Variational Inference from Ranked Samples with Features

Yuan Guo[1] YUANEE20@ECE.NEU.EDU
Jennifer Dy[1] JDY@ECE.NEU.EDU
Deniz Erdoğmuş[1] ERDOGMUS@ECE.NEU.EDU
Jayashree Kalpathy-Cramer[2] KALPATHY@NMR.MGH.HARVARD.EDU
Susan Ostmo[3] OSTMO@OHSU.EDU
J. Peter Campbell[3] CAMPBELP@OHSU.EDU
Michael F. Chiang[3] CHIANGM@OHSU.EDU
Stratis Ioannidis[1] IOANNIDIS@ECE.NEU.EDU
1 *Department of Electrical and Computer Engineering, Northeastern University, MA, USA.*
2 *Department of Radiology, Massachusetts General Hospital, MA, USA.*
3 *Department of Ophthalmology, Casey Eye Institute, Oregon Health & Science University, OR, USA.*

Editors: Wee Sun Lee and Taiji Suzuki

Abstract

In many supervised learning settings, elicited labels comprise pairwise comparisons or rankings of samples. We propose a Bayesian inference model for ranking datasets, allowing us to take a probabilistic approach to ranking inference. Our probabilistic assumptions are motivated by, and consistent with, the so-called Plackett-Luce model. We propose a variational inference method to extract a closed-form Gaussian posterior distribution. We show experimentally that the resulting posterior yields more reliable ranking predictions compared to predictions via point estimates.
Keywords: Variational inference, Plackett Luce, Softmax Bound.

1. Introduction

In many supervised learning settings, elicited labels comprise pairwise comparisons or rankings of samples (Kamishima and Akaho, 2009; K-Cramer et al., 2016; Sculley, 2010; Negahban et al., 2018). For example, labelers may sort objects according to an underlying preference order, or select maximal or minimal items within a given set. Rankings and comparisons are, in general, more informative than class labels (Guo et al., 2019; Yıldız et al., 2019; Tian et al., 2019). They are also preferable as, in practice, human labelers find it easier to make comparative–rather than absolute–class membership judgments; this has been observed in several experimental contexts, e.g., in movie (Brun et al., 2010; Desarkar et al., 2010), travel (Zheng et al., 2009), and music (Koren and Sill, 2011) recommendations, as well as in labeling medical images (Stewart et al., 2005; K-Cramer et al., 2016; Tian et al., 2019).

The Placket-Luce model (Luce, 1959; Plackett, 1975) is a popular probabilistic model for performing inference over ranking datasets. Intuitively, it postulates that the probability that a sample is ranked higher than another is proportional to an inherent parameter. Traditionally, inference in this setting amounts to learning these sample parameters from ranking data (Hunter et al., 2004; Negahban et al., 2016, 2018; Maystre and Grossglauser, 2015). When samples have features, the Plackett-Luce model can naturally be extended to regress parameters, and thereby rankings, from features (Yıldız et al., 2019; Tian et al., 2019; Guo et al., 2019, 2018; Cheng et al., 2010).

Nevertheless, virtually all prior literature on regressing ranking from sample features has focused on maximum a posteriori (MAP) estimation. In this work, we approach ranking inference from a Bayesian point of view, allowing us to produce a posterior distribution on the learned Plackett-Luce model linking ranking to features. This distribution can be used to reason about parameter uncertainty (e.g, by constructing confidence intervals, etc.), which cannot be accomplished via point estimates. In summary, we make the following contributions:

- We propose a Bayesian inference model for ranking datasets, allowing us to take a probabilistic approach to ranking inference. Our probabilistic assumptions are motivated by, and consistent with, the Plackett-Luce model (Luce, 1959; Plackett, 1975).
- We study the posterior distribution of a Plackett-Luce parametric model linking rankings to sample features. We propose a variational inference method to extract a closed-form Gaussian posterior distribution under the Plackett-Luce model.
- Finally, we extensively evaluate the resulting variational inference method over real-life datasets. We show that the resulting closed-form posterior yields more reliable ranking predictions compared to predictions via point estimates.

