A Bayesian Approach to Efficient Ranking in Sponsored Search

Sébastien Lahaie and R. Preston McAfee

Yahoo! Research
{lahaies, mcafee}@yahoo-inc.com

Abstract. In the standard model of sponsored search auctions, an ad is ranked according to the product of its bid and its estimated click-through rate (known as the quality score), where the estimates are taken as exact. This paper re-examines the form of the efficient ranking rule when uncertainty in click-through rates is taken into account. We provide a sufficient condition under which applying an exponent—strictly less than one—to the quality score improves expected efficiency. The condition holds for a large class of distributions known as natural exponential families, and for the lognormal distribution. An empirical analysis of Yahoo's sponsored search logs suggests that exponent settings substantially smaller than one can be efficient for both high and low volume keywords, implying substantial deviations from the traditional ranking rule.

1 Introduction

Sponsored search is today considered one of the most effective marketing vehicles available online. As the stakes have grown, the auction mechanism has seen several revisions over the years to improve efficiency and revenue. When first introduced by GoTo in 1998, ads were ranked purely by bid; later, in 2002, Google adopted the mechanism and introduced a quality score to weigh bids in proportion to clicks received [5], a practice now shared by every major search engine. In the basic model of sponsored search auctions [10], the quality score corresponds to an ad's position-normalized click-through rate (CTR). Under the assumption that CTRs are measured *exactly*, it is simple to verify that ranking ads in order of quality score times bid is economically efficient.

In this paper we re-examine the form of the efficient ranking rule, taking into account the inherent uncertainty in CTR estimates. Even for high-volume keywords, CTRs are notoriously difficult to estimate because clicks are rare events and new ads constantly enter the system. We consider a parametrized family of ranking rules that order ads according to scores of the form $e^{\gamma}b$, where e is the estimated position-normalized CTR, b is the bid, and $\gamma \in [0, 1]$. This family was first introduced by Lahaie and Pennock [9], who showed that settings of γ strictly less than 1 can improve *revenue*. Their model assumes that CTR estimates are exact. In this work we show that, in the presence of CTR uncertainty, using γ less than 1 can be justified on *efficiency* grounds.

Our main result identifies a sufficient condition under which setting γ strictly less than 1 improves efficiency. The condition relates quality scores based on historical click data (e.g., taking e to be the empirical CTR, normalized for position) to a Bayes estimator of the CTR. We show that the condition holds for a wide class of distributions known as natural exponential families, which includes the normal, Poisson, gamma, and binomial distributions among others. We further show that it holds for the lognormal distribution, which we found to be the best model of Yahoo's CTR estimates. We observe that γ is linked to the concept of *shrinkage* in Bayesian inference [4], and draw on this connection to empirically estimate the efficient γ for several keywords in Yahoo's sponsored search market. Our empirical analysis suggests that settings of γ substantially smaller than 1 can be efficient for both high and low volume keywords.

The remainder of the paper is organized as follows. Section 2 introduces the model, including the manner in which we incorporate uncertainty in CTR estimates. Section 3 proves the result that identifies when using γ less than 1 improves efficiency. Section 4 shows that the result applies to natural exponential families as well as the lognormal distribution; it also provides concrete examples of the efficient ranking rules for the beta and lognormal distributions. Section 5 reports on our data analysis of Yahoo's sponsored search logs to uncover the efficient settings of γ in practice. Section 6 concludes.

2 The Model

In this paper we restrict our attention to a single keyword, with a fixed set of agents competing for ad placement whenever a query on the keyword is performed. There are K slots on the page to be allocated among N agents, where $N > K$. In a sponsored search auction each agent i places a bid b_i, and the ads are ranked in decreasing order of $w_i b_i$ where w_i is a weight, or *quality score*, assigned by the search engine. When an ad is clicked, the corresponding agent pays the lowest bid it could have placed while maintaining its position; this is known as the *second-price* payment rule.

While the second-price rule amounts to the Vickrey payment with a single slot, this is no longer the case with multiple slots, and it is well-known that for $K > 1$ sponsored search auctions are not truthful [1]. In general an agent has an incentive to shade its bid b_i below its true value per click (i.e., willingness to pay) v_i. Nonetheless, under the widely accepted solution concept of *envy-free equilibrium* [3, 14], it is the case that agents bid in such a way that they are ranked according to $w_i v_i$, because $w_i b_i$ is an increasing function of $w_i v_i$. Therefore, in what follows, our results and statements in terms of bids will continue to hold if these are replaced with values, assuming envy-free equilibrium, and we can set aside incentive concerns to focus on the problem of efficient ranking.

The determine an efficient ranking the search engine develops an estimate of the *click-through rate* (CTR) c_{ij} that ad i would obtain if placed in slot j. We assume that CTRs are *separable*, meaning they factor according to $c_{ij} = e_i x_j$ into an advertiser effect e_i and a position effect x_j. Because clicks are stochastic,

the advertiser effect is treated as a random variable that follows a probability model $e_i \sim p(\cdot|\theta_i)$, parametrized by θ_i, with mean $\mu_i = \mathbf{E}[e_i \,|\, \theta_i]$. Position effects could also be modeled as random variables in principle, but in this work we treat them as known constants.

