Abstract—Cold-start is a notoriously difficult problem which can occur in recommendation systems, and arises when there is insufficient information to draw inferences for users or items. To address this challenge, a contextual bandit algorithm – the Fast Approximate Bayesian Contextual Cold Start Learning algorithm (FAB-COST) – is proposed, which is designed to provide improved accuracy compared to the traditionally used Laplace approximation in the logistic contextual bandit, while controlling both algorithmic complexity and computational cost. To this end, FAB-COST uses a combination of two moment projection variational methods: Expectation Propagation (EP), which performs well at the cold start, but becomes slow as the amount of data increases; and Assumed Density Filtering (ADF), which has slower growth of computational cost with data size but requires more data to obtain an acceptable level of accuracy. By switching from EP to ADF when the dataset becomes large, it is able to exploit their complementary strengths. The empirical justification for FAB-COST is presented, and systematically compared to other approaches on simulated data. In a benchmark against the Laplace approximation on real data consisting of over 670,000 impressions from autotrader.co.uk, FAB-COST demonstrates at one point increase 16% recommendations for a user. On the basis of these results, it is argued that FAB-COST is likely to be an attractive approach to cold-start recommendation systems in a variety of contexts.

Keywords—cold-start, expectation propagation, multi-armed bandits, thompson sampling, variational inference.

I. INTRODUCTION

When making recommendations to website users, it is important to learn as efficiently as possible which content is most appropriate to display, including the ‘cold-start’ case when there is little or no prior history of the user and/or the content. Content recommendation systems and online advertising are examples contexts, in which there is an intrinsic trade-off between exploiting current knowledge by e.g. displaying adverts believed most likely to be clicked on, and exploring other content that might have higher rewards by e.g. displaying adverts which there is currently little information about.

Multi-armed bandits (MABs) are a class of algorithm which aim to balance the exploration-exploitation dilemma present whenever an intelligent system must make decisions in an uncertain environment [1]. For a recent and detailed overview on Multi-armed bandits see Lattimore & Szepesvari (2018) [2]. The most effective and conceptually simple MAB algorithm is known as Thompson Sampling [3], which uses sampling from a posterior distribution obtained through Bayesian inference. While Thompson’s original (non-contextual) model for such inference has the benefit of being analytically tractable, it has the drawback of each action being assumed independent; when working in a large action space, it will be much more efficient to share information among similar content, which is the main motivation behind the contextual bandit. Chapelle et al [4] propose such a contextual bandit based on Bayesian logistic regression, which is not in general analytically tractable, leading to the authors’ use of the Laplace approximation.

Making use of the Bernstein-von Mises and central limit theorems [5], the Laplace approximation will have errors that are asymptotically $O(T^{-1})$, where $T$ is number of impressions observed. For extremely large datasets, these errors will therefore become negligible, however they may be very large early on in the learning process, and as is shown, these errors may be large enough to have significant practical consequences even after many thousands of observations.

The accuracy of the inference procedure is improved upon by using a combination of Expectation Propagation (EP), which was contemporaneously devised by Minka (2001) [6] and Opper & Winther (2000) [7], and Assumed density filtering (ADF) as presented by Opper (1998) [8]. These are both moment projection variational inference methods, meaning that they work by iteratively projecting the intractable posterior distribution onto a tractable one (usually belonging to the exponential family) via minimisation of the forward Kullback-Leibler divergence. ADF is a one-pass, online method, and observations are processed one-by-one, updating the posterior distribution which is then approximated before processing the next observation. EP – which is an extension to ADF – iteratively refines the approximation by making additional passes through the dataset giving much better accuracy, but at the same time incurring a greater computational cost.

Another class of methods considered are Markov chain Monte Carlo (MCMC), which are capable of generating arbitrarily accurate representations of the Bayesian posterior. These are useful to provide a ‘ground truth’ for comparison of approximate methods in a study such as this, however they are not suitable for online use. This is because MCMC requires large amounts of computational effort, and also typically involves algorithmic parameters that currently cannot be tuned automatically, but rather need to be adjusted until convergence can be diagnosed [9].

The paper is structured as follows. Section II provides details of the multi-armed bandit approach to recommender systems, and Section III provides details of the inferential systems used in Bayesian online learning. In Section IV the FAB-COST algorithm is introduced, which is systemati-
cally compared to Laplace, EP, ADF and MCMC inference procedures, showing an attractive balance of accuracy and computational effort. It is then demonstrated that the increased accuracy of the posterior leads to better performance when used in the logistic bandit setting and results in more clicks generated when used in an online advertising scenario on real data from autotrader.co.uk. Finally the details are concluded on in Section V.

The code used to generate the results in this paper are available on GitHub at https://github.com/JackMack21/FAB-COST.

II. BANDIT ALGORITHMS

A. Notation and general setup

Here and throughout, the shorthand $\mathcal{N}(\mu, \Sigma)$ represents the Gaussian distribution, where $\mu \in \mathbb{R}^D$ is the mean vector of first moments and $\Sigma \in \mathbb{R}^{D \times D}$ is the covariance matrix of centred second moments. $\mathbb{E}_{p(x)}[\cdot]$ represents an expectation over the probability distribution $p(x)$.

Steps in the bandit algorithm are called iterations, and these are indexed by integers $i = 1, \ldots, T$. $T$ is the total iterations over the learning process, and $\tau \leq T$ refers to a current, but not necessarily final, iteration. At each iteration, the bandit algorithm - also referred to as the learner, and which is assumed to serve adverts for exositional simplicity - selects an advert from the set of eligible adverts $\mathcal{A}_i$ with cardinality $|\mathcal{A}_i| = K$. This set is indexed by $j = 1, \ldots, K$. Each advert has $D$ features, examples of which in the context of automobile sales being the colour, model and age of the car advertised. As such, the options available can be represented as a matrix $\mathbf{A}^T \in \mathbb{R}^{K \times D}$, with the $j$-th row, $\mathbf{a}_j^T$, having elements corresponding to the features of an eligible advert.