The remainder of this paper is organized as follows. We discuss related work in Sec. 2. Our problem formulation and variational inference algorithm under the Plackett-Luce model can be found in Sec. 3 and Sec. 4, respectively. Our numerical evaluations are in Sec. 5, and we conclude in Sec. 7.

2. Related Work

The Placket-Luce Model (Plackett, 1975; Luce, 1959) is a classic generative model used for inference over ranking datasets. In the absense of features, it postulates that the probability that a sample is ranked higher than another is proportional to an inherent parameter, the so-called Plackett-Luce score. Maximum Likelihood Estimator (MLE) can be used to estimate these scores; though the log-likelihood function is not concave, there exists a reparametrization that converts MLE to a convex optimization problem (Hunter et al., 2004). Alternatively, Expectation-Maximization (EM) and Minorize-Maximation (MM) algorithms have been proposed to accelerate the computation of the MLE solution (Hunter et al., 2004; Gormley and Murphy, 2008). Alternative algorithms for MLE through shrinkage methods have also been proposed (Ragain et al., 2018; Rajkumar and Agarwal, 2014). Maystre and Grossglauser (2015) and Agarwal et al. (2018) propose spectral algorithms that significantly accelerate Plackett–Luce score estimation. In the Maximum A Posteriori (MAP) setting, given a Gamma prior distribution, Caron and Doucet (2012) provide iterative algorithms for obtaining MAP estimates for Placket-Luce scores.

When samples have features, several works propose regressing Plackett-Luce scores as either shallow (Cheng et al., 2010; Sculley, 2010; Tian et al., 2019; Guo et al., 2018) or deep (Yıldız et al., 2019; Sun et al., 2017) functions of features. In the shallow case, an appropriate parameterization again makes the log-likelihood function concave (Tian et al., 2019), and the Newton method can be used for parameter estimation. All of these works focus on MLE or MAP estimation, and none produce a Bayesian posterior on model parameters, as we do in this paper.

Bayesian inference has been applied to Plackett-Luce model scores, i.e., in the absence of sample features. Guiver and Snelson (2009) propose an inference scheme based on power expectation propagation, which is robust and can be applied to large datasets. By using an alternative Thurstonian

interpretation, Caron and Doucet (2012); Caron and Teh (2012) introduce latent variables that allow them to derive simple Gibbs samplers for the posterior distribution. Wang et al. (2017) propose an variational inference model to learn the mixtures in the feature-less setting. Alquier et al. (2016) discuss variational approximations of the Gibbs posterior. None of the above approaches readily generalize to the regression setting (i.e., one where samples have features) that we consider here.

Using variational inference as a means to approximate an intractable posterior is also classic (Bishop, 2006). In their seminal work, Jaakkola and Jordan (1997) consider a logistic regression model with a Gaussian prior distribution over the parameters. Our approach can be seen as an extension to the Plackett-Luce model; our algorithm reduces to the one by Jaakkola and Jordan (1997) when sets are pairs, i.e., in the so-called Bradley-Terry setting (Bradley and Terry, 1952). Khan et al. (2012) apply variational inference to multivariate categorical data using a stick-breaking likelihood function. Khan and Lin (2017) propose a conjugate computation variational inference which uses stochastic-gradient methods for non-conjugate terms. In our work, we exploit an upper bound of softmax function due to Bouchard (2007), also used by Ahmed and Campbell (2010) and Park and Choi (2010) for variational inference in a multi-class classification setting. We leverage and combine these techniques to show that, given a Gaussian prior distribution on model parameters, the approximation bound for the softmax function leads to a variational Gaussian posterior distribution for the Plackett-Luce model.

3. Problem Formulation

We consider a dataset of samples labeled by an expert as follows: labels are collected via comparisons that the expert makes among alternatives presented to her. We consider two different labeling settings, the *top-query* and the *ranking* setting. In the top-query setting, given a subset of the samples, the expert returns her top-choice among the presented alternatives. In the ranking setting, the expert is again given a subset of samples, but she returns a ranking, i.e., an ordering of the alternatives from highest to lowest. For example, the dataset could comprise medical images. When a subset of images is presented to a medical expert, she can select the image in which a disease is most prominent (top-query setting) or order the images w.r.t. the prevalence of the disease (ranking setting).

