Position-Normalized Click Prediction in Search Advertising

Ye Chen
Microsoft Corporation
1065 La Avenida
Mountain View, CA 94043
yec@microsoft.com

Tak W. Yan
Microsoft Corporation
1065 La Avenida
Mountain View, CA 94043
takyan@microsoft.com

ABSTRACT

Click-through rate (CTR) prediction plays a central role in search advertising. One needs CTR estimates unbiased by positional effect in order for ad ranking, allocation, and pricing to be based upon ad relevance or quality in terms of click propensity. However, the observed click-through data has been confounded by positional bias, that is, users tend to click more on ads shown in higher positions than lower ones, regardless of the ad relevance. We describe a probabilistic factor model as a general principled approach to studying these exogenous and often overwhelming phenomena. The model is simple and linear in nature, while empirically justified by the advertising domain. Our experimental results with artificial and real-world sponsored search data show the soundness of the underlying model assumption, which in turn yields superior prediction accuracy.

Categories and Subject Descriptors

I.5.1 [**Pattern Recognition**]: Models—*Statistical*

General Terms

Algorithms, Theory, Experimentation

Keywords

Search advertising, factor models, exogeneity

1. INTRODUCTION

A click-through rate (CTR) prediction system for sponsored search advertising aims to estimate the CTR given a query-ad pair, typically along with other contextual knowledge such as about the user. The CTR prediction is pivotal for ad ranking, allocation, pricing, and the payoff of users and advertisers as well [7]. The estimated CTR serves as a measure of query-ad relevance, and hence should be made independent of other non-relevance factors. In practice, however, there are factors exogenous to the relevance-based click-through system, and often playing a dominant role in the observed click-through data. One classic example is the ad presentation position. A less cautious treatment of these exogenous factors may lead to a sub-optimal CTR prediction, and many real-world systems have been seen with these flaws.

We propose a probabilistic factor model as a general principled approach to studying these exogenous and often overwhelming phenomena. The model is simple and linear in nature, while empirically justified by the advertising domain. Extensive research has been undergone for correcting positional bias in algorithmic search, among which representative works are the examination model [12], the cascade model [5], and the dynamic Bayesian network (DBN) model [3], but less so for search advertising. Our approach adopts the same factorization assumption as in the examination model, that is, the probability of clicking on an item (result link for algorithmic search and ad for sponsored search) is the product of a positional prior probability and a relevance-based probability which is independent of position. Moreover, we specialize the concept of position into the ad domain, by incorporating other advertising specific yet important signals, e.g., the query-ad keyword match type and the total number of ads shown.

Other models originating from algorithmic search typically assume that the estimated CTR of an item is dependent on the relevance of items shown above on the search result page, as in the cascade and DBN models. These more sophisticated assumptions are appropriate for algorithmic search results where users have a high probability of clicking on one of the result links. For ads, however, the probability of clicking on ads generally is extremely low, usually a fraction of a percent. As a consequence, the effect of (not clicking) higher ads is a product of factors which are extremely close to one. In this case for example, the DBN positional prior reduces to a negative exponential function, which is a good fit to the empirical distribution derived from our factor model as described below.

2. THE FACTOR MODEL

Let i denote a query-ad pair, j be the ad position, c be the number of clicks, and v be the number of impressions. The observed CTR is an empirical conditional probability $p(\text{click}|i,j)$. We make the following simplifying yet classic assumptions in sponsored search advertising [2, 6, 11]:

1. Clicking an ad is independent of its position, given that it is physically examined;

KDD'12, August 12–16, 2012, Beijing, China.

2. Examining an ad is independent of its content or relevance, given its position.

Formally, the position-dependent CTR is factorized as

$$p(\text{click}|i,j) = p(\text{click}|\text{exam},i)p(\text{exam}|j). \quad (1)$$

The first factor $p(\text{click}|\text{exam},i)$, simply denoted as p_i, is a position-normalized CTR that represents the relevance of ad. The second factor $p(\text{exam}|j)$, simply denoted as q_j, reflects the positional bias. With this CTR factorization, we now proceed with two natural stochastic models of clicking behavior, and then arrive at the deployed model smoothed by a spike and slab prior [1, 9].

3. THE BINOMIAL MODEL

It is natural to assume that the number of clicks follows a binomial distribution

$$c_{ij} \sim \text{Binomial}(v_{ij}, p_i q_j), \forall i,j. \quad (2)$$

Given a training data set $D = \{(c_{ij}, v_{ij})\}$, we wish to learn the model parameters $\theta = (p,q)$, where p is a vector of relevance CTR p_i's and q is a vector of positional priors q_j's. The likelihood of the training data is

$$p(D|\theta) = \prod_{i,j} \binom{v_{ij}}{c_{ij}} (p_i q_j)^{c_{ij}} (1 - p_i q_j)^{v_{ij} - c_{ij}}. \quad (3)$$

The log likelihood is

$$\begin{aligned}\ell(\theta) = &\sum_{i,j} \left(\log \binom{v_{ij}}{c_{ij}} + c_{ij} \log(p_i q_j) \right. \\ &\left. + (v_{ij} - c_{ij}) \log(1 - p_i q_j) \right).\end{aligned} \quad (4)$$

We regard one of model parameters as latent variable, e.g., q, and derive an EM algorithm to estimate the MLE of both $\theta = (p,q)$. Notice though the objective function is not necessarily concave, even w.r.t. one univariate p_i or q_j, hence we only seek local optimal. By taking the partial derivatives w.r.t. p_i and q_j, respectively, we have

$$\frac{\partial \ell}{\partial q_j} = \sum_i \left(\frac{c_{ij}}{q_j} - \frac{(v_{ij} - c_{ij}) p_i}{1 - p_i q_j} \right); \quad (5)$$

$$\frac{\partial \ell}{\partial p_i} = \sum_j \left(\frac{c_{ij}}{p_i} - \frac{(v_{ij} - c_{ij}) q_j}{1 - p_i q_j} \right). \quad (6)$$