The observed reward of the advert selected by the learner at iteration $i$ is the binary outcome of a non-click/click, $y_i \in \{0, 1\}$. Each $y$ is treated as a random variable, the expected value for which at the $i$-th iteration corresponds to the selected advert’s click-through-rate (CTR).

$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]^T$ is called the design matrix for the entire learning process. Each row $\mathbf{x}_i^T \in \mathbb{R}^D$ represents the features of the advert displayed at the $i$-th iteration. For some algorithms the matrix $\mathbf{X}^T = [\mathbf{x}_1, \ldots, \mathbf{x}_T]^T$ will be used; this contains the information about the history of adverts chosen up to the current iteration $\tau$. The history of observations up to iteration $\tau$ is represented by the vector $\mathbf{y}_\tau \in \mathbb{R}^\tau$.

$\Theta \in \Theta$ is the vector of parameters that is to be learnt. In the non-contextual case, each advert has a local parameter $\theta_j$ that is learnt and so $\Theta = \mathbb{R}^K$. In the contextual case $\Theta = \mathbb{R}^D$ and $\theta$ is a global parameter vector which shares information between adverts.

B. Thompson Sampling

Thompson sampling [3] is a Bayesian approach to MABs. Its use requires the quantification of belief about the CTR for each advert via a posterior probability distribution. The adverts shown and the binary reward of click/no-click observed up to the current time $\tau$ are denoted via the vectors $\mathbf{a}_\tau = [a_1, \ldots, a_\tau]^T$ and $\mathbf{y}_\tau = [y_1, \ldots, y_\tau]^T$ respectively. As data arrives, which is comprised of the tuple $(a_i, y_i)$, the latent parameter $\theta$ can be learnt via Bayes’ rule

$$p(\theta|\mathbf{y}_\tau, \mathbf{a}_\tau) \propto \prod_{i=1}^\tau p(y_i|a_i, \theta)p(\theta).$$

Once a posterior distribution has been calculated for each advert, an advert should be chosen which maximises the expected CTR, where the expected CTR is calculated as

$$\mathbb{E}[y_{\tau+1}|a_{\tau+1}, \mathbf{a}_\tau, \mathbf{y}_\tau] = \int \mathbb{E}[y_{\tau+1}|\mathbf{y}_{\tau+1}, \mathbf{a}_\tau, \theta]p(\theta|\mathbf{y}_{\tau+1}, \mathbf{a}_\tau, \theta)d\theta.$$  

In Thompson sampling it is not necessary to evaluate this integral, but rather a sample is generated from each advert’s CTR posterior, and the advert which corresponds to the maximum sample generated is displayed. Although the integral [1] could be evaluated explicitly, displaying content that simply maximises it would involve constantly exploiting existing knowledge and would never explore to learn. The sampling-based approach, however, allows exploration and exploitation to be traded off.

A well studied example of Thompson Sampling is the non-contextual Beta-Bernoulli bandit. In this simple case, each advert is assumed to have an independent Bernoulli likelihood. This has a conjugate prior, the Beta distribution, meaning that Bayesian inference can be performed analytically to give a closed form solution. When working in a very large action space (i.e. with many adverts) it is, however, much more efficient to share information among similar adverts rather than assuming independence, which is the main motivation behind the contextual bandit.

C. The Contextual Bandit

Instead of learning an independent posterior distribution for each advert, the contextual bandit instead learns a global posterior $p(\theta|\mathbf{X})$ where $\theta \in \mathbb{R}^D$, with $D$ representing the number of covariates.

As discussed above, at the current iteration $\tau$, the learner is presented with the features of the adverts available, stored in the matrix $\mathbf{A}^T$. The learner then chooses an advert from $\mathcal{A}_\tau$ which is expected to have the highest CTR when combining the context received with the sample generated from the posterior

$$a_\tau = \text{argmax}(\mathbf{A}^T \theta),$$

i.e. if the $j$-th row of $\mathbf{A}^T \theta$ as in (2) is the maximum, advert $a_j$ is shown. After observing the reward $y_i$, the context of the advert chosen $a_j$ is then added to the history of chosen adverts (corresponding to the $i$-th row of $\mathbf{X}$), and the posterior is updated either in batch using $\mathbf{X}^T$ or online using the chosen adverts covariates which is now denoted $\mathbf{x}_\tau$.

The global parameter vector $\theta$ therefore acts as a projection of the features onto the real numbers. Since the outcome is a binary reward $y_i \in \{0, 1\}$, an approach is needed that relates a continuous projection to such an outcome. Bayesian logistic regression fulfils this requirement, which is the next point of discussion.
D. Bayesian Logistic regression

Logistic regression has proven to be a very popular and effective two class ‘soft’ classification method \[^{10}, {11}\]. The goal of logistic regression is to find the best fitting model to describe the relationship thus far observed between the binary response variables \(y \in \mathbb{R}^T\), and the design matrix \(X \in \mathbb{R}^{T \times D}\). Although in statistics literature it is referred to as logistic regression, by virtue of modelling a binary outcome it is also a method for classification \[^{12}\]. While there has recently been much interest in learning techniques based on e.g. neural networks, as Dacrema et al. \[^{13}\] point out, interpretability and reproducibility are two very important issues that need to be addressed, and are important advantage of using well-understood statistical techniques.