Formally, we have N samples, indexed by $i \in \mathcal{N} \equiv \{1, 2, \cdots, N\}$, each associated with a vector $\boldsymbol{x_i} \in \mathbb{R}^d$. The expert is presented with a set of alternatives $A \subseteq \mathcal{N}$, where $m = |A| \geqslant 2$ is the size of set A. In the top-query setting, the expert chooses the top item $c \in A$; in the ranking setting, the expert ranks the samples in A into an ordered sequence $(a_1 \succ a_2 \succ \cdots \succ a_m)$. Hence, in the top-query setting, for $\mathcal{L} = \{1, 2, \cdots, L\}$, we are given a dataset

$$D = \Big\{(A_l, c_l)\Big\}_{l \in \mathcal{L}}, \tag{1}$$

where $c_l \in A_l$ is the top choice in the set of alternatives $A_l \subseteq \mathcal{N}$. In contrast, in the ranking setting, we are given a dataset

$$D = \Big\{\big(A_l, \{a_i^l\}_{i=1}^{m_l}\big)\Big\}_{l \in \mathcal{L}}, \tag{2}$$

where $a_i^l \in A_l, i = 1, \cdots, m_l = |A_l|$, indicates the preference order of A_l: a_1^l is the most preferred alternative, a_2^l is the second preferred alternative, $a_{m_l}^l$ is the least preferred, etc. In both cases, we wish to perform Bayesian inference over dataset D. To do so, we describe our discriminative model in more detail below. For notational convenienve, we partition set $\mathcal{L}$ into two sets $\mathcal{L}_2, \mathcal{L}_{>2}$, defined as $\mathcal{L}_2 = \{l : |A_l| = 2\}$, $\mathcal{L}_{>2} = \{l : |A_l| > 2\}$. Note that $\mathcal{L} = \mathcal{L}_2 \cup \mathcal{L}_{>2}$ and $\mathcal{L}_2 \cap \mathcal{L}_{>2} = \emptyset$.

N	number of samples	L	number of sets of alternatives	i, j	sample indices
A_l	the set of alternatives	l	index of a set of alternatives	c_l	top-query of set A_l
$\boldsymbol{x_i}$	feature vector of sample i	D	dataset of comparison labels	s_i	Plackett Luce score for i
$\alpha, \boldsymbol{\zeta}, \boldsymbol{\xi}$	variational parameters	σ	sigmoid function	β, a	gamma parametric
n_i	total times that sample i is top query	ε	exponential distribution	Γ	gamma distribution
$\mathcal{L}$	$\{1, 2, \cdots, L\}$	$\mathcal{L}_2$	$\{l : \lvert A_l\rvert = 2\}$	$\mathcal{L}_{>2}$	$\{l : \lvert A_l\rvert > 2\}$
$\mathcal{N}$	the set for all samples	m_l	the size of set A_l	$\boldsymbol{s}$	Plackett Luce model score
$\mathbf{L}(\boldsymbol{s}; D)$	negative log-likelihood function	$\mathbb{1}$	indicator function	$p_0(\boldsymbol{s})$	prior distribution of $\boldsymbol{s}$
δ_{li}	whether i is in the set A_l	$\boldsymbol{\theta}$	parameter vector	$p_0(\boldsymbol{\theta})$	prior distribution of $\boldsymbol{\theta}$
$\overline{\mathbf{L}}_k$	variational lower bound at iteration k	$\lambda(t)$	$\lambda(t) = \frac{1}{4t}\tanh(\frac{t}{2})$	$\lvert E\rvert$	number of experts
$\overline{\mathbf{L}}_\infty$	variational lower bound ceiling				