While separability is only an approximation to actual CTR patterns [2], it is still relevant for the search engine to estimate position-normalized advertiser effects because $w_i = \mu_i$ is a natural choice for the quality score. If $s : K \to N$ is an allocation of slots, where slot j goes to agent $s(j)$, then under separability the efficiency of the allocation is:

$$\mathbf{E}\left[\sum_{j=1}^{K} x_j e_{s(j)} b_{s(j)} \,\middle|\, \theta_1, \ldots, \theta_N\right] = \sum_{j=1}^{K} x_j \mu_{s(j)} b_{s(j)}.$$

As it is (typically) the case that $x_1 > x_2 > \cdots > x_K$, it is then efficient to take $w_i = \mu_i$ and rank agents in decreasing order of $\mu_i b_i$ [8]. In this work, we relax the assumption that the probability model for each e_i is known exactly and consider how this uncertainty can affect the form of the efficient ranking rule. When discussing CTR modeling, we will often suppress the subscript i when not referring to a specific advertiser, as we do until the end of this section.

To incorporate uncertainty in the probability model due to limited data, we introduce a prior $\theta \sim q(\cdot)$ on the model parameter. Given a vector of m observations $\mathbf{e} = (e^1, \ldots, e^m)$ for the advertiser effect, a generic approach to ranking is to compute a statistic $t(\mathbf{e})$ of the data, and set the weight w to be a function of the statistic. For instance, one could compute the maximum likelihood estimate $\hat{\theta}(\mathbf{e})$ given the data and use the corresponding statistic

$$t_M(\mathbf{e}) = \mathbf{E}[e \,|\, \hat{\theta}(\mathbf{e})] \tag{1}$$

as a weight in order to rank the agents. We will refer to (1) as the *maximum likelihood statistic*. This is often straightforward to compute (e.g., for distributions such as the Bernoulli, normal, and Poisson it is the empirical mean). The maximum likelihood approach is unbiased as the amount of data grows, but in practice click observations are limited. To properly incorporate uncertainty in the presence of limited data, we can instead use a Bayesian approach. In this case the parameter distribution is updated via Bayes rule which sets $q(\theta|\mathbf{e}) \propto p(\mathbf{e}|\theta)q(\theta)$, where $p(\mathbf{e}|\theta) = \prod_{i=1}^{m} p(e^i|\theta)$, and the posterior mean is then

$$t_B(\mathbf{e}) = \mathbf{E}[e \,|\, \mathbf{e}] = \int_{\Theta} \mathbf{E}[e \,|\, \theta]\, q(\theta \,|\, \mathbf{e})\, d\theta, \tag{2}$$

where Θ is the domain of the parameter θ. We will refer to (2) as the *Bayes statistic*. While this statistic leads to efficient ranking incorporating all uncertainty, it can be more challenging to compute depending on the probability model for advertiser effects and the prior used.

In the remainder of the paper we will focus our attention on ranking rules that set $w = t(\mathbf{e})^\gamma$ for $\gamma \in [0, 1]$. With $\gamma = 1$, using statistic (2) is efficient,

and using statistic (1) is efficient in the limit as the amount of data grows. This is the usual form of ranking rule used in sponsored search, taking the statistic as a quality score. At $\gamma = 0$, on the other hand, we rank purely by bid, a rule that was used in the very first sponsored search auctions [5]. As we will see, the virtue of this class of ranking rules is that it allows one to use γ to incorporate uncertainty into the ranking, increasing efficiency, while using simpler statistics such as (1) rather than (2).

Formally, assuming bids have been fixed, a *ranking rule* σ defines an allocation of slots to agents for every set of observations $\underline{\mathbf{e}} = (\mathbf{e}_1, \ldots, \mathbf{e}_N)$ of advertiser effects, so that $\sigma(\cdot\,; \underline{\mathbf{e}}) : K \to N$. The expected efficiency of a ranking rule is defined as

$$\mathbf{E}\left[\sum_{j=1}^{K} x_j t_B(\mathbf{e}_{\sigma(j;\underline{\mathbf{e}})}) b_{\sigma(j;\underline{\mathbf{e}})}\right],$$

where the expectation is with respect to the distribution over sampled observations. In what follows, we use $V(\gamma)$ to denote the expected efficiency of the ranking rule that uses $w = t(\mathbf{e})^\gamma$ to weigh bids, for a given statistic t. We are interested in the settings of γ that are most efficient.

3 Main Condition

Our main result provides a sufficient condition for the use of a $\gamma < 1$ exponent on the chosen ranking statistic $t(\mathbf{e})$ on efficiency grounds, rather than revenue grounds as in Lahaie and Pennock [9]. Intuitively, the exponent reflects the contribution of the prior in the Bayes statistic (2). For the sake of simplicity we state the theorem for the case of two agents and one slot ($N = 2$, $K = 1$), and for this case we can ignore position effects.