Logistic regression is a likelihood-based method in which the parameter vector \(\theta\) describes the probabilistic relationship between the input vector \(x_i\), and a binary response \(y_i \in \{0, 1\}\). Due to the response being binary, a Bernoulli model is used for probabilities:

\[
\Pr(y_i|x_i, \theta) = \pi(x_i, \theta)^{y_i}(1 - \pi(x_i, \theta))^{(1-y_i)},
\]

where \(\pi(x_i, \theta) = \mathbb{E}[y_i|x_i, \theta] = p(y_i = 1|x_i, \theta)\). As \(\pi(x_i, \theta)\) is the probability of observing a positive outcome, the inner product \(\theta^\top x_i\) is mapped from the real line to the interval \([0, 1]\) via the (sigmoidal) logistic function, \(\sigma(x) = 1/(1 + \exp(-x))\),

\[
\pi(x_i, \theta) = \sigma(\theta^\top x_i) = \frac{1}{1 + \exp(-\theta^\top x_i)}, \quad \text{(4)}
\]

Assuming conditional independence at each iteration, the likelihood function for the current iteration is obtained as

\[
p(y_{\tau}|X^{\tau}, \theta) = \prod_{i=1}^{\tau} \Pr(y_i|x_i, \theta), \quad \text{(5)}
\]

where the historical observations are represented as \(y_{\tau} = [y_1, \ldots, y_{\tau}]^\top\) and \(X^{\tau} = [x_1, \ldots, x_{\tau}]^\top\). Equations (5), (4) and (6) between them define logistic regression.

To select adverts via Thompson sampling, \(\theta\) must be estimated, including an appropriate quantification of uncertainty, which motivates the use of a Bayesian treatment of logistic regression. Here, it is assumed that there is a distribution over \(\theta\) that is sequential learnt as data arrives via Bayes’ rule:

\[
p(\theta|y_{\tau}, X^{\tau}) \propto p(y_{\tau}|X^{\tau}, \theta)p(\theta), \quad \text{(6)}
\]

where \(p(\theta)\) is a probability density function representing the prior belief about the parameters, \(p(\theta|y_{\tau}, X^{\tau})\) is a probability density function representing the posterior beliefs about the parameters, and all other quantities are as defined above. Since \(\theta\) is passed through a non-linear mapping, inference is not straightforward; more explicitly, the logistic likelihood function does not permit a conjugate prior. This leads onto the next section which explains methods for dealing with such situations.

III. INFERENTIAL METHODOLOGY

Suppose that one is trying to solve the equation (6) for the posterior distribution – dependence on \(y_{\tau}, X^{\tau}\) is suppressed the exact distribution is denoted \(p(\theta)\). When there is a conjugate prior, the constant of proportionality in (6) can be calculated exactly, but otherwise it must be approximated.

Markov chain Monte Carlo (MCMC) methods are very popular in Bayesian statistics; and provided there is sufficient computational resources, they guarantee an arbitrarily accurate approximation to the posterior distribution. The need for large computational resources can, however, be problematic, especially when working at scale. An overview of MCMC is provided by Brooks et al. \[^{9}\].

Other approaches that will be used and detail below make a Gaussian approximation, which is justified via the Bernstein Von Mises Theorem. This states that the posterior converges to a Gaussian asymptotically \[^{14}\], however different methods will achieve different accuracies at a given amount of data.

A. The Laplace approximation

The Laplace approximation is a very popular inference method in Bayesian statistics. Its popularity is due to its simplicity: given a target distribution \(p(\theta) = \exp(-\xi(\theta))\), which in this case will be the posterior distribution, a tractable Gaussian approximation \(q(\theta)\) is made, centred at the mode of the original target density with variance equal to the curvature of the negative log-target. Explicitly, let

\[
\mu^* = \text{argmin}_{\theta} \xi(\theta), \quad \Lambda^* = \left. \frac{\partial^2 \xi(\theta)}{\partial \theta^2} \right|_{\theta = \mu^*},
\]

and then the two moments of the Gaussian approximation are matched to the above:

\[
p(\theta) \approx q(\theta) = N(\mu^*, \Lambda^{*^{-1}}).
\]

The Laplace approximation is used as an inference procedure in the logistic contextual bandit described by Chapelle et al \[^{4}\] and very commonly when making an approximation in Bayesian GLMs. Asymptotically, the errors are expected to be \(O(T^{-1})\) \[^{15}\].

B. Variational Inference

Variational approximations turn what was an inference problem into one of optimisation. They work by minimising a distance measure \(D\) between the target distribution \(p(\theta)\) and an approximation \(q(\theta) \in Z\), where \(Z\) is the family of densities the approximation is to be restricted to, so that

\[
q^*(\theta) = \text{argmin}_{q \in Z} D(q(\theta), p(\theta|x)). \quad \text{(7)}
\]

The more general there is in \(Z\), the more complex the optimisation procedure becomes, and in this application, the approximating family is restricted to the set of Gaussians.

The distance measure used is typically the Kullback-Leibler divergence \[^{16}\], also known as the relative entropy, which is non-symmetric. Here, the forward KL-divergence, defined as

\[
\text{KL}(p(\theta|x)||q(\theta)) = \int p(\theta|x) \log \left( \frac{p(\theta|x)}{q(\theta)} \right) d\theta \quad \text{(8)}
\]

where

\[
\text{KL}(p(\theta|x)||q(\theta)) = \mathbb{E}_{p(\theta|x)} [\log \frac{p(\theta|x)}{q(\theta)}]
\]

is used as the distance measure, which when used in the objective in (7), gives a solution known as the moment projection.
Note that the reverse KL-divergence is obtained by swapping $p$ and $q$ in (5) and when used as the objective in (7), the solution is known as the information projection. This form is not necessarily convex in $\theta$ and can therefore yield different solutions depending on how the optimisation procedure is initialised. While the information projection is most commonly used in variational inference due to its simplicity, there are problems with reliability and accuracy, and no advantage was found by using this approach – see Bishop (2006) [12] and Blei et al. (2017) [17] for further discussion.

