous colleagues at Google, the editor, the anonymous referee, and seminar attendees at UC Davis for helpful comments and discussions.

References

Arrow KJ. 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*, Olkin I (ed). Stanford University Press: Stanford.

Bartlett PL, Jordan MI, McAuliffe JD. 2006. Convexity, classification, and risk bounds." *Journal of the American Statistical Association* 101(473): 138-156.

Basu S, Markov S. 2004. Loss function assumptions in rational expectations tests on financial analysts' earning forecasts. *Journal of Accounting and Economics* 38: 171-203.

Bax E, Kuratti A, McAfee RP, Romero J. 2012. Comparing predicted prices in auctions for online advertising. *International Journal of Industrial Organization* 30(1): 80-88.

Boudt K, Croux C. 2010. Robust M-Estimation of multivariate GARCH models." *Computational Statistics and Data Analysis* 54(11): 2459-2469.

Chapelle O. 2015. Offline evaluation of response prediction in online advertising auctions. *Proceedings of the* 24^{th} *International Conference on the World Wide Web* (WWW) 919-922.

Cohen MA, Ho TH, Ren ZJ, Terwiesch C. 2003. Measuring imputed cost in the semiconductor imputed supply chain. *Management Science* 49(12): 1653-1670.

Denuit M, Dhaene J. 2001. Bonus-Malus scales using exponential loss functions. *Blätter der Deutschen Gesellschaft für Versicherungs und Finanzmathematik* 25(1): 13-27.

Dmochowski JP, Sajda P, Parra LC. 2010. Maximum likelihood in cost-sensitive learning: model specification, approximations, and upper bounds. *Journal of Machine Learning Research* 11(Mar): 3313-3332.

Elliott G, Komunjer I, Timmerman A. 2005. Estimation and testing of forecast optimality under optimal loss. *Review of Economic Studies* 72(4): 1107-1125.

Elliott G, Lieli RP. 2013. Predicting binary outcomes. *Journal of Econometrics* 174(1): 15-26.

Fu H, Jordan P, Mahdian M, Nadav U, Talgam-Cohen I, Vassilvitskii S. 2012. Ad auctions with data. *Proceedings of the* 5^{th} *International Symposium on Algorithmic Game Theory* (SAGT) 184-189.

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102(477): 359-378.

Granger CWJ. 1969. Prediction with a generalized cost of error function. *Journal of the Operations Research Society* 20(2): 199-207.

Hilbe JM. 2009. *Logistic Regression Models.* Chapman & Hall: London.

Hummel P, McAfee RP. 2015. Loss functions for predicted click-through rates in auctions for online advertising. Google Inc. Typescript.

Hummel P, McAfee RP. 2016a. Machine learning in an auction environment. *Journal of Machine Learning Research* 17(197): 1-37.

Hummel P, McAfee RP. 2016b. When does improved targeting increase revenue? *ACM Transactions on Economics and Computation* 5(1): Article 4.

Koh K, Kim SJ, Boyd S. 2007. An interior-point method for large-scale ℓ_1-regularized logistic regression. *Journal of Machine Learning Research* 8(Jul): 1519-1555.

Lahaie S, McAfee RP. 2011. Efficient ranking in sponsored search. *Proceedings of the* 7^{th} *International Workshop on Internet and Network Economics* (WINE) 254-265.

Lieli RP, White H. 2010. The construction of empirical credit scoring models based on maximization principles. *Journal of Econometrics* 157(1): 110-119.

Manski CF. 2004. Statistical treatment rules for heterogeneous populations. *Econometrica* 72(4): 1221-1246.

McMahan BH, Holt G, Sculley D, Young M, Ebner D, Grady J, Nie L, Phillips T, Davydov E, Golovin D, Chikkerur S, Liu D, Wattenberg M, Hrafnkelsson AM, Boulos T, Kubica J. 2013. Ad click prediction: a view from the trenches. *Proceedings of the* 19^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD) 1222-1230.

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

Ng AY. 2004. Feature selection, L_1 vs. L_2 regularization, and rotational invariance. *Proceedings of the* 21^{st} *International Conference on Machine Learning* (ICML) 78-85.

Ostrovsky M, Schwarz M. 2016. Reserve prices in Internet advertising auctions: a field experiment. Stanford Graduate School of Business Typescript.

Patton AJ, Timmerman A. 2007. Testing forecast optimality under unknown loss. *Journal of the American Statistical Association* 102(480): 1172-1184.