Table 1: Summary of Notation

3.1. Plackett-Luce Model

Our descriminative model is based on the so-called Plackett-Luce model. Luce's choice axiom (Hunter et al., 2004; Luce, 1959; Maystre and Grossglauser, 2015) states that the relative preference of one item over another is not affected by the presence or absence of other items in the set of alternatives. Formally, let $p(c = i|A)$ be the probability of choosing item i when faced with alternatives in the set A. The Plackett Luce model postulates that: (a) events (c_l, A_l) are independent and (b) there exist parameters $s_i, i \in \mathcal{N}$, s.t. each event has probability:

$$p(c = i|A) = s_i / \left(\textstyle\sum_{j\in A} s_j\right). \tag{3}$$

A special case of Plackett Luce is the case when $|A_l| = 2$, i.e. the set of alternatives comprises pairwise-comparisons; this is also known as the Bradley-Terry model (Bradley and Terry, 1952). This is important in practice, as datasets often contain only pairs. Moreover, as we will see later on, our bounds become sharper in this case (see Lemma 2).

Under the Plackett-Luce model, the ranking setting can be reduced to the top-query setting by treating a ranking as the outcome of multiple independent top-queries: that is, ranking $(a_1 \succ a_2 \succ \cdots a_{K-1} \succ a_K)$ can be seen as the outcome of a_1 being selected as the top among the set of alternatives A, a_2 being the top among $A \setminus \{a_1\}$, etc. Assuming these selections are independent, Eq. (3) yields a joint probability:

$$p(a_1 \succ a_2 \succ \cdots a_{K-1} \succ a_K|A) = \frac{s_{a_1}}{\sum_{j=1}^{K} s_{a_j}} \frac{s_{a_2}}{\sum_{j=2}^{K} s_{a_j}} \cdots \frac{s_{a_{K-1}}}{s_{a_{K-1}} + s_{a_K}}. \tag{4}$$

Note that under the independence assumption of the Plackett-Luce model Eq. (4) implies that a dataset of form Eq. (2) can be converted to a dataset of form (1) having the *same* joint probability distribution. This amounts to breaking each ranking of a set A_l to the equivalent $|A_l| - 1$ independent top queries. For this reason, we focus on the top-query setting from this point on, keeping this equivalence in mind.

3.2. Inference over the Plackett Luce Model

In the absence of features, inference amounts to the determination of parameters $\boldsymbol{s}$ given (top-query) dataset D defined as in Eq. (1). As alternative sets are independent, the total probability is:

$$p(D|\boldsymbol{s}) = \textstyle\prod_{l\in\mathcal{L}} p(A_l|\boldsymbol{s}) = \prod_{l\in\mathcal{L}} \frac{s_{c_l}}{\sum_{i\in A_l} s_i}, \tag{5}$$

where $\boldsymbol{s} = [s_i]_{i\in\mathcal{N}} \in \mathbb{R}_+^N$. Given a prior distribution $p_0(\boldsymbol{s})$ for $\boldsymbol{s}$, we can infer $\boldsymbol{s}$ through maximum a posteriori estimation (MAP) over the model (5). Then, the estimation of $\boldsymbol{s}$ amounts to minimizing the following negative log-likelihood function:

$$\mathbf{L}(\boldsymbol{s}; D) = -\sum_{l=1}^{L} \log p(A_l|\boldsymbol{s}) - \log p_0(\boldsymbol{s}). \tag{6}$$

Under the Plackett Luce model, the negative log-likelihood function for $\boldsymbol{s}$ is not convex. For Maximum Likelihood estimation (MLE) i.e., when we do not introduce a prior, we can write $s_i = e^{\theta_i}$, which makes negative log-likelihood convex with respect to $\boldsymbol{\theta} = [\theta_i]_{i\in\mathcal{N}} \in \mathbb{R}^N$ (Hunter et al., 2004; Khetan and Oh, 2016; Rajkumar and Agarwal, 2014; Negahban et al., 2016). There also exist fast iterative algorithms to solve problem (5). For example, Hunter et al. (2004) proposes the minorize-maximization (MM) algorithm for Eq. (5), while recent spectral algorithms accelerate this further (Maystre and Grossglauser, 2015; Agarwal et al., 2018; Kumar et al., 2015). A commonly used prior distribution for $\boldsymbol{s}$ is the Gamma distribution: $p_0(\boldsymbol{s}) = \prod_{i=1}^{N} \frac{\beta^a}{\Gamma(a)} e^{-\beta s_i} s_i^{(a-1)}$, where $a, \beta > 0$. There is an iterative algorithm to minimize $\mathbf{L}(\boldsymbol{s}; D)$ in Eq. (6) under a Gamma prior (Hunter et al., 2004; Caron and Doucet, 2012): at step $k+1$,