Since the model parameters are non-negative, we apply the following multiplicative recurrence [4, 10]

$$\text{E-step:} \quad q'_j \leftarrow \frac{\sum_i c_{ij}}{\sum_i (v_{ij} - c_{ij}) p_i / (1 - p_i q_j)}; \quad (7)$$

$$\text{M-step:} \quad p'_i \leftarrow \frac{\sum_j c_{ij}}{\sum_j (v_{ij} - c_{ij}) q_j / (1 - p_i q_j)}. \quad (8)$$

4. THE POISSON MODEL

If the number of trials n is sufficiently large and the success probability p is sufficiently small, $\text{Binomial}(n,p) \rightsquigarrow \text{Poisson}(np)$. Since ad is such a domain, we now derive the Poisson model which yields a similar yet more efficient update. The generative model is

$$c_{ij} \sim \text{Poisson}(v_{ij} p_i q_j), \forall i,j. \quad (9)$$

The data likelihood is

$$p(D|\theta) = \prod_{i,j} \frac{(v_{ij} p_i q_j)^{c_{ij}} \exp(-v_{ij} p_i q_j)}{c_{ij!}}. \quad (10)$$

The log likelihood is

$$\ell(\theta) = \sum_{i,j} (c_{ij} \log(v_{ij} p_i q_j) - v_{ij} p_i q_j - \log(c_{ij}!)). \quad (11)$$

By taking the partial derivatives w.r.t. p_i and q_j we have

$$\frac{\partial \ell}{\partial q_j} = \sum_i \left(\frac{c_{ij}}{q_j} - v_{ij} p_i \right); \quad (12)$$

$$\frac{\partial \ell}{\partial p_i} = \sum_j \left(\frac{c_{ij}}{p_i} - v_{ij} q_j \right). \quad (13)$$

We then derive the following multiplicative recurrence by leveraging the non-negativity of model parameters

$$\text{E-step:} \quad q'_j \leftarrow \frac{\sum_i c_{ij}}{\sum_i v_{ij} p_i}; \quad (14)$$

$$\text{M-step:} \quad p'_i \leftarrow \frac{\sum_j c_{ij}}{\sum_j v_{ij} q_j}. \quad (15)$$

It is clear, by comparing with the binomial recurrence, what the Poisson approximation implies for the EM recurrence. Moreover, it can be shown that the Poisson EM recurrence is globally convergent w.r.t. a univariate p_i or q_j, thus guaranteed to find a unique global maximum. When the E and M-steps are combined, however, the recurrence is still convergent but a global maximum is not guaranteed. Further, the E-step in Eq. (14) reveals the tempting "double-discounting" trap found in the naïve estimator as empirical positional CTR $q_j = \sum_i c_{ij} / \sum_i v_{ij}$.

5. THE GAMMA-POISSON MODEL

For both empirical and regularization purposes, we impose a gamma prior on the positional factor in the Poisson model

$$q_j \sim \text{Gamma}(\alpha, \beta), \forall j. \quad (16)$$

Empirically, the observed CTR is geometrically decreasing as the position lowers down [11], exhibiting a good fit to the gamma signature. In practice, inferior positions (e.g., bottom positions in side bar) might suffer from severe data sparsity, particularly clicks; thus regularizing or smoothing those noisy estimates will yield better generalization. The gamma distribution is a convenient choice, since it is a conjugate prior of Poisson.

The data likelihood (posterior for q) with a gamma prior is

$$\begin{aligned}p(D, q|p, \alpha, \beta) = &\prod_{i,j} \frac{(v_{ij} p_i q_j)^{c_{ij}} \exp(-v_{ij} p_i q_j)}{c_{ij!}} \\ &\times \prod_j \frac{q_j^{\alpha-1} \exp(-q_j/\beta)}{\beta^\alpha \Gamma(\alpha)}.\end{aligned} \quad (17)$$

The log likelihood is

$$\begin{aligned}\ell(\theta) = &\sum_{i,j} (c_{ij} \log(v_{ij} p_i q_j) - v_{ij} p_i q_j - \log(c_{ij}!)) \\ &+ \sum_j ((\alpha - 1) \log q_j - q_j/\beta - \alpha \log \beta - \log \Gamma(\alpha)).\end{aligned} \quad (18)$$

By taking the partial derivatives w.r.t. p_i and q_j we have

$$\frac{\partial \ell}{\partial q_j} = \sum_i \left(\frac{c_{ij}}{q_j} - v_{ij} p_i \right) + (\alpha - 1)/q_j - 1/\beta; \quad (19)$$

$$\frac{\partial \ell}{\partial p_i} = \sum_j \left(\frac{c_{ij}}{p_i} - v_{ij} q_j \right). \quad (20)$$

The EM multiplicative recurrence is

$$\text{E-step:} \quad q'_j \leftarrow \frac{\sum_i c_{ij} + (\alpha - 1)}{\sum_i v_{ij} p_i + 1/\beta}; \quad (21)$$

$$\text{M-step:} \quad p'_i \leftarrow \frac{\sum_j c_{ij}}{\sum_j v_{ij} q_j}. \quad (22)$$

The regularization is on q_j in E-step only and the interpretation of smoothing is obvious. When $\alpha = 1$ and $\beta \to \infty$, the gamma distribution approaches a uniform distribution, i.e., no prior.

6. THE CLICK MODEL

A click model or a CTR prediction model aims to estimate a positional-unbiased CTR for a given query-ad pair, i.e., the relevance CTR $p(\text{click}|\text{exam}, i)$ or p_i. The positional normalized click model described above does exactly this, as well as produces the positional priors $p(\text{exam}|j)$ or q_j. Another view of the factor model is a kNN model smoothed over ad positions; and when the feature space only contains query-ad pairs, $k = 1$. It is plausible that for query-ad pairs with sufficient historical clicks, the factor model might perform reasonably well. For a principled treatment of the cold-start problem, a trivial extension would be appending queries and ads unigram features into the feature space, and backing-off CTR estimates when new pairs are encountered in prediction time.

The positional normalized click model can also be applied independently of and in conjunction with other click models that estimate relevance-only CTR [1, 4]. More rigorously, we make the assumption that the positional factor is independent of other relevance factors. In model training, one shall normalize each ad impression by its positional prior $v_{ij} q_j$. In prediction, the CTR predictor learned from position-normalized training data produces exactly the relevance-only CTR, as we desire.