Unlike the information projection, the objective (5) is convex in $\theta$, and will give a unique solution when minimised. This minimum corresponds to an approximation centred at the mean of the target which is why it is known as mean seeking. The moment projection plays an important role in asymptotic theory and as explained by [18], its minimisation has the desirable effect of minimising the expected loss.

C. Exponential families

A distribution belongs to the set of exponential families if its density can be written as

$$q(\theta|\lambda) = \exp \left( \lambda^T \phi(\theta) - \Phi(\lambda) \right),$$

where $\Phi(\lambda)$ is known as the log-partition function which ensures normalisation.

Assuming that the representation of the exponential family is minimal (i.e. there are no dependencies between the components of $\phi(\theta)$ and $\lambda$), then the following properties hold:

1) **Product of Exponentials:** A product of densities belonging to the set exponential families is also a member of the exponential families:

$$\prod_{i=1}^{T} q_i(\theta|\lambda_i) = q(\theta | \sum_{i=1}^{T} \lambda_i) \exp \left( \phi(\theta) \left( \sum_{i=1}^{T} \lambda_i \right) - \sum_{i=1}^{T} \Phi(\lambda_i) \right)$$

given that $\sum_{i=1}^{T} \lambda_i \in \Theta$.

2) **Moments:** Moments can be found by differentiating the log-partition function with respect to its natural parameters:

$$E_{q(\theta)}[\phi(\theta)] = \nabla_{\lambda} \Phi(\lambda), \quad \text{Var}_{q(\theta)}[\phi(\theta)] = \nabla_{\lambda}^2 \Phi(\lambda).$$

3) **Bijective mapping:** There is a bijective mapping from the natural parameters $\lambda(\eta)$ and the moment parameters $\eta(\lambda)$. The log-partition function $\Phi(\lambda)$ is strictly convex and has a Legendre dual

$$\Psi(\eta) = E_{q(\lambda)}[\log p(\lambda|\eta)|\theta]$$

Conversion between the natural and moments parameters is done via:

$$\eta(\lambda) = \nabla_{\lambda} \Phi(\lambda), \quad \theta(\eta) = \nabla_{\eta} \Psi(\eta)$$

These properties are very useful for message passing algorithms which both EP and ADF are a subset of – see Pearl (1986) [19], Seeger (2007) [20] gives an in-depth discussion of the properties of exponential families.

IV. FAB-COST

In this section a description of the FAB-COST algorithm is given, along with a demonstration of its performance on real automotive website data (AT) taken from the autotrader.co.uk website.

A. The Logistic Contextual Bandit

Having introduced multi-armed bandits, Thompson sampling for these via Bayesian logistic regression, and multivariate normal approximations to the posterior in such regressions, the overall structure of the FAB-COST approach, as in Algorithm [1] can now be provided. It is worth mentioning that the pseudocode for the logistic regression bandit provided is the same as the linear bandit introduced by Russo et al (2014) [21] and Agrawal et al (2013) [22], however the linear case is conjugate and can be solved exactly, meaning that it does not require the work to update moments accurately that have been carried out.

Algorithm 1: Logistic Regression Thompson Sampling

1. for $i = 1 \ldots T$ do
2. 1. Generate a sample from the approximated posterior:
3. $\theta_i \sim N(\mu_{i-1}, \Sigma_{i-1})$
4. 2. Select an advert:
5. $a_i = \arg\max_{j \in A} (A^T \theta_i)$
6. 3. Update moments:
7. $\mu_i \leftarrow \mathbb{E}[\theta_i | X_i, y_i]$
8. $\Sigma_i \leftarrow \mathbb{E}(\theta - \mu_i)(\theta - \mu_i)^T | X_i, y_i$
9. end

Describing this algorithm in long form, it is initialised with a prior belief on $\theta$. At each iteration, the learner is presented with an available set of adverts and random sample is generated from the posterior distribution (or prior at the first iteration), which corresponds to step 1 in Algorithm [1]. An advert is then selected from the set via Thompson Sampling by choosing $a_j$ which maximises the linear combination of the sample generated $\hat{\theta}_i$ and the eligible adverts covariates $A^j_i$; this corresponds to step 2 in Algorithm [1] – note that the monotonicity of the logistic function means that the sampled CTR is maximised when the linear combination is maximised. This is followed by observing a binary reward of a user clicking or not clicking on the chosen advert, at which point the posterior $p(\theta)$ is updated (i.e. within the Gaussian approximation, the first and second moments are updated). This corresponds to step 3 in Algorithm [1].
B. Expectation Propagation

Expectation Propagation (EP) is an iterative approach to minimising the forward KL-divergence between the posterior that is to be approximated, and the Gaussian approximation. It was first generalised by Minka (2001) \cite{Minka2001} and Opper & Winther (2000) \cite{Opper2000}, but has roots further back in Statistical Physics \cite{Dehaine2000}. EP belongs to a group of message passing algorithms and works by essentially propagating the moments of an exponential family - which in the Gaussian case are the mean and variance - between the factors of the posterior. Finding the moments of the target is obviously problematic as the target is intractable - if it weren’t then a closed form solution could be found analytically. EP’s solution to this is to form what is known as the tilted distribution $t_i(\theta)$ (whose moments are much easier to find) and iteratively project the moments from this, onto the tractable approximation $q(\theta)$.