$$s_i^{(k+1)} = (a - 1 + n_i)\Big[\beta + \sum_{l=1}^{L} \frac{\delta_{li}}{\sum_{t\in A_l} s_t^{(k)}}\Big]^{-1}, \quad \forall\, i \in \mathcal{N}, \tag{7}$$

where $n_i = \sum_{l\in\mathcal{L}} \mathbb{1}_{i=c_l}$, $i \in \mathcal{N}$, counts the total times that sample i is the top item amongst alternatives, and $\delta_{li} = \mathbb{1}_{i\in A_l}$ $(i \in \mathcal{N}, l \in \mathcal{L})$ indicates whether sample i is in the set of alternatives A_l.

Bayesian Inference over the Plackett Luce Model. The posterior distribution $p(\boldsymbol{s}|D)$ satisfies:

$$p(\boldsymbol{s}|D) = p_0(\boldsymbol{s})\frac{p(D|\boldsymbol{s})}{p(D)} \propto p_0(\boldsymbol{s}) \prod_{l\in\mathcal{L}} \frac{s_{c_l}}{\sum_{t\in A_l} s_t}. \tag{8}$$

Given a Gamma prior distribution $p_0(\boldsymbol{s})$, the posterior distribution may be intractable, but Gibbs sampling can be used to estimate it (Caron and Doucet, 2012; Caron and Teh, 2012; Caron et al., 2014). In particular, for $l = 1, 2, \cdots, L$, we first sample $z_l^{(k+1)}$ from the exponential distribution $\mathcal{E}(\sum_{t\in A_l} s_t^{(k)})$ (z_l is the auxiliary variable given by Thurstonian interpretation). Then, for $i = 1, \cdots, N$, we sample $s_i^{(k+1)}$ from the Gamma distribution $\Gamma(a + n_i, \beta + \sum_{l=1}^{L} \delta_{li} z_l^{(k+1)})$, where n_i, δ_{li} are defined as the same as Eq. (7).

4. Plackett Luce Model Incorporating Features

In our setting, every sample $i \in \mathcal{N}$ has a feature vector $\boldsymbol{x}_i \in \mathbb{R}^d$. We assume that there exists a parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$, sampled from a prior distribution $p_0(\boldsymbol{\theta})$, so that s_i in Eq. (6) satisfies:

$$s_i = e^{\boldsymbol{\theta}^T \boldsymbol{x}_i}. \tag{9}$$

Our goal is to perform variational inference to estimate the posterior of $\boldsymbol{\theta} \in \mathbb{R}^d$. Given a prior $p_0(\boldsymbol{\theta})$, the posterior distribution $p(\boldsymbol{\theta}|D)$ for parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$ satisfies:

$$p(\boldsymbol{\theta}|D) \propto p_0(\boldsymbol{\theta}) \prod_{l\in\mathcal{L}} \frac{\exp(\boldsymbol{x}_{c_l}^T \boldsymbol{\theta})}{\sum_{j\in A_l} \exp(\boldsymbol{x}_j^T \boldsymbol{\theta})}. \tag{10}$$

To approximate the posterior distribution by a distribution $q(\boldsymbol{\theta})$ that belongs to a restricted family (e.g., it is Gaussian), we identify a $q(\boldsymbol{\theta})$ that maximizes the Evidence Lower Bound (ELBO):

$$\mathbf{L}(q) = \mathbb{E}_{q(\boldsymbol{\theta})}\Big[\log \frac{p(D|\boldsymbol{\theta})p_0(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\Big] \overset{(5)(9)}{=\!=} \mathbb{E}_{q(\boldsymbol{\theta})}\Big[\log \frac{\Big(\prod_{l\in\mathcal{L}} \frac{\exp(\boldsymbol{x}_{c_l}^T\boldsymbol{\theta})}{\sum_{j\in A_l}\exp(\boldsymbol{x}_j^T\boldsymbol{\theta})}\Big)p_0(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\Big]. \tag{11}$$

This is equivalent to maximizing the KL divergence between $q(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|D)$ (Murphy, 2012).