7. EXPERIMENTS

7.1 Simulation with synthetic data

We first simulated the Gamma-Poisson model on an artificial data set generated by a probabilistic model in the same spirit, i.e., given a sound model assumption. The synthetic data largely mimics the real-world search ad data, by carefully designed model parameters. Although the simulated data cannot fully reflect a real-world system, it has at least two advantages: (1) allowing for a quick study of the effects of a large number of parameters while abstracted from real-world noises; and (2) exposing the true distributions underlying the data to verify the learned model, which is impossible with real-world data.

The data was generated as follows:

1. $\forall$ position $j \in [1, \ldots, m]$, generate a $q_j \sim \text{Gamma}(\alpha, \beta)$, sort q in descending order, and scale q by $1/q_1$;
2. $\forall$ query-ad pair $i \in [1, \ldots, n]$, generate a $p_i \sim \text{Beta}(\gamma, \delta)$;
3. $\forall i$, generate a number of impressions $s_i \sim \text{Poisson}(\lambda)$;
4. $\forall i$, construct a multinomial distribution over positions $\phi_i \propto 1/\left(p_i/\mu(p_i)\right)^{j-1}$, to push good ads higher up;
5. $\forall i$, generate an impression allocation vector over positions $v_i \sim \text{Multinomial}(s_i, \phi_i)$, to form an $n \times m$ matrix of impression V;
6. Derive an $n \times m$ matrix of CTRs $Z = pq^\top$;
7. Derive an $n \times m$ matrix of Poisson means $Y = V.Z$, where '.' is element-wise multiplication;
8. Generate an $n \times m$ matrix of clicks $C \sim \text{Poisson}(Y)$, and $C \leftarrow \min(C, V)$, element-wise.

The underlying distributions are $\theta = (p, q)$. We used the same true distributions to generate a training set $D^{trn} = \{C, V\}$ and a testing set $D^{tst} = \{C', V'\}$. The parameters of the generative model are summarized in Table 1.

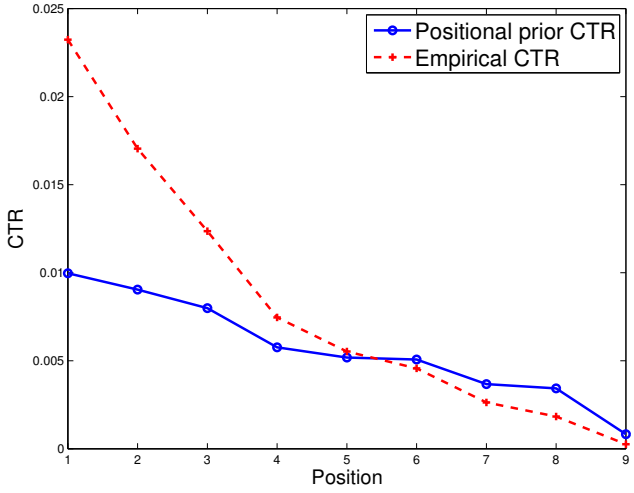
Table 1: Parameters of the generative model

Parameter	Description	Value
n	num of query-ad pairs	$100,000$
m	num of positions	9
(α, β)	Gamma parameters	$(1.2, 0.01)$
(γ, δ)	Beta parameters	$(1, 99)$
λ	Poisson mean of row sum s_i	$m \times 100$

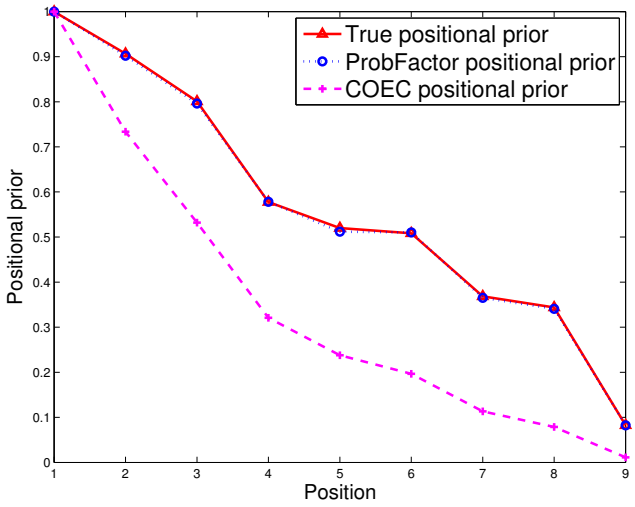
After generating the training data $D^{trn} = \{C, V\}$, we examined the empirical positional CTR $q_j^{emp} = \sum_i c_{ij} / \sum_i v_{ij}$ and the positional prior CTR $q_j^{pri} = \mu(p_i) q_j$. The latter will only realize were ads shown randomly. The coupling of positional bias with ad quality bias is evident from the empirical positional CTR curve (the dashed line) in Figure 1(a), in other words, higher-CTR ads tend to be shown in higher positions. A tempting trap is averaging CTR for each position as the positional prior, referred to as the COEC (clicks over expected clicks) model [8, 13]. This leads to the "double-discounting" problem, as revealed by the COEC learned positional priors (the dashed line) in Figure 1(b). The factor model decouples the positional and relevance factors in a principled way, hence perfectly recovers the true positional priors, as shown in Figure 1(b) (the solid and dotted lines).

Since we know the underlying true distributions $\theta = (p, q)$, we then compared the learned distributions $\hat{\theta} = (\hat{p}, \hat{q})$ with the true ones, as plotted in Figure 2. Both scatter plots for relevance CTR $p_i = p(\text{click}|\text{exam})$ and positional-biased CTR $p_{ij} = p(\text{click}|\text{position})$ show that the true and learned CTRs are well calibrated.