First, it is assumed that the true posterior factorises into a product of $T$ factor terms or sites:

$$p(\theta|\lambda) \propto \prod_{i=1}^{T} p_i(\theta).$$

EP approximates each of these true sites by a Gaussian distribution $q_i(\theta)$, which in natural parameters is expressed as

$$p_i(\theta) \approx q_i(\theta|\lambda_i) \propto \exp \left\{ h_i x_i - \Lambda_i \frac{x_i^2}{2} \right\}.$$  

Then due to property 1 in of exponential families in \S III-C the global approximation is expressed as

$$q(\theta|\lambda) = \prod_{i=1}^{T} q_i(\theta|\lambda_i),$$

and the natural parameters of our global approximation can be calculated as the product of the natural parameters of each site approximation $h = \sum_{i=1}^{T} h_i$, $\Lambda = \sum_{i=1}^{T} \Lambda_i$. It is this ability to simply add and subtract natural parameters of the sites that motivates the use of an exponential family approximation.

The EP algorithm sweeps through the data set, with steps described below.

1) The tilted distribution: At each iteration of the algorithm, the current global approximation $q(\theta|\lambda) = \prod_{i=1}^{T} q_i(\theta|\lambda_i)$ is augmented by replacing one of the sites with a true site $p_i(\theta)$. This can be thought of in two steps: firstly the cavity distribution is defined as

$$q^{\backslash i}(\theta|\lambda^{\backslash i}) = \prod_{j \neq i} q_j(\theta|\lambda_j),$$

which is the global approximation with a site removed, then the tilted distribution is defined as

$$t_i(\theta) \propto p_i(\theta) \prod_{j \neq i} q_j(\theta|\lambda_j),$$

which is the cavity with its ‘hole’ filled in with a true site.

2) The moment projection: EP proceeds to iteratively project the tilted onto the global approximation

$$q^*(\theta) = \arg\min_{q \in Z} \text{KL}(t_i(\theta)||q(\theta|\lambda)).$$  

Equivalently, the first two moments of the tilted can be computed $E_{t_i}[\phi(\theta)] = [\mu, \Sigma]^\top$ and the moments of the global approximation are equated to these: $q^*(\theta) \sim \mathcal{N}(\mu, \Sigma)$.

The moments of the local site approximation $q_i(\theta)$ must then be updated which is generally done by division

$$q_i^*(\theta|\lambda_i) = \frac{q_i^*(\theta|\lambda)}{q^{\backslash i}(\theta|\lambda^{\backslash i})}.$$  

Using property 1 in of exponential families in \S III-C this results in a simple subtraction of the natural parameters.

3) Mapping from natural to moment parameters: As the objective in (9) is found via moment matching, yet the Gaussian approximation is parameterised by its natural parameters, the two different parameterisations but be alternated between at each iteration. Due to property 3 of exponential families in \S III-C there is a bijection mapping between the two. In the Gaussian case these mappings are:

$$\Sigma = \Lambda^{-1}, \quad \mu = \Lambda^{-1} h,$$

where the moment parameters – the mean and the variance – $\mu$ and $\Sigma$ respectively, and the natural parameters – the shift and the precision – are given by $h$ and $\Lambda$ respectively.

4) Overall structure: As shown in Algorithm 2 EP makes multiple sweeps through the dataset, iteratively forming the tilted distribution, matching the moments of the global approximation, and finally updating the site parameters. This is done until convergence which has been shown to usually be after 3-4 sweeps (which coincides with the experiments carried out in this work) although a stopping rule described by Seeger (2008) \cite{Seeger2008} can also be used.

Algorithm 2: The Expectation Propagation algorithm

\begin{algorithm}
\begin{algorithmic}
\State \textbf{while not converged} do
\For {$i = 1 \ldots T$} do
\State 1. Form the tilted distribution:
\State \hspace{.5cm} $t_i(\theta) = \frac{1}{Z} p_i(\theta) q^{\backslash i}(\theta|\lambda)$. 
\State 2. Minimise the forward KL-divergence between the tilted distribution and the global approximation:
\State \hspace{.5cm} $q^*(\theta|\lambda) = \arg\min_{q \in Z} \text{KL}(t_i(\theta)||q(\theta|\lambda))$. 
\State 3. Update the approximating site:
\State \hspace{1cm} $q_i^*(\theta|\lambda_i) = \frac{q_i^*(\theta|\lambda)}{q^{\backslash i}(\theta|\lambda^{\backslash i})}$. 
\EndFor
\EndWhile
\end{algorithmic}
\end{algorithm}

5) Error analysis: While EP has shown tremendous empirical success in many applications – particularly Bayesian General Linear Models and Gaussian Process regression \cite{Seeger2008} – the theoretical understanding of it is less comprehensive than for other approaches. There has, however, been important progress due to Dehaene and Barthelmé who show that EP
behaves like iterations of Newton’s algorithm for finding the mode of a function. Under what they describe in later work as ‘unrealistic assumptions on the model’, Dehaene and Barthelemy also showed that EP can converge at a rate of $O(T^{-2})$. The experiments carried out in this work suggest that while such a rate is indeed likely to be optimistic in more realistic settings, EP is often significantly more accurate than the Laplace approximation.

C. Assumed density filtering

1) Algorithm: EP is a batch method, requiring multiple sweeps through a complete dataset, potentially limiting its usefulness online. Assumed Density filtering (ADF) is a sequential inference method and can be used to work online. As with EP, ADF iteratively minimises the forward KL-divergence between the tilted distribution and the approximation, the difference being how this tilted distribution is formed.

In the setting of data arriving sequentially, the posterior distribution at data point $\tau$ is given as

$$p(\theta|x) = \frac{\prod_{i=1}^{\tau} p(x_i|\theta) p(\theta)}{\int \prod_{i=1}^{\tau} p(x_i|\theta) p(\theta) \, d\theta}.$$  

ADF takes $q(\theta)$ as the prior on $\theta$ and iterates through the data, incorporating each point into the approximate posterior. The conditional distribution of $\theta$ given the first $\tau$ data points can be expressed as

$$p(\theta|x_{1:\tau}) = \frac{p(x_{\tau}|\theta) p(\theta|x_{1:\tau-1})}{\int p(x_{\tau}|\theta) p(\theta|x_{1:\tau-1}) \, d\theta}.$$  

Then assuming that at the previous iteration the approximation $q^{(\tau-1)}(\theta)$ is made to the true posterior $p(\theta|x_{1:\tau-1})$, (11) can be rewritten to give a new tilted distribution

$$t_\tau(\theta|x_{\tau}) = \frac{p(x_{\tau}|\theta) q^{(\tau-1)}(\theta)}{\int p(x_{\tau}|\theta) q^{(\tau-1)}(\theta) \, d\theta}.$$  

Due to there being neither a site update nor a cavity there is no need to map between the natural and moment parameters and therefore no need to perform the expensive matrix inversion seen in equation (10).