4.1. Variational Lower Bound

We first establish a variational lower bound on the ELBO. We make use of an auxiliary lemma:

Lemma 1 *For the Plackett Luce model (3) in which $s_i = e^{\boldsymbol{\theta}^T\boldsymbol{x}_i}, \forall i \in \mathcal{N}$, if $|A| = 2$, i.e, $A = \{c_l, \bar{c}_l\}$ for some $c_l, \bar{c}_l \in \mathcal{N}$, then for all $\xi \in \mathbb{R}_+$, we have that:*

$$p(c = c_l|A) \geqslant \sigma(\xi)e^{(\boldsymbol{x}_{c_l,\bar{c}_l}^T\boldsymbol{\theta}-\xi)/2-\lambda(\xi)((\boldsymbol{x}_{c_l,\bar{c}_l}^T\boldsymbol{\theta})^2-\xi^2)}, \tag{12}$$

where $\sigma(\xi) = \frac{1}{1+e^{-\xi}}$, $\lambda(\xi) = \frac{1}{4\xi}\tanh(\frac{\xi}{2})$ and $\boldsymbol{x}_{c_l,\bar{c}_l} = \boldsymbol{x}_{c_l} - \boldsymbol{x}_{\bar{c}_l} \in \mathbb{R}^d$, for $c_l, \bar{c}_l \in \mathcal{N}$. If $|A| > 2$, for any $\xi_j \in \mathbb{R}_+, j \in A$, and any $\alpha \in \mathbb{R}$, we have that:

$$p(c = c_l|A) \geqslant e^{\boldsymbol{x}_{c_l}^T\boldsymbol{\theta}-\alpha}\prod_{j\in A}\Big(\sigma(\xi_j)e^{(-\boldsymbol{x}_j^T\boldsymbol{\theta}+\alpha-\xi_j)/2-\lambda(\xi_j)((\boldsymbol{x}_j^T\boldsymbol{\theta}-\alpha)^2-\xi_j^2)}\Big). \tag{13}$$

The proof is in Appendix A of the supplement. Lemma 1 allows us to bound the ELBO as follows:

Lemma 2 *Assume that the prior distribution is Gaussian, i.e.: $p_0(\boldsymbol{\theta}) = \frac{1}{B_0}e^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu}_0)^T\boldsymbol{S}_0^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_0)}$, where $B_0 = (2\pi)^{d/2}|\boldsymbol{S}_0|^{1/2}$, $\boldsymbol{\mu}_0 \in \mathbb{R}^d$, and $\boldsymbol{S}_0 \in \mathbb{R}^{d\times d}$. Assume that $q(\boldsymbol{\theta})$ is a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S})$. Then, for all $\boldsymbol{\xi} = [\xi_{lj}]_{l\in\mathcal{L}_{>2}, j\in A_l} \in \mathbb{R}^{\sum_{l\in\mathcal{L}_{>2}}|A_l|}$, $\boldsymbol{\alpha} = [\alpha_l]_{l\in\mathcal{L}_{>2}} \in \mathbb{R}^{|\mathcal{L}_{>2}|}$ and $\boldsymbol{\zeta} = [\zeta_l]_{l\in\mathcal{L}_2} \in \mathbb{R}^{|\mathcal{L}_2|}$, the ELBO $\mathbf{L}(q)$ in Eq. (11) is lower-bounded by:*