We then evaluated the trained factor model using the testing set $D^{tst} = \{C', V'\}$ generated from the same underlying distributions, with the only statistical fluctuations from Steps 3, 5, and 8, the Poisson and multinomial processes in the generative model. Figure 3(a) plots the true vs. learned relevance CTR $p_i = p(\text{click}|\text{exam})$ by both the factor and COEC models. In the higher CTR range ($p_i \geq \mu(p_i)$), both models calibrate well with the observed. In the lower CTR range ($p_i < \mu(p_i)$), however, both models tend to overestimate, in particular the COEC model performs worse. The first reason is legitimate and applies to both models, that

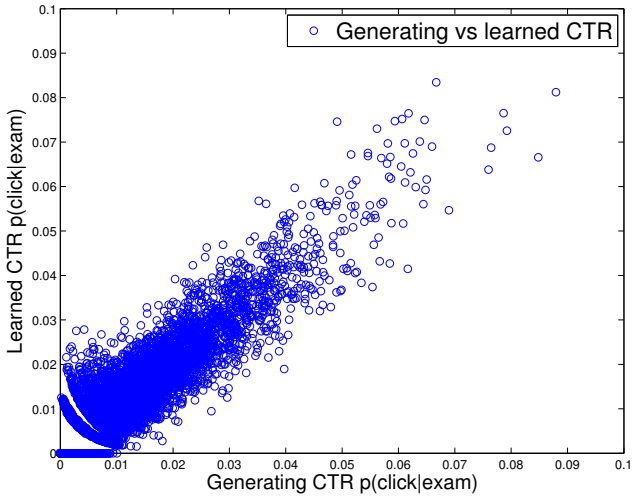


(a) Positional prior and empirical CTR

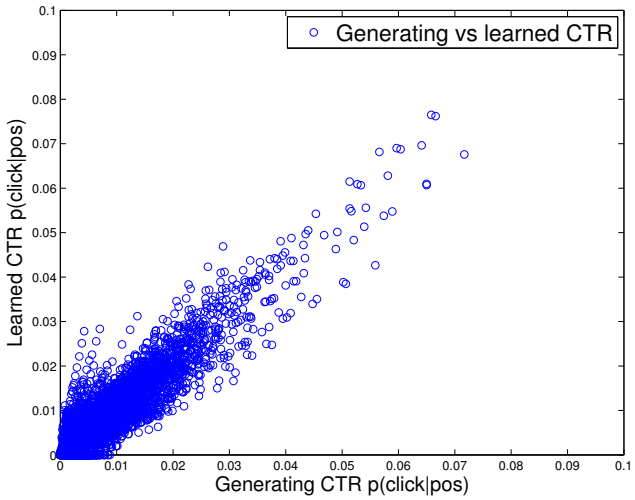


(b) True positional, Factor and COEC learned

Figure 1: Positional bias coupled with ad relevance bias in empirical data.



(a) True vs. learned $p_i = p(\text{click}|\text{exam})$



(b) True vs. learned $p_{ij} = p(\text{click}|\text{pos})$

Figure 2: True vs. learned CTR.

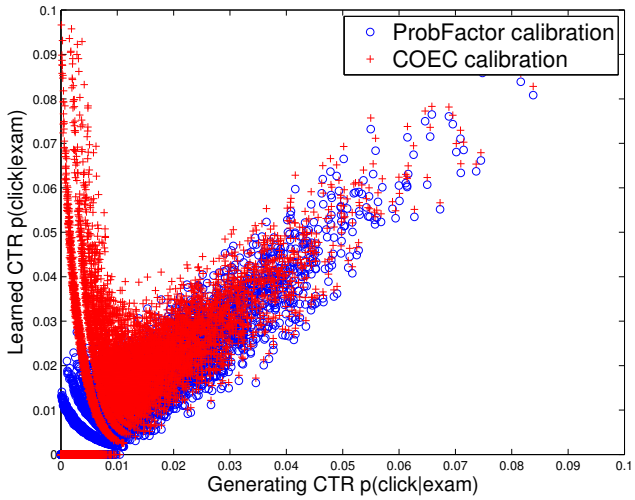
is, we impose a gamma prior on q_j and seek MAP estimate. The second reason exposes the flaw in the COEC model. Impressions with lower-than-average CTRs tend to be allocated to lower positions by the design of the generative process. The COEC model performs poorly because of the "double-discounted" positional effect, that is, q_j's for lower j's are over-penalized, and hence p_i's allocated to those lower j's (lower-CTR i's) are more widely overestimated. Figure 3(b) plots the observed vs. predicted positional biased CTR $p_{ij} = p(\text{click}|\text{position})$ for both the factor and COEC models. We observed a similar pattern as the relevance CTR calibration plots.

One primary objective of CTR prediction is to rank ads (RankScore = Bid × CTR)[1], hence we plot the ROC curves of click recall vs. view recall[2] in Figures 4 and 5. The ROC curves for the factor and COEC models across all positions are shown in Figure 4, and the area under the curve (AUC) results are summarized in the "AUC@all" column in Table 2. The factor model performs better than the COEC model by a small margin. The gain is small because ranking among different positions is relatively easy, given the strong positional effect.

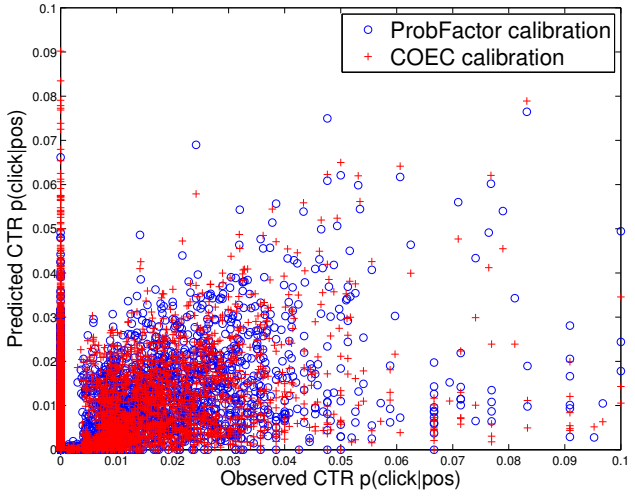
We further compared the ranking performance between the two models at individual positions, mainline (ML) 1 and 4, and sidebar (SB) 8, as plotted in Figure 5 and numerically reported in Table 2. Mainline refers to the positions above the algorithmic results, and sidebar refers to the positions right to the algorithmic results. The factor model performs better than the COEC model for every individ-

[1]Real-world sponsored search systems may have other variants such as CTR exponent and relevance terms, but our discussion carries over.

[2]The CTR term in rank score is estimated to capture the ad relevance in terms of click feedback, hence for evaluating a CTR predictive model, it is sufficient to rank ads by CTR estimates alone.