Theorem 1. *Assume that agents are ranked according to $t(\mathbf{e}_i)$ for $i = 1, 2$. Then we have $V'(1) < 0$ if the quantity*

$$\frac{\mathbf{E}\left[t_B \,|\, t\right]}{t} \tag{3}$$

is decreasing *in the statistic $t \equiv t(\mathbf{e})$, where $t_B \equiv t_B(\mathbf{e})$.*

Proof. To simplify notation, we write μ_i for random variable $t_B(\mathbf{e}_i)$, and t_i as short-hand for $t(\mathbf{e}_i)$. Let $f(t_i, \mu_i)$ denote the joint distribution between the ranking statistic and the Bayes statistic, for $i = 1, 2$, and let f_t and f_μ be the marginals; variables with different subscripts are independently distributed. Agent 1 is chosen over agent 2 if $t_1^\gamma b_1 > t_2^\gamma b_2$, or $t_1 > t_2 \left(b_2/b_1\right)^{1/\gamma}$. The expected efficiency can be written as

$$V(\gamma) = \mathbf{E}[\mu_2 b_2] + \mathbf{E}[\mu_1 b_1 - \mu_2 b_2]\mathbf{1}_{\left\{t_1 > t_2 (b_2/b_1)^{1/\gamma}\right\}},$$

where $\mathbf{1}_A$ is the characteristic function of the set A. Differentiating with respect to γ, we obtain

$$V'(\gamma) = \mathbf{E}\left[(\mu_1 b_1 - \mu_2 b_2) t_2 \, (b_2/b_1)^{1/\gamma} \, \frac{1}{\gamma^2} \log(b_2/b_1)\right], \tag{4}$$

where the expectation is over the random variables μ_1, μ_2, and t_2. Evaluating this at $\gamma = 1$, we obtain

$$\begin{aligned} V'(1) &= \mathbf{E}\left[(\mu_1 b_1 - \mu_2 b_2) t_2 \, (b_2/b_1) \log(b_2/b_1)\right] \\ &= b_2 \log\left(\frac{b_2}{b_1}\right) \mathbf{E}\left[\left(\mu_1 \frac{b_1}{b_2} - \mu_2\right) t_2 \, (b_2/b_1)\right] \\ &= b_2 \log\left(\frac{b_2}{b_1}\right) \mathbf{E}\left[\left(\mu_1 t_2 - \mu_2 t_2 \left(\frac{b_2}{b_1}\right)\right)\right] \\ &= b_2 \log\left(\frac{b_2}{b_1}\right) \mathbf{E}\left[(\mu_1 t_2 - \mu_2 t_1)\right]\Big|_{t_1 = t_2(b_2/b_1)} \end{aligned} \tag{5}$$

Let M and T denote the domains of definition for variables μ_i and t_i respectively $(i = 1, 2)$. We now have

$$\begin{aligned} \mathbf{E}\left[\mu_1 t_2\right] &= \int_T \int_M \int_M \mu_1 t_2 f(\mu_1, t_1) f(\mu_2, t_2) \, d\mu_2 \, d\mu_1 \, dt_2 \\ &= \int_M \int_M \mu_1 t_2 f(\mu_1, t_1) f_t(t_2) \, d\mu_1 \, dt_2 \\ &= \int_M \int_M \mu_1 t_2 f_\mu(\mu_1 | t_1) f_t(t_1) f_t(t_2) \, d\mu_1 \, dt_2 \\ &= \int_M t_2 f_t(t_1) f_t(t_2) \mathbf{E}[\mu_1 | t_1] \, dt_2 \\ &= \mathbf{E}\left[f_t(t_1) t_2 \mathbf{E}[\mu_1 | t_1]\right]. \end{aligned} \tag{6}$$

The outer expectation in the latter is with respect to t_2. By an analogous derivation we find that

$$\mathbf{E}\left[\mu_2 t_1\right] = \mathbf{E}\left[f_t(t_1) t_1 \mathbf{E}[\mu_2 | t_2]\right]. \tag{7}$$

Combining (6) and (7), we find that the expectation in (5) evaluates to

$$\begin{aligned} \mathbf{E}\left[(\mu_1 t_2 - \mu_2 t_1)\right] &= \mathbf{E}\left[f_t(t_1)(t_2 \mathbf{E}[\mu_1 | t_1] - t_1 \mathbf{E}[\mu_2 | t_2])\right] \\ &= \mathbf{E}\left[f_t(t_1) t_1 t_2 \left(\frac{\mathbf{E}[\mu_1 | t_1]}{t_1} - \frac{\mathbf{E}[\mu_2 | t_2]}{t_2}\right)\right]. \end{aligned} \tag{8}$$

Recall that this is evaluated at $t_1 = t_2(b_2/b_1)$. Now assume $b_1 > b_2$, so that $t_1 < t_2$. Under condition (3) we see from (8) that the expectation term in (5) is positive, while the leading term $b_2 \log(b_2/b_1)$ is negative, so (5) is negative. By a symmetric argument, the derivative (5) is negative when $b_1 < b_2$, which completes the proof.

The conditions given in the theorem imply that efficiency is improved by using $\gamma = 1 - \epsilon$ rather than $\gamma = 1$, for some $\epsilon > 0$. The theorem does not claim that using $t(\mathbf{e})^\gamma$ as a weight, with a properly chosen $\gamma < 1$, is exactly efficiency. When using a statistic such as the empirical advertiser effect for ranking, the condition that (3) be decreasing should hold, intuitively, because t_B is a mixture of the empirical effect and the prior. Therefore the expectation t_B should not respond strongly to a change in the observation t. This intuition is corroborated for a large class of distributions in the next section.