$$\mathbf{L}(\boldsymbol{\zeta}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{S}) = \mathbb{E}_{q(\boldsymbol{\theta})}\Big[\textstyle\sum_{l\in\mathcal{L}} Q_l\Big] + \frac{1}{2}\log\frac{|\boldsymbol{S}|}{|\boldsymbol{S}_0|} + \mathbb{E}_{q(\boldsymbol{\theta})}\Big[\frac{(\boldsymbol{\theta}-\boldsymbol{\mu})^T\boldsymbol{S}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})}{2} - \frac{(\boldsymbol{\theta}-\boldsymbol{\mu}_0)^T\boldsymbol{S}_0^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_0)}{2}\Big], \tag{14}$$

where Q_l is the logarithm of the lower bound in Lemma 1, given by:

$$Q_l = \begin{cases} \boldsymbol{x}_{c_l}^T\boldsymbol{\theta} - \alpha_l + \sum_{j\in A_l}\Big[\log\sigma(\xi_{lj}) - \frac{\boldsymbol{x}_j^T\boldsymbol{\theta}-\alpha_l+\xi_{lj}}{2} - \lambda(\xi_{lj})\big((\boldsymbol{x}_j^T\boldsymbol{\theta}-\alpha_l)^2 - \xi_{lj}^2\big)\Big], & l \in \mathcal{L}_{>2}, \\ \log\sigma(\zeta_l) + (\boldsymbol{x}_{c_l,\bar{c}_l}^T\boldsymbol{\theta} - \zeta_l)/2 - \lambda(\zeta_l)((\boldsymbol{x}_{c_l,\bar{c}_l}^T\boldsymbol{\theta})^2 - \zeta_l^2), & l \in \mathcal{L}_2, \end{cases} \tag{15}$$

where $A_l = \{c_l, \bar{c}_l\}$ and $\boldsymbol{x}_{c_l,\bar{c}_l} = \boldsymbol{x}_{c_l} - \boldsymbol{x}_{\bar{c}_l}$ for $l \in \mathcal{L}_2$. The proof is in Appendix B of the supplement.

4.2. Variational Lower Bound Optimization

To produce our estimate of the posterior, we follow the classic approach (Bishop, 2006; Jaakkola and Jordan, 1997) of minimizing the variational lower bound (14) on the ELBO rather the ELBO itself. We do so using am alternating maximization algorithm, i.e., alternately optimizing (14) w.r.t. its distribution and bound parameters. Formally, for $k \in \mathbb{N}$:

$$\boldsymbol{S}^{(k)}, \boldsymbol{\mu}^{(k)} = \mathbf{argmax}_{\boldsymbol{S},\boldsymbol{\mu}} \;\; \mathbf{L}(\boldsymbol{\zeta}^{(k)}, \boldsymbol{\xi}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\mu}, \boldsymbol{S}) \tag{16a}$$

$$\boldsymbol{\xi}^{(k+1)}, \boldsymbol{\zeta}^{(k+1)}, \boldsymbol{\alpha}^{(k+1)} = \mathbf{argmax}_{\boldsymbol{\xi},\boldsymbol{\zeta},\boldsymbol{\alpha}} \;\; \mathbf{L}(\boldsymbol{\zeta}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}^{(k)}, \boldsymbol{S}^{(k)}). \tag{16b}$$

Algorithm 1 Variational Inference Alternating Maximization

Prior parameter $\boldsymbol{\mu}_0, \boldsymbol{S}_0$, feature vector $\boldsymbol{X}$, top-query data $D = \{c_l, A_l\}_{l \in \mathcal{L}}$.

Initialize $\boldsymbol{\xi}, \boldsymbol{\zeta}, \boldsymbol{\alpha}$
while stopping criterion (26) is not satisfied **do**
  $\boldsymbol{\mu}, \boldsymbol{S} =$ MINORIZATION$(\boldsymbol{\xi}, \boldsymbol{\zeta}, \boldsymbol{\alpha})$.
  $(\boldsymbol{\xi}, \boldsymbol{\zeta}, \boldsymbol{\alpha}) =$ MAXIMIZATION$(\boldsymbol{\mu}, \boldsymbol{S})$.
end while
return $\boldsymbol{\mu}, \boldsymbol{S}$

procedure MINORIZATION($\boldsymbol{\xi}, \boldsymbol{\zeta}, \boldsymbol{\alpha}$)
  Compute $\boldsymbol{S}$ via (17).
  Compute $\boldsymbol{\mu}$ via (18).
end procedure

procedure MAXIMIZATION($\boldsymbol{\mu}, \boldsymbol{S}$)
  Compute $\boldsymbol{\zeta}$ as (19).
  while not converged **do**
    fix $\boldsymbol{\xi}$, compute $\boldsymbol{\alpha}$ via (22).
    fix $\boldsymbol{\alpha}$, compute $\boldsymbol{\xi}$ via (23).
  end while
end procedure