2) Error analysis: Figure 1 shows that ADF makes a very poor approximation to the posterior unless it is trained on sufficient data. EP on the other hand makes an accurate approximation even on the first 1,000 data points.

Theoretical results are given by Opper (1998) calculates the asymptotic convergence of ADF by showing that the inverse of the covariance matrix approaches the fisher information matrix as $T \to \infty$. By assuming that the difference between the change in the covariance matrix between time points is negligible, he models its evolution as a matrix differential equation to give asymptotic accuracy of $O(T^{-1})$ for the mean, although no convergence rate is given for the variance.

These empirical and theoretical considerations align with the intuition that multiple sweeps through a dataset as in EP are expected to mitigate against inaccuracies from sites that lead to tilted distributions that are poorly approximated better than in a one-sweep algorithm such as ADF. Discussion by Gelman et al. is relevant in this context.

D. Gaussian filtering

1) Moment matching: As mentioned earlier, Equation (9), which takes the form:

$$q^*(\theta|\lambda) = \arg\min_{q \in \mathcal{Q}} KL(t_\tau(\theta)||q(\theta|\lambda)),$$

can be solved at each iteration via moment matching

$$E_q[\phi(\theta)] = E_{t_\tau}[\phi(\theta)].$$

Due to the approximation being restricted to the set of Gaussian distributions, this requires only the first two moments of the tilted distribution to be found: the mean $\mu$ and the variance $\Sigma$.

Note that if all moments associated with $q(\theta)$ existed and were equal to those associated with $t(\theta)$, then the KL-divergence would be zero and the approximation would be exact. As the approximation used is Gaussian, it is only characterised by the first two moments; any difference in higher moments (e.g. skew, kurtosis) between the two is will lead to errors since the Gaussian lacks the flexibility to capture these.

In the logistic regression case, the tilted function is

$$t_\tau(\theta) = \frac{1}{Z(\lambda_\tau)} p_\tau(\theta) \prod_{j \neq i} q_j(\theta|\lambda_j),$$

where the true site distribution functions are $p_i(\theta) = \sigma(y_i \theta^\top x_i)$ and the approximating site functions are $q_j(\theta|\lambda_j) = \mathcal{N}(\theta; \mu_j, \Sigma_j)$. The normalisation constant is

$$Z(\lambda_i) = \int p_i(\theta) \prod_{j \neq i} q_j(\theta|\lambda_j) \, d\theta.$$  

Because of product of exponentials property (see §III-C1), the cavity can be written as $\prod_{j \neq i} q_j(\theta) = \mathcal{N}(\theta; \mu^\lambda, \Sigma^\lambda)$ where the complement notation $\mu^\lambda := (\mu_j)_{j \neq i}$ has been used.

Using the moments property from §III-C2 the moments of the tilted can be found via differentiation of the log partition

![Fig. 1: Convergence of both EP and ADF on AT dataset. ADF is slow to converge, motivating its combination it with EP, which is typically much more accurate.](image-url)
function as

$$\mathbb{E}_t[\phi(\theta)] = \nabla_{\hat{\lambda}_i} \log \hat{Z}(\hat{\lambda}_i).$$

The remaining problem is that the natural parameters \(\hat{\lambda}_i\) are not known; if they were the moments could be found by a simple bijective mapping. Fortunately there is a recursive formulation derived by Herbrich (2005) \[28\], which enables the moments of the tilted to be calculated as

$$\mathbb{E}_t[\phi(\theta)] = \mathbb{E}_q[\phi(\theta)] + \nabla_{\hat{\lambda}_i} \log \hat{Z}(\hat{\lambda}_i), \quad (12)$$

meaning that the cavity natural parameters \(\lambda^i\) are needed instead of the tilted natural parameters \(\hat{\lambda}_i\).

After some calculations, [12] can be used to give an iterative update formula for the first and second Gaussian moments of the global approximation:

$$\mu = \mu^i + \Sigma^i \alpha_i, \quad \Sigma = \Sigma^i - \Sigma^i (\alpha_i \alpha_i^T - 2B_i) \Sigma^i, \quad (13)$$

where

$$\alpha_i = \nabla_{\mu^i} \log \hat{Z}(\hat{\lambda}_i) \in \mathbb{R}^D \quad \text{and} \quad B_i = \nabla_{\Sigma} \log \hat{Z}(\hat{\lambda}_i) \in \mathbb{R}^{D \times D} \quad (14)$$

2) Linear subspace property: The normalising constant to the tilted distribution \(\hat{Z}(\hat{\lambda}_i)\) is, in general, intractable – its moments can be evaluated via numerical quadrature or MCMC, however in this work an approximation described by MacKay (1992) \[29\] is used. In MacKay’s estimation framework, known as the ‘evidence framework’ or ‘moderated output’, the normalising constant is approximated as

$$\hat{Z}(\hat{\lambda}_i) = \int \sigma(y_i \theta^T x_i) \mathcal{N}(\theta; \mu^i, \Sigma^i) d\theta \approx \sigma(\kappa(s_i^2) a_i), \quad (15)$$

where

$$\kappa(s_i^2) = \left(1 + (\pi s_i^2/8)\right)^{-1/2}, \quad s_i^2 = x_i^T \Sigma^i x_i, \quad a_i = x_i^T \mu^i.$$  