Reid MD, Williamson RC. 2010. Composite binary losses. *Journal of Machine Learning Research* 11(Sep): 2387-2422.

Reid MD, Williamson RC. 2011. Information, divergence, and risk for binary experiments. *Journal of Machine Learning Research* 12(Mar): 731-817 (2011).

Shalit H, Yitzhaki S. 2002. Estimating beta. *Review of Quantitative Finance and Accounting* 18(2): 95-118.

Shamir GI. 2015. Minimum description length (MDL) regularization for online learning. *Journal of Machine Learning Research: Workshop and Conference Proceedings* 44: 260-276.

Skalak DB, Niculescu-Mizil A, Caruna R. 2007. Classifier loss under metric uncertainty. *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing* (EMNLP) 310-322.

Steinwart I. 2007. How to compare different loss functions and their risks. *Constructive Approximation* 26(2): 225-287.

Sun Y, Zhou Y, Deng X. 2014. Optimal reserve prices in weighted GSP auctions. *Electronic Commerce Research and Applications* 13(3): 178-187.

Varian, HR. 1974. A Bayesian approach to real estate assessment. In *Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage*, Fienberg SE, Zellner A (eds). North-Holland: Amsterdam.

Wainwright MJ, Ravikumar P, Lafferty JD. 2006. High-dimensional graphical model selection using ℓ_1-regularized logistic regression. *Proceedings of the* 19^{th} *Conference on Neural Information Processing Systems* (NIPS) 1465-1472.

Weiss AA. 1996. Estimating time series models using the relevant cost function. *Journal of Applied Econometrics* 11(5): 539-560.

Yahoo! Labs. 2013. "Project: Display Inventory Allocation Optimization." http://labs.yahoo.com/project/display-inventory-allocation-optimization/. Accessed July 2013.

Zhang T. 2004. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Mathematical Statistics* 32(1): 56-134.

Web Appendices: Not for Publication

Appendx A: Proofs of Theorems

Proof of Theorem 2: The derivative of the empirical loss function with respect to q is $b^2(p-q)g(bq|b,x)$. If $\lim_{A\to 0} g(A|b,x) = 0$ and $\lim_{A\to b} g(A|b,x) = 0$, 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 $q > p$. The result follows. □

Proof of Theorem 3: Note that the pCTR q should be chosen to optimize the value of the expression $E[\int_{bp}^{bq}(bp-A)g(A|b,x)\ dA]$, where the expectation is taken over the uncertain realization of p. Differentiating this expression with respect to q then gives $E[b^2(p-q)g(bq|b,x)]$, which equals zero if and only if $q = E[p]$. Thus the magnitude of the expected empirical loss function is minimized by a pCTR equal to the expected CTR. □

Proof of Theorem 4: If $g(\cdot) \equiv g(\cdot|b,x)$ is single-peaked at some eCPM bid $\hat{A}$ and $bp > \hat{A}$, then $g(bp-\epsilon) > g(bp+\epsilon)$ for small values of $\epsilon > 0$, so $|(bp-A)g(A)|$ is greater when $A = b(p-\epsilon)$ than when $A = b(p+\epsilon)$ for 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 empirical loss function imposes stricter penalties for making small underpredictions than for making small overpredictions if $bp > \hat{A}$. A similar argument proves that the empirical loss function imposes stricter penalties for making small overpredictions than for making small underpredictions if $bp < \hat{A}$. □

Proof of Theorem 5: The total surplus that results from using a pCTR of q is $bcPr(bq \geq A) + APr(bq < A) = \int_{bq}^{\infty} A\ g(A|b,x)\ dA + \int_0^{bq} bc\ g(A|b,x)\ dA$. And the total surplus that results from correctly predicting whether an ad will receive a click in any given auction is $\int_{bc}^{\infty} A\ g(A|b,x)\ dA + \int_0^{bc} bc\ g(A|b,x)\ dA$. From this it follows that the loss in efficiency that

results from using a pCTR of q rather than correctly predicting whether the ad will receive a click is $\int_{bq}^{\infty} A\ g(A|b,x)\ dA + \int_0^{bq} bc\ g(A|b,x)\ dA - [\int_{bc}^{\infty} A\ g(A|b,x)\ dA + \int_0^{bc} bc\ g(A|b,x)\ dA] = \int_{bc}^{bq}(bc - A)g(A|b,x)\ dA$. The result then follows. $\square$