4 Exponential Families

To usefully apply our main theorem, one needs the ability to evaluate the expectation of the Bayes statistic given the value of the ranking statistic used in practice. As suggested in Section 2, a convenient choice for the latter is the maximum likelihood statistic, which often evaluates to the empirical mean of the observed advertiser effects. In this section we consider a rich collection of distributions, known as *exponential families*, to which the theorem applies and which cover most of the standard distributions one might use for CTR modeling. Exponential families have closed forms for the maximum likelihood statistic, and have convenient conjugate priors which make the Bayes statistic tractable to analyze. The properties of exponential families that we introduce here are standard and can be found in [12, 15].

An exponential family is a parametrized distribution with density that takes the form

$$p(e|\theta) = f(e) \exp\left[\theta \cdot \phi(e) - g(\theta)\right]. \tag{9}$$

Here f is a base density over advertiser effects, and θ is known as the *natural parameter*. The term $\phi(e)$ is the *sufficient statistic*. We will restrict our attention to families with scalar-valued sufficient statistics; this implies that the natural parameter θ is also a scalar. The term $g(\theta)$ is a normalizing constant given by

$$g(\theta) = \log \int f(e) \exp\left[\theta \cdot \phi(e)\right] \, de.$$

The domain of the natural parameter is those θ for which the normalizer is finite: $\Theta = \{\theta : g(\theta) < +\infty\}$. It is known to be convex—for the case of a scalar natural parameter, the domain is a (possibly unbounded) interval. It is straightforward to check that the first derivative of the normalizer evaluates to the expectation of the sufficient statistic, a fact we will use later on:

$$g'(\theta) = \mathbf{E}\left[\phi(e) \,|\, \theta\right]. \tag{10}$$

In general, the maximum likelihood estimate $\hat{\theta}(\mathbf{e})$ for the natural parameter, given a vector of m observations $\mathbf{e} = (e^1, \ldots, e^m)$, cannot be evaluated analytically. However, the expectation of the sufficient statistic under this estimate is simply

$$\mathbf{E}[\phi(e) \,|\, \hat{\theta}(\mathbf{e})] = \frac{1}{m} \sum_{i=1}^{m} \phi(e^i), \tag{11}$$

namely the empirical mean of the sufficient statistic. An exponential family has a conjugate prior of the form

$$p(\theta|\nu, n) = \exp\left[\nu \cdot \theta - n \cdot g(\theta) - h(\nu, n)\right].$$

This is again an exponential family, but with a two-dimensional natural parameter (ν, n), and here $h(\nu, n)$ is the normalizing constant. Given the m observations $(e^1, \ldots, e^m)$, the parameters of the conjugate distribution are updated according to the rule:

$$n \leftarrow n + m$$
$$\nu \leftarrow \nu + \sum_{i=1}^{m} \phi(e^i)$$

Note that the latter parameter is essentially updated according to the maximum likelihood statistic (11). Therefore, exponential families provide a tractable form for the maximum likelihood statistic, and define a clear relationship between this statistic and the posterior distribution. This makes them amenable to the application of Theorem 1.

4.1 Natural Exponential Families

A *natural* exponential family is one where the sufficient statistic is simply $\phi(e) = e$. In this case, the maximum likelihood statistic coincides with the empirical mean, because according to (11) we have

$$t_M(\mathbf{e}) = \mathbf{E}[e \,|\, \hat{\theta}(\mathbf{e})] = \frac{1}{m} \sum_{i=1}^{m} e^i.$$

Many of the most prominent univariate distributions are natural exponential families, such as the normal, Poisson, gamma, exponential, Weibull, binomial, and Bernoulli distributions [12]. For all of these distributions, the condition (3) in our main theorem applies when using the maximum likelihood statistic for ranking, as the next result shows.

Proposition 1. *Assume advertiser effects are distributed according to a natural exponential family, and that advertisers are ranked according to weights* $t_M(\mathbf{e})^\gamma$. *Then there is an* $\epsilon > 0$ *such that using* $\gamma = 1 - \epsilon$ *improves expected efficiency over* $\gamma = 1$.

Proof. For succinctness let $\tilde{e} = \sum_{i=1}^{m} e_i$, and let $\bar{e} = \tilde{e}/m$. As just mentioned, $t_M(\mathbf{e}) = \bar{e}$ for a natural exponential family; denote this empirical mean by $\bar{e}$. We will show that (3) is decreasing in $\bar{e}$, and the result will then follow from Theorem 1. After a Bayes update, we have

$$\begin{aligned} \frac{\mathbf{E}[e \,|\, \bar{e}]}{\bar{e}} &= \frac{1}{\bar{e}} \int_\Theta \mathbf{E}[e \mid \theta]\, p(\theta \,|\, \nu + \tilde{e}, n + m)\, d\theta \\ &= \frac{1}{\bar{e}} \int_\Theta g'(\theta) \exp[(\nu + \tilde{e})\theta - (n + m)g(\theta) - h]\, d\theta \end{aligned} \tag{12}$$

$$= \frac{\nu}{(n+m)\bar{e}} + O(1)$$
$$-\frac{1}{(n+m)\bar{e}} \int_{\Theta} [(\nu + \tilde{e}) - (m+n)g'(\theta)] \exp[(\nu+\tilde{e})\theta - (n+m)g(\theta) - h]\, d\theta$$
$$= \frac{\nu}{(n+m)\bar{e}} - \frac{1}{(n+m)\bar{e}} \int_{\Theta} p'(\theta \,|\, \nu + \tilde{e}, n+m)\, d\theta + O(1). \tag{13}$$

In the above we have used h as short-hand for $h\,(\nu + \tilde{e}, n+m)$. Note that the first term in (13) is decreasing in $\bar{e}$. We will have proved condition (3) if we can establish that the second term vanishes. But this is the case because the posterior density integrates to 1, and therefore we have the identity

$$\frac{d}{d\theta} \int_{\Theta} p(\theta \,|\, \nu + \tilde{e}, n+m)\, d\theta = 0.$$

Interchanging the differentiation and integration operations, which is admissible because the posterior density is continuous, completes the proof.