(a) True vs. learned $p_i = p(\text{click}|\text{exam})$



(b) Observed vs. predicted $p_{ij} = p(\text{click}|\text{pos})$

Figure 3: Calibratedness of Factor and COEC.

ual position, and the gain is widened as the position lowers down. The reason is actually revealed in Figure 1(b). Ranking ad position-dependent CTRs within a same position is essentially ranking their relevance-only CTRs, given a fixed positional prior. An over-penalized positional prior, as the COEC model estimates, will yield overestimated relevance CTRs, especially for those low-quality ads mostly pushed down. On the other hand, this error does not appear as a universal scaling for all ads, since some good-quality ads will be shown in higher positions sometimes. Consequently, the true relevance ranking would be more polluted in lower positions, in the presence of a poorly estimated positional prior.

Table 2: Simulation AUC of Factor and COEC

Model	AUC@all	AUC@ML1	AUC@ML4	AUC@SB8
Factor	0.8219	0.6130	0.5870	0.5899
COEC	0.8187	0.6045	0.5509	0.5467

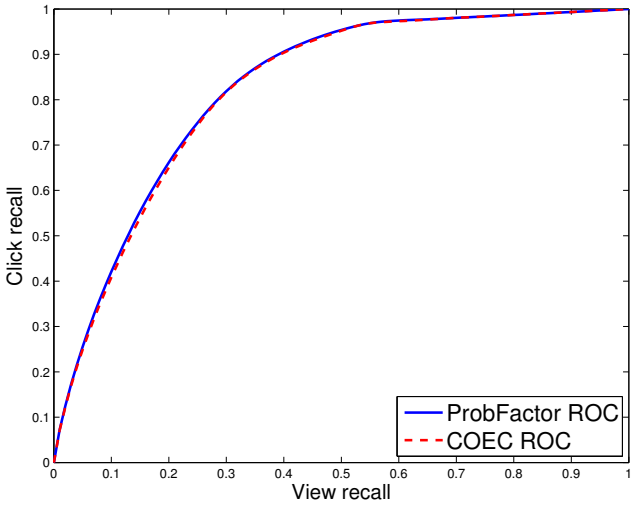


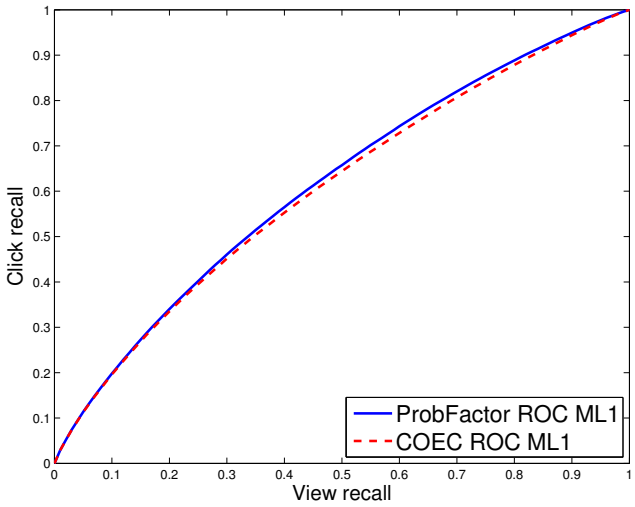
Figure 4: Simulation ROC curves for Factor and COEC.

7.2 Experiments with real-world data

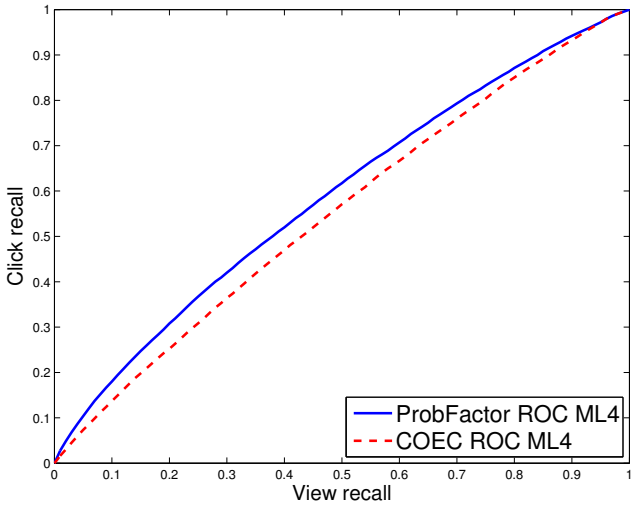
We then evaluated the Gamma-Poisson model on a real-world search advertising data set from Microsoft adCenter. We collected one week (10/01/2011-10/07/2011) ad click and impression count data for training, and the following day (10/08/2011) for testing. Each training example is uniquely identified by the composite key (UserId, Query, AdId, OrderItemId, Position, MainlineAds, MatchTypeId), and the data contains approximately $50M$ examples per day (for all users having Windows Live Id). A query-ad pair is uniquely identified by (Query, AdId, OrderItemId). The Position value is the absolute order from the ad showed in the top position. The MainlineAds value is the number of ads showed in mainline (also jargoned as Dude State from Yahoo!). An absolute position and a mainline ad count (or simply noted as dude) jointly determine the relative position, e.g., position = 3 along with dude = 4 gives ML3, while position = 3 when dude = 2 gives SB1[3]. In addition, we leverage the match type by which an ad is retrieved for a given query, e.g., exact match or broad match.