Proof of Theorem 6: To prove this result, it suffices to show that the loss function that reflects the expected revenue loss from using a pCTR of q when this ad's actual CTR is 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 using a pCTR of q when the actual CTR is p is $\frac{p}{q}\int_0^{bq} A\ g(A|b,x)\ dA + bq(1 - G(bq|b,x))$, and the loss function that reflects the expected revenue loss from using such a pCTR is $\frac{p}{q}\int_0^{bq} A\ \ g(A|b,x)\ \ dA + bq(1 - G(bq|b,x)) - [\int_0^{bp} A\ \ g(A|b,x)\ \ dA + bp(1 - G(bp|b,x))]$. Differentiating this loss function with respect to q gives $-\frac{p}{q^2}\int_0^{bq} A\ g(A|b,x)\ dA + b^2p\ g(bq|b,x) + b(1 - G(bq|b,x)) - b^2q\ g(bq|b,x)$, and in the case where $G(\cdot|b,x)$ 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 loss function to be minimized at a prediction $q \neq p$. The result then follows. $\square$

Proof of Theorem 7: We know from the proof of Theorem 6 that the derivative of the loss function with respect to q is $-\frac{p}{q^2}\int_0^{bq} A\ g(A)\ dA + b^2p\ g(bq) + b(1 - G(bq)) - b^2q\ g(bq)$ when $G(\cdot) \equiv G(\cdot|b,x)$ and $g(\cdot) \equiv g(\cdot|b,x)$, 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 function is optimized by making overpredictions of CTRs 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 $\frac{E[A]}{p}$. 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 CTRs for CPC ads with large bids. The result then follows. $\square$

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

Proof of Theorem 8: If one uses the loss function $L(q,p)$, then this loss function will result in some distribution of pCTRs given the actual CTRs which we can model by the cumulative distribution function $F(q|p)$. The machine learning system will select a 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 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 9: Note that if we are fitting a model where q is of the form $q = \frac{1}{1+e^{-\sum_{i=1}^{m}\beta_i x_i}}$, then the loss function will be concave in the coefficients $\vec{\beta} = (\beta_1, \ldots, \beta_m)$ 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 $\vec{\beta} = (\beta_1, \ldots, \beta_m)$, 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 $\frac{\partial q}{\partial \beta} = \frac{Cxe^{-\beta x}}{(1+Ce^{-\beta x})^2} = xq(1-q)$, and we also have $\frac{\partial^2 q}{\partial \beta^2} = \frac{-Cx^2e^{-\beta x}(1+Ce^{-\beta x})^2+2C^2x^2(1+Ce^{-\beta x})(e^{-\beta x})^2}{(1+Ce^{-\beta x})^4} = \frac{Cx^2e^{-\beta x}(2Ce^{-\beta x}-(1+Ce^{-\beta x}))}{(1+Ce^{-\beta x})^3} = \frac{Cx^2e^{-\beta x}(Ce^{-\beta x}-1)}{(1+Ce^{-\beta x})^3} = x^2q(1-q)(1-2q)$. From this it follows that $\frac{\partial^2 L(q,p)}{\partial \beta^2} = \frac{\partial^2 L(q,p)}{\partial q^2}(\frac{\partial q}{\partial \beta})^2 + \frac{\partial L(q,p)}{\partial q}\frac{\partial^2 q}{\partial \beta^2} = \frac{\partial^2 L(q,p)}{\partial q^2}x^2q^2(1-q)^2 + \frac{\partial L(q,p)}{\partial q}x^2q(1-q)(1-2q)$. This in turn implies that $\frac{\partial^2 L(q,p)}{\partial \beta^2} \leq 0$ if and only if $\frac{\partial^2 L(q,p)}{\partial q^2}q(1-q) + \frac{\partial L(q,p)}{\partial q}(1-2q) \leq 0$.