Using the moderated output given in [15], \(\alpha_i\) and \(B_i\) are expressed as

$$\alpha_i = \nabla_{\mu^i} \log \hat{Z}(\hat{\lambda}_i) = \frac{\rho_i}{\sigma'(\rho_i \mu^i)} \sigma'(\rho_i \mu^i), \quad (16)$$

$$B_i = \nabla_{\Sigma} \log \hat{Z}(\hat{\lambda}_i) = -\frac{\pi}{16\kappa^2} x_i x_i^T \mu^i \alpha_i^T,$$

where

$$\rho_i = y_i \kappa(s_i^2) x_i^T, \quad \sigma(z) = \frac{1}{1 + \exp(-z)}, \quad \sigma'(z) = \frac{\exp(-z)}{(1 + \exp(-z)}.$$  

3) Filtering algorithm: Using the results [16] along with the iterative update equations [13], Gaussian approximations to the posterior in Bayesian logistic regression can be computed for both EP and ADF. These are shown in Algorithms 3 and 4 respectively.

E. Combining methods in FAB-COST

An outline is now given of how the considerations above lead us to the FAB-COST approach to recommendation systems, which will turn out to provide improved performance on real data for well-controlled computational effort.

1) Data: Auto Trader PLC is the UK’s largest digital automotive marketplace, and the ‘AT’ dataset was constructed from the website autotrader.co.uk. This consists of a day’s

![Fig. 2: FAB-COST vs Laplace. The Laplace approximation fails to capture the true variance, meaning that in the bandit setting a failure to balance the exploration-exploitation tradeoff is expected.](image-url)
‘featured listing’ user click data, totalling at $T = 678,446$
 impressions. After a user makes a search for a car they
 are presented with a single ‘featured listing’ which appears at
 the top of the search result, meaning that there is no positional
 bias. The user then proceeds to either click or not click on
 the presented advert. Many dozens of covariates for each advert
 are available, which was reduced to the 15 most important
 using a random forest algorithm for feature selection, although
 other possibilities for feature selection can be used [11].

2) Computational Cost: The structure of Algorithms 3 and 4
 makes clear that one iteration ADF is expected to
 be computationally cheaper than one iteration of EP. This
 is because, at each iteration, EP requires the expensive matrix
 inversion required to map between the natural and moment
 parameters. Although both matrix multiplication and matrix
 inversion come at cubic computational complexity $O(D^3)$, the
 pre-multiplication constant for ADF is much smaller. Calcula-
 tions show that per site, EP’s flop count of $\frac{3}{2}D^3 + O(D^2)$
 is over three times greater than that of ADF’s flop count
 of $4D^3 + O(D^2)$. Due to matrix multiplication being easily
 parallelised (which Python’s Numpy library exploits), as well
 as EP requiring multiple sweeps over the data set, we recorded
 a 75-fold speed-up when using ADF in batch over EP.

Now consider the number of iterations involved in an online
 learning context. Let $\tau$ be the number of data points used to
 make the last posterior approximation and $m$ is the number of
data points arriving since the last posterior approximation. EP
 requires the entire data set up to the current iteration to make
 a posterior approximation leading to a computational cost of
 $O((\tau + m)D^3)$ compared to ADFs $O(mD^3)$. For $m \ll \tau$
 this is going to result in a dramatic increase in computational cost
 choosing EP over ADF.

The computational cost vs accuracy trade-off is shown in
 the right column of Figure 3. The requirement for EP to be ran
 in batch is what causes the significant increase in the FLOPs
 use over the learning process. In these plots, an EP update is
 performed every 5,000 impressions ($m = 5,000$) rather than
 in a true sequential manner. This batch size was chosen purely
 for illustrative purposes; an EP update would have to be done
 more frequently early on in a bandit setting. If these batch
 sizes decreased then the computational cost would increase as a
 result.

3) Accuracy: As discussed above, MCMC can be used to
 provide an arbitrarily accurate representation of the posterior,
 although the computational effort involved in doing this is
 prohibitive in an online context. To assess accuracy for the
 purposes of this study, however, the No-U-Turn sampler
 (NUTS) was implemented using PyMC3 [30], and takes the
 output of this as the ground truth of the posterior distributions
 we are trying to approximate. By measuring the absolute
 error between the moments of the ground truth and each
 inference procedure, an empirical estimate for the asymptotic
 convergence rate of each can be computed by measuring the
 gradient of a log-log plot (see the left column of Figure 3).

The results show that EP gives better asymptotic accuracy
 to the posterior than other methods, as would be expected from
 the theoretical results discussed above, although for the real
 dataset considered, a simple power law in $T$ is not observed. In
 general it is clear that the main improvements for EP and ADF
 over Laplace come from estimation of the variance, as Figure
 2 shows more explicitly. This to be particularly important for
 recommendation systems, where balancing the explore-exploit
 tradeoff is crucial.

Also as expected, once data availability becomes large, one
 can switch to ADF as proposed in the FAB-COST algorithm
 without significant loss of accuracy. Although EP can be used
 periodically to update the posterior approximation in the FAB-
 COST algorithm, only two EP updates were performed in the
 simulations above (at 100 and 10,000 iterations) although the
 posterior accuracy would be improved with more frequent EP
 updates. Theses updates at 100 and 10,000 iterations out of the
 $T = 670,000$ sized data set shows that if given a reasonable
 prior, ADF achieves good accuracy.