We maintain the individual identities of features (i.e., query, ad id, and order item id) present in a number of examples above a threshold $c * d$, where d is the number of training days and c is a tuning factor; and group features below the threshold into a "minority" bin. By cross validation, we set $c = 100$ and the training data contains $1.2M$ query-ad pairs. The match type feature shall carry important information for predicting CTR, and intuitively can be leveraged in two ways: either as an additional feature to form the query-ad-matchtype triplet or as an additional dimension in the relative position. Empirically we found the second approach yielded better results. The training data contains 200 distinct position-dude-matchtype triplets, with 12 absolute positions $[1 \ldots 12]$, 5 dude states $[1 \ldots 5]$ (we encode zero ML ads as dude = 5 for convenience of illustration), and 4 match types $[1 \ldots 4]$. We used the same parameters for the gamma prior ($\alpha = 1.2, \beta = 0.01$). For the EM recurrence,

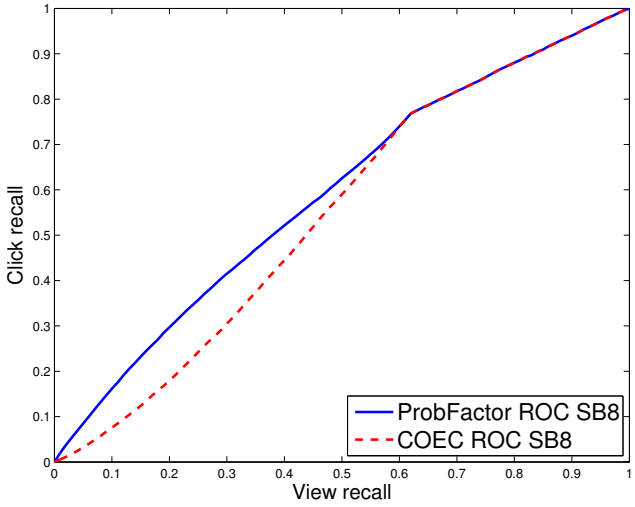
[3]We ignored bottom ads for simplicity and their negligible revenue impact.



(a) ROC of Factor and COEC for ML1

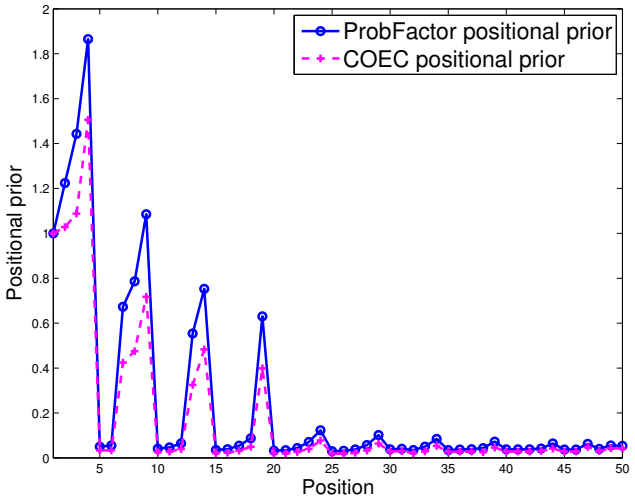


(b) ROC of Factor and COEC for ML4

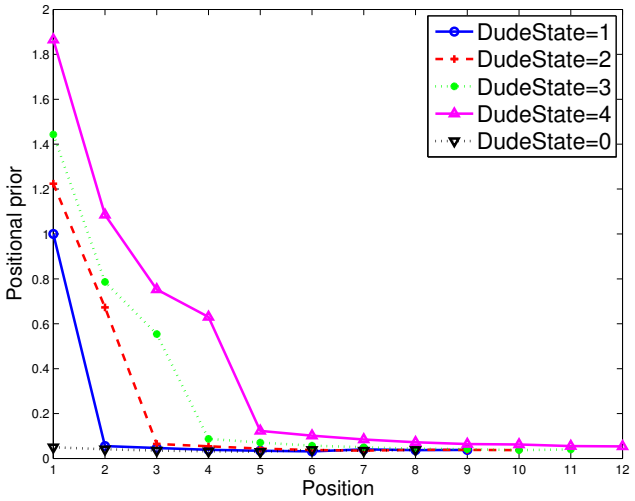


(c) ROC of Factor and COEC for SB8

Figure 5: Positional ROC curves Factor and COEC.



(a) Positional priors per (position-dude) pair by factor and COEC models



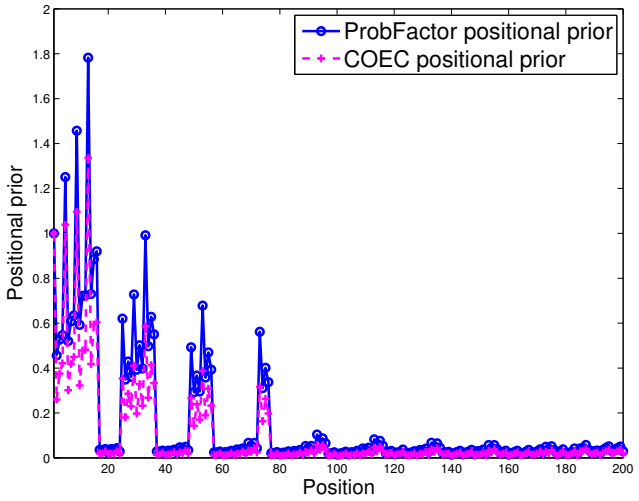
(b) Positional priors per absolute position by factor model for different dude states

Figure 6: Positional priors per (a) relative positions defined by (position-dude) for factor and COEC models, (b) absolute positions by factor model for different dude states.

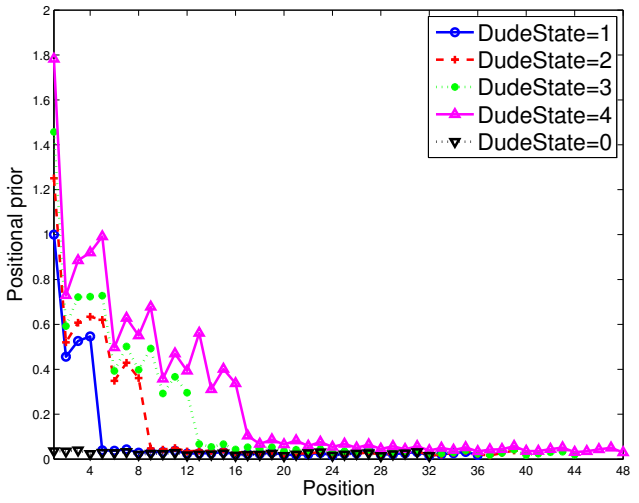
we initialize the model parameters $\theta = (p, q)$ to be empirical expectation. This data-driven initialization leverages data sparsity well since, under the multiplicative recurrence as in Eq. (22), initial zero p_i's (due to no clicks) will remain zero.