To gain some intuition for the result, it is helpful to consider a concrete instance of a natural exponential family. In one interpretation of the separable CTR model, the position effect is the probability that the user will look at a slot, and the advertiser effect is the probability the ad is clicked given that it is viewed [8]. As clicks are binary events, the Bernoulli distribution—a natural exponential family—is then a straightforward choice of model for advertiser effects. Assume that $e \sim \text{Bernoulli}(p)$ and that $p \sim \text{Beta}(n\mu, n(1-\mu))$—the beta distribution is the conjugate prior for the Bernoulli. The mean of the latter is μ, while the empirical mean $\bar{e}$ is both the maximum likelihood statistic and a sufficient statistic for the Bayes update. After the update we have

$$p \,|\, \bar{e} \sim \text{Beta}\left(n\mu + m\bar{e}, n(1-\mu) + m(1-\bar{e})\right),$$

which has a mean of $\gamma\bar{e} + (1-\gamma)\mu$ where $\gamma = \frac{m}{n+m}$. Because the parameter p for the Bernoulli is its mean, the posterior mean of p is also the posterior mean of e. The term (3) in our main theorem therefore evaluates to

$$\gamma + (1-\gamma)\frac{\mu}{\bar{e}},$$

which is decreasing in $\bar{e}$, as expected. However, Theorem 1 only states that using some $\gamma < 1$ as an exponent on $\bar{e}$ improves efficiency here—it does *not* state that ranking according to $\bar{e}^{\gamma} b$ is efficient. The closed form solution to the update implies that to rank two bidders efficiently, we should make the comparison

$$b_1 \cdot [\gamma\bar{e}_1 + (1-\gamma)\mu] \overset{?}{>} b_2 \cdot [\gamma\bar{e}_2 + (1-\gamma)\mu], \tag{14}$$

which takes a linear rather than exponential form. We see that when the prior is uninformative ($n = 0$) or there is ample data ($m \to \infty$), then $\gamma \to 1$ and we rank by $\bar{e}b$. When there is no data, $\gamma = 0$ and we rank purely by bid. Note that to rank efficiently according to (14), one needs an estimate of the prior mean μ.

4.2 Lognormal Distribution

While the probability interpretation of the advertiser and position effects is intuitively appealing, in practice the search engine may use a different factorization of CTRs that does not lead to effects in $[0, 1]$. However, it is clear that the effects should be non-negative. The lognormal distribution has support on the positive reals and so could prove a convenient choice to model advertiser effects—this turned out to be the case in our empirical analysis, as we report in Section 5 later on. We will show in this section that Theorem 1 applies to this distribution as well; in fact, using a certain $\gamma \in (0, 1)$ exponent is *exactly* efficient for this distribution.

The lognormal is an exponential family, but not a *natural* exponential family, because it has sufficient statistic $\phi(e) = \log e$. Recall that an effect e is lognormal if $\log e \sim \mathcal{N}(\mu, \sigma_e^2)$. We assume the variance is known, and that $\mu \sim \mathcal{N}(\nu, \sigma_\mu^2)$—the normal distribution is the conjugate prior for the normal. Given m observations, let $\bar{\ell} = \frac{1}{m}\sum_{i=1}^{n} \log e_i$ denote the empirical mean of the sufficient statistic. Let $\hat{e} = \left(\prod_{i=1}^{m} e_i\right)^{1/m}$ denote the geometric mean of the observations, and observe that we have $\hat{e} = \exp(\bar{\ell})$. It is known that the expected value of $\exp(y)$ for $y \sim \mathcal{N}(\mu, \sigma^2)$ is $\exp(\mu + \sigma^2/2)$, so we have

$$t_M(\mathbf{e}) = \exp(\bar{\ell} + \sigma_e^2/2) = \hat{e} \cdot \exp(\sigma_e^2/2). \tag{15}$$

That is, the maximum likelihood statistic is proportional to the geometric mean, so the latter is a natural ranking statistic in this context. On the other hand, letting $\tau_e = \sigma_e^{-1}$ and $\tau_\mu = \sigma_\mu^{-1}$, the Bayes update leads to the posterior

$$\mu \,|\, \bar{\ell} \sim \mathcal{N}\left((1-\gamma)\nu + \gamma\bar{\ell},\ \left(\tau_\mu^2 + \tau_e^2\right)^{-1}\right), \tag{16}$$

where $\gamma = m\tau_e^2/(\tau_\mu^2 + m\tau_e^2)$. A straightforward evaluation of (2) therefore gives

$$\begin{aligned} t_B(\mathbf{e}) &= \exp[(1-\gamma)\nu + \gamma\bar{\ell} + \sigma_\mu^2/2 + \sigma_e^2/2] \\ &= \hat{e}^\gamma \cdot \exp[(1-\gamma)\nu + \sigma_\mu^2/2 + \sigma_e^2/2] \end{aligned} \tag{17}$$

The next result is now immediate, but because of its relevance in practice we record it as a proposition.