4) Recommendation system and performance: Algorithm 5
 shows the pseudocode for FAB-COST. This takes the logistic
 bandit (shown in Algorithm 1) and adds moment updating
 using both EP and ADF. In the experiments, two EP updates
 were chosen at $E = 100$ and $E = 10,000$ (although this

![Fig. 3: The left column shows log-log plots of the mean
 and variance error for the Laplace approximation, ADF, EP
 and FAB-COST where the No U-Turn sampler was used to
 establish the ground truth. The grey lines show asymptotic
 error of $O(T^{-1})$ and $O(T^{-2})$ for reference. The right column
 shows the asymptotic accuracy computational cost trade-off
 between the methods as discussed in section IV-E2.](image-url)
Algorithm 5: FAB-COST

1. Set $E$ as the iteration(s) at which you want to make an EP update to the posterior.
2. Initialise the prior distribution $q_0(\theta) = \mathcal{N}(\mu_0, \Sigma_0)$.
3. for $t = 1 \ldots T$ do
   1. Generate a sample from the approximated posterior:
      $\tilde{\theta}_t \sim \mathcal{N}(\mu_{t-1}, \Sigma_{t-1})$
   2. Select an advert:
      $a_t = \arg\max_{j \in A}(A^\top \tilde{\theta}_t)$
   3. Update moments via ADF:
      $\mu_t = \mu_{t-1} + \Sigma_{t-1}^{-1} \alpha_t$
      $\Sigma_t = \Sigma_{t-1}^{-1} - \Sigma_{t-1}^{-1} (\alpha_t \alpha_t^\top - 2B_t) \Sigma_{t-1}^{-1}$
   4. IF $t = E$ , perform an EP approximation as shown in algorithm [3] using the last $E$ data points.
4. end

---

**REFERENCES**

[1] W. Chen, Y. Wang, and Y. Yuan, “Combinatorial multi-armed bandit: General framework and applications,” in *International Conference on Machine Learning*, 2013, pp. 151–159.

[2] T. Lattimore and C. Szepesvári, “Bandit algorithms,” preprint 2018.

[3] W. R. Thompson, “On the likelihood that one unknown probability exceeds another in view of the evidence of two samples,” *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

[4] O. Chapelle and L. Li, “An empirical evaluation of Thompson sampling,” in *Advances in neural information processing systems*, 2011, pp. 2249–2257.

[5] J. L. Doob, “Application of the theory of martingales,” *Le calcul des probabilités et ses applications*, pp. 23–27, 1949.

[6] T. P. Minka, “Expectation propagation for approximate bayesian inference,” in *Advances in Neural Information Processing Systems*, vol. 12, no. 11, pp. 2655–2664, 2000.

[7] M. Opper and O. Winther, “Gaussian processes for classification: Mean-field algorithms,” *Neural computation*, vol. 12, no. 11, pp. 2655–2684, 2000.

[8] ——, “A bayesian approach to on-line learning,” *On-line learning in neural networks*, pp. 363–378, 1998.

[9] S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, Eds. New York: Chapman and Hall/CRC.

[10] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, ser. Applied Logistic Regression. John Wiley & Sons, 2004.

---

**ACKNOWLEDGMENT**

A big thank you to David Hoyle and Robert Trickey from Auto Trader for their meets and discussions at the beginning of my PhD. Thanks to Edward Pyzer-Knapp from IBM research for his advice on how to take this paper forward. Acknowledgements also go to Simon Barthaléme for his emails and discussion on EP.
[11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer-Verlag, 2009.
[12] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.
[13] M. F. Dacrema, P. Cremonesi, and D. Jannach, "Are we really making much progress? a worrying analysis of recent neural recommendation approaches," in *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM, 2019, pp. 101–109.
[14] A. DasGupta, *Asymptotic theory of statistics and probability*. Springer Science & Business Media, 2008.
[15] J. D. Murray, *Asymptotic Analysis*. New York: Springer, 2012.
[16] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.
[17] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.
[18] J. M. Bernardo, "Approximations in statistics from a decision-theoretical viewpoint," in *Probability and Bayesian statistics*. Springer, 1987, pp. 53–60.
[19] J. Pearl, "Fusion, propagation, and structuring in belief networks," *Artificial intelligence*, vol. 29, no. 3, pp. 241–288, 1986.
[20] M. Seeger, "Expectation propagation for exponential families," Tech. Rep., 2005.
[21] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," *Mathematics of Operations Research*, vol. 39, no. 4, pp. 1221–1243, 2014.
[22] S. Agrawal and N. Goyal, "Thompson sampling for contextual bandits with linear payoffs," in *International Conference on Machine Learning*, 2013, pp. 127–135.
[23] M. Mézard, G. Parisi, and M. Virasoro, "Random free energies in spin glasses," *Journal de Physique Lettres*, vol. 46, no. 6, pp. 217–222, 1985.
[24] M. W. Seeger, "Bayesian inference and optimal design for the sparse linear model," *Journal of Machine Learning Research*, vol. 9, no. Apr, pp. 759–813, 2008.
[25] A. Gelman, A. Vehtari, P. Jylänki, C. Robert, N. Chopin, and J. P. Cunningham, "Expectation propagation as a way of life," *arXiv preprint arXiv:1412.4869*, vol. 157, 2014.
[26] G. Dehaene and S. Barthelmé, "Expectation propagation in the large data limit," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 80, no. 1, pp. 199–217, 2018.
[27] G. P. Dehaene and S. Barthelmé, "Bounding errors of expectation-propagation," in *Advances in Neural Information Processing Systems*, 2015, pp. 244–252.
[28] R. Herbrich, "Minimising the kullback-leibler divergence," *Microsoft, Tech. Rep.*, 2005.
[29] D. J. MacKay, "The evidence framework applied to classification networks," *Neural computation*, vol. 4, no. 5, pp. 720–736, 1992.
[30] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, "Probabilistic programming in python using PyMC3," *PeerJ Computer Science*, vol. 2, p. e55, 2016.