The positional effect is a function of relative position. More precisely, the positional prior q_j depends on at least three factors: (1) the page section, e.g., ML or SB; (2) the relative position or the number of ads above in section, e.g., ML2; and (3) the number of ads below in section, e.g., ML2 with dude = 4 has two ads below in ML. The learned positional priors as shown in Figures 6 and 7 support the above observations.

Figure 6(a) plots the positional priors q_j where a position j is defined by an (absolute position, dude) pair. The x-axis indexes (absolute position, dude) pairs in ascending order,



(a) Positional priors per (position-dude-matchtype) triplet by factor and COEC models



(b) Positional priors per (position-matchtype) pair by factor model for different dude states

Figure 7: Positional priors per (a) relative positions defined by (position-dude-match) for factor and COEC models, (b) relative positions defined by (position-match) by factor model for different dude states.

e.g., $x = 1$ for $(1, 1)$, $x = 2$ for $(1, 2)$, and so forth. The global descending trend (by examining local peaks $x = 4, 9, 14, \ldots$) reflects the positional bias. The discrepancy between the factor and COEC models confirms the "double-discounting" trap. The dude factor (the number of ads in ML) comes into play in two ways. Firstly and intuitively, when a dude value moves an ad to SB, the positional prior drops drastically (by examining the points at $x = 5, 6, 10, 11, 12, \ldots$). Secondly and more interestingly, when the number of ads below in section increases, the positional prior rises substantially. For example, the positional priors at $x = 1 \ldots 4$ are all for ML1 but with dude $= 1 \ldots 4$, and the increase is almost two folds from dude $= 1$ to 4. One plausible explanation is that more ML ads prevents a user from moving away from ads to algorithmic results. Another evidence for the effect of the number of ads below can be appreciated by comparing the positional priors at $x = 1$ (for ML1 with dude $= 1$) and $x = 9$ (for ML2 with dude $= 4$). In this case, ML2 has a higher positional prior than ML1. Figure 6(b) plots the positional priors q_j where a position j is defined by an absolute position, and one line per dude state. It is evident that, for each dude line, the positional prior drops most drastically at the dude value, that is, moving from ML to SB.

Figure 7(a), where a position is defined by an (absolute position, dude, match type) triplet, shows the effect of match type. For instance, the positional priors at $x = 1 \ldots 4$ are all for ML1 and dude $= 1$ but with matchtype $= 1 \ldots 4$. The difference between matchtype $= 1$ (exact match) and matchtype $= 2$ (broad match) is greater than two folds. Figure 7(b) separate one line per each dude state, by defining position as a (absolute position, match type) pair. The impact of match type given a same position can be appreciated by examining local variations for each position (the points at $x = 1 \ldots 4, 5 \ldots 8$, and so on.

We now evaluate the CTR prediction accuracy of the trained models. Besides the ranking metric AUC, we use the metric of relative information gain (RIG), defined as follows.

$$\text{RIG} = \frac{\sum_i (y_i \log p_i + (1 - y_i) \log (1 - p_i))}{nH(p)} + 1, \qquad (23)$$

where i indexes impressions in the testing data set, $y_i \in \{0, 1\}$ indicates click or not, p_i is the predicted CTR, n is the number of testing impressions, and $H(p)$ is the entropy of the empirical expected CTR, i.e., $\sum_i y_i/n$. RIG is a normalized average log likelihood under the binomial assumption, or a normalized negative cross entropy. RIG can be interpreted as the relative gain of a model $p_i, \forall i$ over a trivial yet foresighted one-parameter model by predicting every impression as the empirical expected CTR, that is, when $p_i = \sum_{i'} y_{i'}/n, \forall i$, RIG $= 0$.

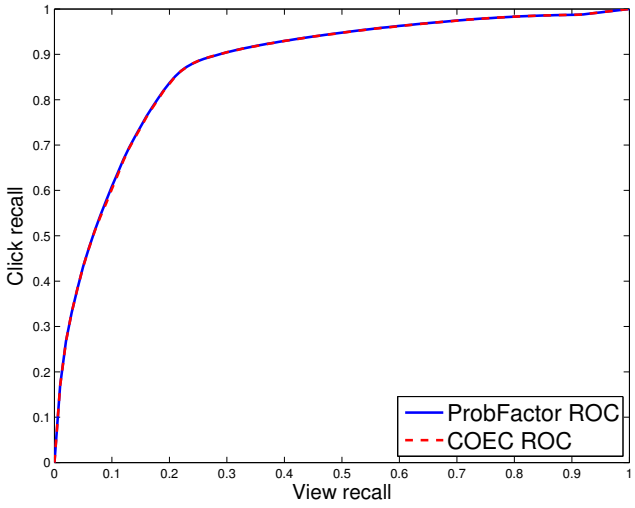
We compare models with different definitions of relative position, factor or COEC positional handling, and treating match type as exogenous or endogenous variables, as summarized in Table 3. The results are shown in Table 4.

Table 3: Different modeling methods

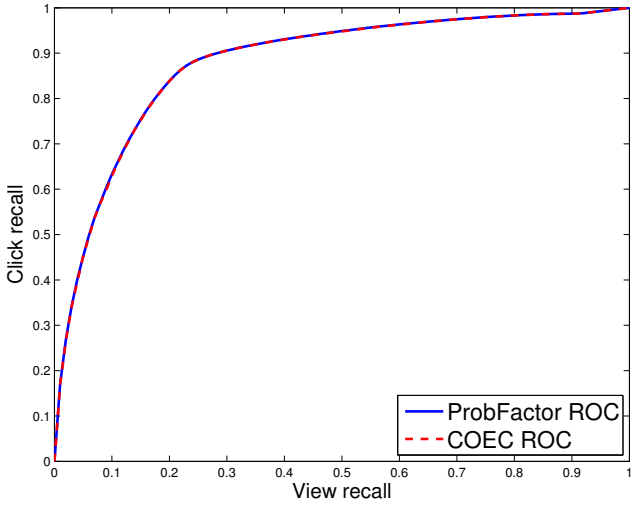
Model	Position def.	Learning	Match
Factor(pos)	abs. position	factor	none
COEC(pos)	abs. position	COEC	none
Factor(pos-dud)	(position,dude)	factor	none
COEC(pos-dud)	(position,dude)	COEC	none
Factor(pos-dud-mat)	(position,dude,match)	factor	exog.
COEC(pos-dud-mat)	(position,dude,match)	COEC	exog.
Factor(admat-pos-dud)	(position,dude)	factor	endog.
COEC(admat-pos-dud)	(position,dude)	COEC	endog.