Proposition 2. *Assume advertiser effects follow a lognormal distribution. Then ranking according to $\hat{e}^\gamma$, with $\gamma = m\tau_e^2/(\tau_\mu^2 + m\tau_e^2) \in (0, 1)$, maximizes expected efficiency.*

Proof. From (17) we see that $t_B(\mathbf{e})/\hat{e} \propto \hat{e}^{\gamma-1}$, which is decreasing in $\hat{e}$ because $\gamma < 1$. From (16), we see that $\hat{e} = \exp(\bar{\ell})$ is a sufficient statistic to perform the Bayes update, so $\mathbf{E}[t_B|\hat{e}] = t_B(\mathbf{e})$. Therefore we know from Theorem 1 that using some exponent strictly smaller than 1 on the geometric mean improves efficiency. However, we can in fact achieve exact efficiency, because when ranking two bidders we make the comparison

$$b_1 \cdot t_B(\mathbf{e}_1) \overset{?}{>} b_2 \cdot t_b(\mathbf{e}_2) \quad \Leftrightarrow \quad b_1 \cdot \hat{e}_1^\gamma \overset{?}{>} b_2 \cdot \hat{e}_2^\gamma \tag{18}$$

where $\gamma = m\tau_e^2/(\tau_\mu^2 + m\tau_e^2)$. This completes the proof.

When there is ample data ($m \to +\infty$) or the prior is uninformative ($\tau_\mu \to 0$), it is efficient to rank according to $\hat{e}b$. When there is no data ($m = 0$), we rank purely by bid. Note that in making the comparison (18), the contribution of the prior mean cancels out. This compares favorably to the linear form of the efficient ranking rule we derived for the beta distribution in (14), where it is necessary to estimate the prior mean; however, the prior variance is still needed to determine the efficient γ.

5 Empirical Data Analysis

In this section we report on an empirical analysis of Yahoo's sponsored search logs to get a sense of the settings of γ that are efficient in practice. The theory so far has established that, under reasonable modeling assumptions, using an exponent of $\gamma = 1 - \epsilon$ on the empirical advertiser effect would improve efficiency, for some $\epsilon > 0$. However, if the ϵ need only be very small according to the data, these results would have little bearing on real sponsored search auctions.

5.1 Data Description

We collected data by considering all the keywords in the month of June 2010 that had at least one advertisement. From these keywords we retained those where, over the month, the total number of clicks on ads was at least 2, and the average depth was at least 2. The depth of a query is the number of ads shown, which can range from 0 to 12 on Yahoo. The keywords were stratified into 10 deciles by search volume, and we randomly selected 20 from each decile for a total of 200 keywords. While the sampling is not proportional, we are not interested in aggregating statistics across deciles; proportional sampling would lead to a dataset overwhelmed by tail keywords with sparse click data.

For each ad shown on a keyword, and every position the ad was placed in, we have the total number of searches and clicks as well as the position effect. A position here is defined not just by the rank of the ad, but also where it was placed on the page (top, bottom, side), and how its competitors were laid out. For instance, showing an ad at the third rank when there are two ads at the top (i.e., first on the side) is not the same as showing the ad at that same rank when no ads are at the top (i.e., third on the side): the different positioning leads to a different position effect. There are a total of 60 distinct positions in our dataset. For each position we have a position effect hard-coded by Yahoo; while these were occasionally revised over the month, the changes were typically minimal. The relative standard deviations of the position effects over the month had a median of 0% and mean of 2% over the keywords and advertisers. We therefore take these effects as constants, consistent with our earlier assumptions.

Our dataset has 117K records, one for each keyword-ad-position triplet, and contains information on 19K distinct ads, for an average of 95 ads per keyword over the month and 587 records per keyword (naturally the distribution is heavily

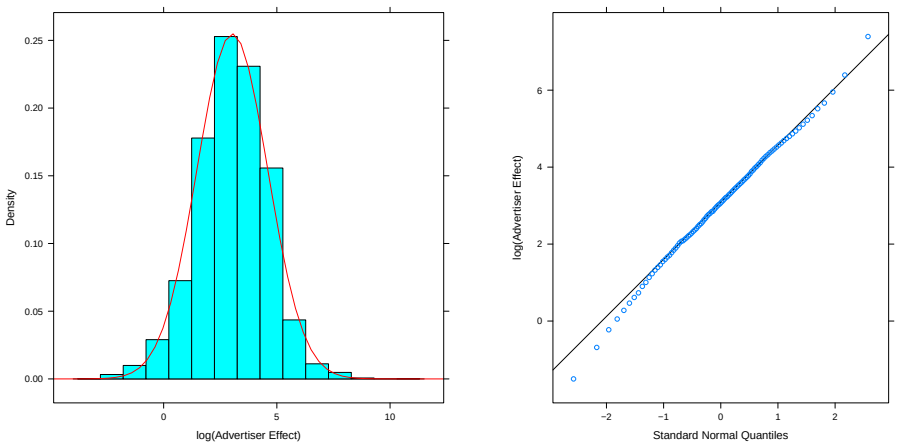

Fig. 1. Lognormality of the observed advertiser effects (position-normalized CTRs). The left panel shows the empirical distribution for ads that have at least one click over the month, together with the best-fit normal distribution. The right panel gives the theoretical quantile-quantile plot.