Now we know by the reasoning in the proof of Theorem 8 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 Lemma 1: 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 $\vec{\beta} = (\beta_1, \ldots, \beta_m)$, we know from the reasoning in the proof of Theorem 9 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 $-\frac{2b(q)}{q(1-q)}+\frac{b'(q)}{q}-\frac{(1-2q)b(q)}{q^2(1-q)} \leq 0$, 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 $q(1-q)h(q) + q^2(1-q)h'(q) - qh(q) \leq 0$, 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 10: We know from the reasoning in the proof of Theorem 8 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 10, 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 $q(1-q)g'(q) = 0$ when q is 0 or 1, but $(3q-1)g(q)$ is negative for values of q near 0 and $(3q-2)g(q)$ is positive when q is 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 $q(1-q)g(q)$ as possible 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. $\square$

Appendix B: Estimating Average Errors in Predicted Click-Through Rates

This section describes techniques that could be used to estimate the average errors in predicted click-through rates. While it is not possible to ever identify the average errors in a model's predicted click-through rates on an impression by impression basis, it is possible to estimate the average error in the predicted number of clicks that an advertiser receives over many impressions. We illustrate these points in this appendix.

First we explain why it is not possible to precisely identify the true average errors in a model's predicted click-through rates on an impression by impression basis. To see this, suppose we are in a setting in which, on average, half of impressions are clicked and we make use of a model that always predicts a probability of a click equal to $\frac{1}{2}$. Here there are at least two possibilities that will be consistent with the data that we have observed.

One possibility is that the true probability of a click on every single impression is $\frac{1}{2}$. In this case, the average error in our machine learning system is 0. But another possibility is that there is some feature which we do not observe that is present in a random half of the auctions such that the probability of a click is 1 if the feature is present and 0 otherwise. In this case, the average error in our predicted click-through rates is substantial. Since both of these possibilities are consistent with the data, one cannot identify the true average errors in a model's predicted click-through rates on an impression by impression basis in this example. And more generally, since the true probabilities of a click on any given impression are unobserved, it is not possible to identify the true average errors in a model's predicted click-through rates on an impression by impression basis.

But it is possible to estimate the average error in the advertisers' predicted numbers of clicks over many impressions. To see this, let c denote the actual number of clicks that an advertiser has received, let q denote the number of clicks that the advertiser was predicted to receive, and let p denote the advertiser's true expected number of clicks given the true underlying probabilities with which each of the impressions are clicked. If an advertiser has received many impressions, then given the actual value of p, c will be a random variable that can be reasonably modeled as a draw from a Poisson distribution with parameter p. Thus

$E[c|p] = p$ and $Var[c|p] = E[(c-p)^2] = p$. Also, if q is an unbiased estimate of p, then $E[q-p] = 0$.

To estimate the average error in the advertisers' predicted number of clicks, we wish to estimate the value of $E[(p-q)^2]$, where the expectation is taken over different advertisers. Note that $E[(c-q)^2] = E[(c-p+p-q)^2] = E[(c-p)^2] + 2E[(c-p)(p-q)] + E[(p-q)^2]$ and $E[(c-p)(p-q)] = 0$ since $E[c-p] = 0$ and $c-p$ is a random variable that is uncorrelated with the random variable $p-q$ (because conditional on p, the actual number of clicks an ad receives is uncorrelated with the prediction errors made by the machine learning system). Thus this expression for $E[(c-q)^2]$ simplifies to $E[(c-q)^2] = E[(c-p)^2] + E[(p-q)^2]$, which in turn implies that $E[(p-q)^2] = E[(c-q)^2] - E[(c-p)^2] = E[(c-q)^2] - E[p] = E[(c-q)^2 - p]$.

Now for any given advertiser, we observe the value of $(c-q)^2$ because we observe this advertiser's actual number of clicks c and the advertiser's predicted number of clicks q. We also observe an unbiased estimate of p because an advertiser's actual number of clicks c is an unbiased estimate of p. Thus we can compute an unbiased estimate of $(c-q)^2 - p$ for any given advertiser, which we can in turn use to estimate the value of $E[(p-q)^2] = E[(c-q)^2 - p]$ by averaging over a large number of advertisers. By doing this, we can reasonably approximate the average percentage errors in the advertisers' predicted number of clicks.

These average percentage errors will differ for different buckets of advertisers, as there will be smaller percentage errors in the predicted number of clicks for advertisers with larger amounts of data. We believe it is plausible that the percentage errors in predicted click-through rates on an impression by impression basis would be comparable to the estimated percentage errors in predictions of the total number of clicks for new ads. While the prediction errors for new ads are larger than the prediction errors for other ads, it is also more difficult to predict click-through rates on an impression by impression basis than it is to estimate the average number of clicks for an ad over many impressions. Thus it is not clear whether the true average percentage errors in predicted click-through rates are greater or lower than the average errors in predictions of the total number of clicks for new ads. The estimates given in footnote 8 thus seem like a reasonable point to analyze.