Table 4: Experimental results

Model	RIG	AUC@all	@ML1	@ML4
Factor(pos)	0.1126	0.8225	0.7725	0.6942
COEC(pos)	0.1126	0.8216	0.7723	0.6935
Factor(pos-dud)	0.1797	0.8716	0.8363	0.8207
COEC(pos-dud)	0.1797	0.8712	0.8357	0.8200
Factor(pos-dud-mat)	**0.1861**	**0.8750**	0.8435	0.8257
COEC(pos-dud-mat)	0.1861	0.8744	0.8424	0.8252
Factor(admat-pos-dud)	0.1430	0.8713	0.8419	0.8190
COEC(admat-pos-dud)	0.1430	0.8706	0.8414	0.8183



(a) ROC of Factor and COEC with (position-dude) prior



(b) ROC of Factor and COEC with (position-dude-matchtype) prior

Figure 8: Experimental ROC curves for Factor and COEC with (a) (position-dude) positional prior, and (b) (position-dude-matchtype) positional prior.

The best empirical results were obtained by the factor model using the (position, dude, match) positional prior, with a RIG 0.1861 and an AUC 0.8750. When only using absolute position to derive positional prior, the prediction accuracy is much worse, with a RIG 0.1126 and an AUC 0.8225. The factor model consistently outperforms the COEC model, yet by a small margin. Adding match type into feature vector does not gain incremental ranking performance (from an AUC of 0.8716 by Factor(pos-dud) to an AUC of 0.8713 by Factor(admat-pos-dud)), and even impairs the testing data log likelihood metric (from a RIG of 0.1797 by Factor(pos-dud) to a RIG of 0.1430 by Factor(admat-pos-dud)), likely due to overfitting as a consequence of increasing the number of model parameters by four folds. The ROC curves of the Factor(pos-dude) and Factor(pos-dud-match) models are shown in Figure 8.

To evaluate the impact of the size of training data on the prediction performance, we varied the number of training days from 1, 7, to 13 preceding a same testing day using the empirically best model (Factor(pos-dud-match)), and the results are shown in Table 5.

Table 5: The impact of training data size

Model	RIG	AUC@all
Factor(pos-dud-match) 1-day train	−0.1058	0.8519
Factor(pos-dud-match) 7-day train	0.1861	0.8750
Factor(pos-dud-mat) 13-day train	**0.2219**	**0.8809**

It is evident that more training data gives better model generality, under the position-smoothed kNN model we are experimenting with. In particular, up to 13-day training, both RIG and AUC metrics monotonically increase. With only one day of training data, although the ranking metric is reasonable (AUC = 0.8519), the model likely suffers from overfitting (RIG = −0.1058). Another tuning parameter influencing the trade-off between model capacity and generality is the feature selection threshold factor c, and its choice should be made empirically or by cross validation. For this evaluation, we intentionally used the data from a smaller non-US market (UK), and the comparison results are shown in Table 6.

Table 6: The impact of feature selection

Model	RIG	AUC@all
Factor(pos-dud) $c = 20$	−0.0912	0.8508
Factor(pos-dud-match) $c = 20$	−0.0887	0.8521
Factor(pos-dud) $c = 100$	0.1693	0.8612
Factor(pos-dud-mat) $c = 100$	**0.1755**	**0.8648**

8. CONCLUSIONS

We have presented a probabilistic factor model to estimate CTR for search advertising, particularly to normalize out the positional bias. Our approach is simple yet principled, and empirically justified by the advertising domain. Moreover, we have generalized the concept of position to other important factors specific to ads, including keyword match type and ad presentation layout. Hence our approach shall serve as a general framework to study changes exogenous to the relevance-based CTR model. Finally, we conducted extensive experiments with synthetic and real-world ad data, and established that removing those confounding factors is critical to achieve accurate CTR prediction.

9. REFERENCES

[1] D. Agarwal, R. Agrawal, R. Khanna, and N. Kota. Estimating rates of rare events with multiple hierarchies through scalable log-linear models. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010)*, pages 213–222, 2010.

[2] S. Athey and D. Nekipelov. A structural model of sponsored search advertising auctions. *Proceedings of the 6th Ad Auctions Workshop*, 2010.

[3] O. Chapelle and Y. Zhang. A dynamic Bayesian network click model for web search ranking.

Proceedings of the 18th International Conference on World Wide Web (WWW 2009), pages 1–10, 2009.

[4] Y. Chen, M. Kapralov, D. Pavlov, and J. F. Canny. Factor modeling for advertisement targeting. *Advances in Neural Information Processing Systems (NIPS 2009)*, 22:324–332, 2009.

[5] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. *Proceedings of the First ACM International Conference on Web Search and Data Mining (WSDM 2008)*, pages 87–94, 2008.

[6] B. Edelman, M. Ostrovsky, and M. Schwarz. Internet advertising and the generalized second price auction: selling billions of dollars worth of keywords. *American Economic Review*, 97:242–259, 2005.

[7] T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich. Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine. *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, pages 13–20, 2010.

[8] D. Hillard, E. Manavoglu, H. Raghavan, C. Leggetter, E. Cantú-Paz, and R. Iyer. The sum of its parts: reducing sparsity in click estimation with query segments. *Information Retrieval*, 14:315–336, 2011.

[9] H. Ishwaran and J. Rao. Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics*, 33(2):730–773, 2005.

[10] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems (NIPS 2000)*, 13:556–562, 2000.

[11] F. Pin and P. Key. Stochastic variability in sponsored search auctions: observations and models. *Proceedings of the 12th ACM Conference on Electronic Commerce (EC 2011)*, pages 61–70, 2011.

[12] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. *Proceedings of the 16th International Conference on World Wide Web (WWW 2007)*, pages 521–530, 2007.

[13] W. V. Zhang and R. Jones. Comparing click logs and editorial labels for training query rewriting. *WWW 2007 Workshop on Query Log Analysis: Social And Technological Challenges*, 2007.