skewed). We define the observed advertiser effect for an ad at certain position on a given keyword as the position-normalized empirical click-through rate:

$$\frac{\text{clicks}}{\text{searches} \cdot \text{position effect}}$$

The observed effects do not all lie in $[0, 1]$: they have a median of 0.002 and mean of 8.12 in our data. Figure 1 indicates that the observed ad effects are well modeled by a lognormal distribution, restricting our attention to ads that received at least one click. For this probability model, the results of Section 4.2 show that there is in principle a setting of γ for each keyword that is exactly efficient.

5.2 Hierarchical Model

To empirically estimate the optimal γ for different keywords we develop a hierarchical Bayesian model of advertiser effects. We have seen through (16) that with the lognormal distribution (among others), γ can be viewed as the weight on the empirical advertiser effect in a convex combination between it and the prior mean. In Bayesian inference this is known as the *shrinkage* factor [4, 11], and we can obtain shrinkage estimates as a by-product of a hierarchical model.

We fit a model to each individual keyword. Given a keyword, the units are ad-position pairs i, and we denote the position-normalized empirical CTR for this pair by y_i. Let $j[i]$ denote the ad in unit i. We fit the following basic one-way

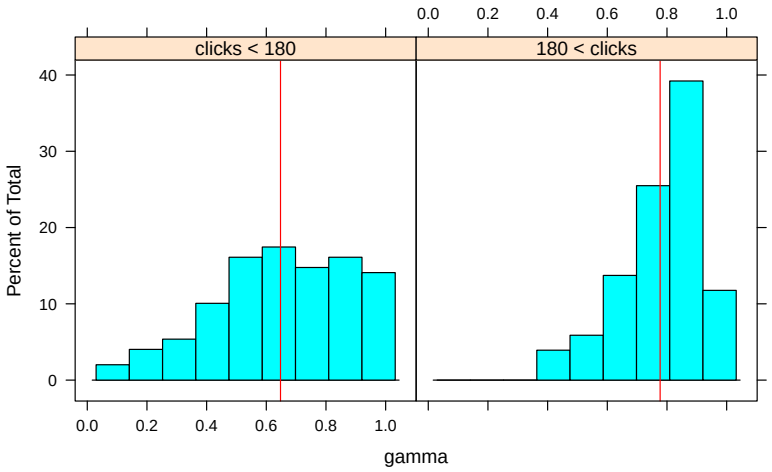

Fig. 2. Empirical distribution of estimated γ's for keywords with small and large numbers of clicks over the month. The reference lines indicate the means. For keywords with small numbers of clicks, the distribution is more uniform, whereas for keywords that attract many clicks γ skews towards 1.

hierarchical model [6]:

$$\log y_i \sim \mathcal{N}(\alpha_{j[i]}, \sigma_y^2) \tag{19}$$

$$\alpha_j \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2) \tag{20}$$

where i ranges over all the units and j over all the ads. (To avoid taking the log of 0, we recoded empirical effects of 0 to 10^{-5}, which is an order of magnitude smaller than the smallest positive observed effect in our dataset.) We assign uninformative uniform priors to σ_y, μ_α, and σ_α. The posterior distribution was evaluated using the Gibbs sampler provided by the JAGS program [13], and 1000 draws from the posterior were taken to estimate model statistics, in particular γ. For each draw γ was estimated using the following approach proposed by Gelman and Pardoe [7]. Letting $\epsilon_j = \alpha_j - \mu_j$ for each advertiser j, we set

$$\gamma = \frac{\mathbf{V}_j \mathbf{E}[\epsilon_j]}{\mathbf{E}[\mathbf{V}_j \epsilon_j]}, \tag{21}$$

where $\mathbf{V}$ represents the finite-sample variance operator, $\mathbf{V}_j \epsilon_j = \frac{1}{n-1} \sum_j (\epsilon_j - \bar{\epsilon}_j)$, and $\mathbf{E}$ in this context is the finite-sample mean. The denominator in (21) is the unexplained component of the variance in the α_j's, while the numerator is the variance among the point estimates of the ϵ_j's. We will have γ close to 1 if the latter is large relative to the former, meaning that α_j's usually lie closer to the empirical mean of the advertiser's effect. On the other hand, if the latter is small relative to the former, then the estimated α_j cluster more closely to μ_j and so

the prior mean is given higher weight. Gelman and Pardoe [7] demonstrate that (21) can be viewed as a Bayesian analog to the definition of γ we saw earlier: $\gamma = m\tau_e^2/(\tau_\mu^2 + m\tau_e^2)$. We report on the mean γ evaluated according to (21) over the 1000 draws.

Figure 2 shows the distribution of the resulting γ's over the 200 keywords. We identified different patterns in the distribution depending on whether we consider low or high click keywords; here high means greater than 180 clicks per month, or 6 clicks per day on average. For low click keywords the distribution of γ is more uniform, with mean and median both at 0.64. High click keywords see γ more skewed towards 1, as one would intuitively expect, with a mean of 0.78 and a median of 0.82. Note that under both regimes the mean is substantially below 1, which suggests that using a rule of the form $e^\gamma b$ could improve efficiency for many keywords.

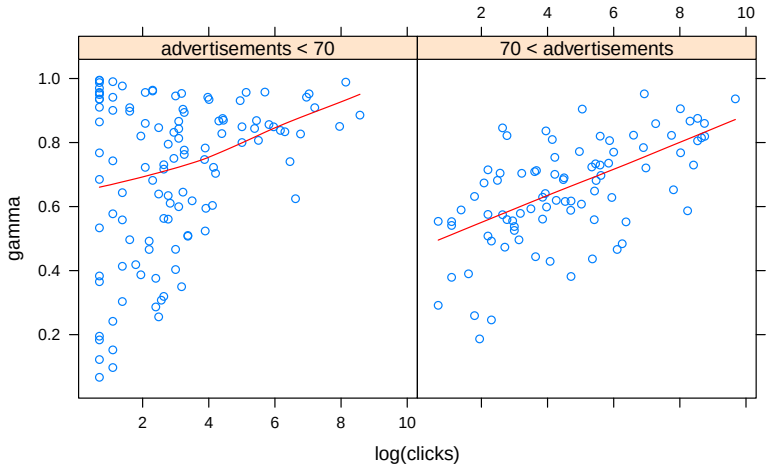

Fig. 3. Estimated γ for keywords with small and large numbers of advertisers over the month. The Loess curves show that under both regimes γ increases on average as the keyword receives more clicks, but for keywords with small numbers of advertisers and clicks there is substantial variability.

Figure 3 shows the empirical results from a different perspective. We again have two different regimes: keywords with few and many ads. Here a keyword has many ads if more than 70 distinct ads were shown over the month. For keywords with many ads there is a clear relationship between the volume of clicks and γ. This is intuitive since more clicks means more accurate CTR estimates. For keywords with few ads there is still a general upward trend, but there is substantial variability in the γ estimates, attributable to the dearth of data. In both cases the most relevant range for tuning γ seems to be $[0.6, 1]$.

6 Discussion

To conclude let us discuss a few limitations and extensions of this analysis. A key assumption implicit in the use of (21), and throughout the paper, is that each ad sees the same amount of observations m. In practice this is of course not the case, especially as ads are constantly added to the system. With uneven amounts of data among ads on a keyword, the estimate (21) amounts to a weighted combination of the different shrinkage factors for the individual ads. To rank efficiently, one would have to use ad-specific γ's. This is not very appealing because the contribution of the prior mean in (17) no longer cancels out in the comparison (18), leading to a more complicated ranking rule. A better understanding of the efficiency trade-offs between keyword- and ad-specific γ's is in order.

In our analysis, we base our estimate of the shrinkage factor γ on the empirical advertiser effects, but in practice the search engine uses machine-learned effects to rank. While these correlate well with realized advertiser effects, it would be informative to understand exactly how γ should be set given the search engine's estimates. One possibility is to introduce them into (19) as a linear predictor for realized effects. However, the resulting γ from such a model would not be the recommended exponent for the machine-learned effects. In fact, because the predictor would reduce the errors in the numerator of (19), this would misleadingly pull (21) towards 0. Developing sound ways to estimate γ with machine-learned effects is an important next step in this line of research.

References

[1] Gagan Aggarwal, Ashish Goel, and Rajeev Motwani. Truthful auctions for pricing search keywords. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, pages 1–7, 2006.

[2] Susan Athey and Denis Nekipelov. A structural model of sponsored search advertising auctions. Technical report, Microsoft Research, May 2010.

[3] Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. Internet advertising and the Generalized Second Price auction: Selling billions of dollars worth of keywords. *American Economic Review*, 97(1), March 2007.

[4] Bradley Efron and Carl Morris. Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, June 1975.

[5] Daniel C. Fain and Jan O. Pedersen. Sponsored search: A brief history. In *Second Workshop on Sponsored Search*, 2006.

[6] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2003.

[7] Andrew Gelman and Iain Pardoe. Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, 48(2):241–251, May 2006.

[8] Sébastien Lahaie. An analysis of alternative slot auction designs for sponsored search. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, pages 218–227, 2006.

[9] Sébastien Lahaie and David M. Pennock. Revenue analysis of a family of ranking rules for keyword auctions. In *Proceedings of the 8th ACM Conference on Electronic Commerce*, pages 50–56, 2007.

[10] Sébastien Lahaie, David M. Pennock, Amin Saberi, and Rakesh V. Vohra. Sponsored search auctions. In Noam Nisan, Tim Roughgarden, Éva Taros, and Vijay V. Vazirani, editors, *Algorithmic Game Theory*, pages 699–716. Cambridge University Press, 2007.

[11] Thomas A. Louis. Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association*, 79(386):393–398, June 1984.

[12] Carl N. Morris. Natural exponential families with quadratic variance functions. *The Annals of Statistics*, 10(1):65–80, March 1982.

[13] Martyn Plummer. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. `www-ice.iarc.fr/~martyn/software/jags/`.

[14] Hal R. Varian. Position auctions. *International Journal of Industrial Organization*, 25:1163–1178, 27.

[15] Martin J. Wainwright and Michael I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., 2008.