In the Bradley-Terry case (i.e., over pairwise comparisons), we can compute both (16a) and (16b) in closed form, using the approach outlined in Jaakkola and Jordan (1997). The general Plackett-Luce model, however, cannot be addressed by this approach, as the step (16b) does not admit a closed-form solution. Note that the conversion of rankings to top-queries described in Section 3.1 yields both pairwise (i.e., Bradley-Terry) and general top-queries; we therefore describe how to treat both cases. We outline our approach below; a summary of the algorithm is given in Alg. 1.

Step (16a) (MINORIZATION): Step (16a) admits a closed form, following the same argument as Bishop (2006); this is given by the following lemma, proved in Appendix C of the supplement.

Lemma 3 *The solution $\boldsymbol{S}^k$, $\boldsymbol{\mu}^k$ to (16a) can be written as:*

$$[\boldsymbol{S}^{(k)}]^{-1} = \boldsymbol{S}_0^{-1} + 2\sum_{l \in \mathcal{L}_2} \lambda(\zeta_l^{(k)}) \boldsymbol{x}_{c_l,\bar{c}_l} \boldsymbol{x}_{c_l,\bar{c}_l}^T + 2\sum_{l \in \mathcal{L}_{>2}} \sum_{j \in A_l} \lambda(\xi_{lj}^{(k)}) \boldsymbol{x}_j \boldsymbol{x}_j^T, \tag{17}$$

$$\boldsymbol{\mu}^{(k)} = \boldsymbol{S}^{(k)} \left(\boldsymbol{S}_0^{-1} \boldsymbol{\mu}_0 + \sum_{l \in \mathcal{L}_2} \boldsymbol{x}_{c_l,\bar{c}_l}/2 + \sum_{l \in \mathcal{L}_{>2}} \left(\boldsymbol{x}_{c_l} + \sum_{j \in A_l} \left(2\lambda(\xi_{lj}^{(k)}) \alpha_l^{(k)} \boldsymbol{x}_j - \boldsymbol{x}_j/2 \right) \right) \right). \tag{18}$$

Step (16b) (MAXIMIZATION): For step (16b), we first observe that the maximization of $\boldsymbol{\zeta}$ can be done independently of the optimization w.r.t. $\boldsymbol{\xi}$ and $\boldsymbol{\alpha}$; moreover, it again admits a closed-form solution; the proof is identical to the one by Jaakkola and Jordan (1997), which we prove in Appendix D of the supplement; this is precisely because, in the Bradley-Terry case (i.e., when $|A_l| = 2$), inferring $\boldsymbol{\theta}$ amounts to logistic variational inference over the feature differences $\boldsymbol{x}_{c_l,\bar{c}_l}$:

Lemma 4 *Given $\boldsymbol{S}^{(k)}, \boldsymbol{\mu}^{(k)}$, the solution of (16b) w.r.t. $\boldsymbol{\zeta}$ can be written as*

$$\zeta_l^{(k+1)} = \sqrt{\boldsymbol{x}_{c_l,\bar{c}_l}^T (\boldsymbol{S}^{(k)} + \boldsymbol{\mu}^{(k)} \boldsymbol{\mu}^{(k)T}) \boldsymbol{x}_{c_l,\bar{c}_l}}, \qquad l \in \mathcal{L}_2. \tag{19}$$

Given $\boldsymbol{S}^{(k)}, \boldsymbol{\mu}^{(k)}$, optimizing (16b) w.r.t. $\boldsymbol{\xi}$ and $\boldsymbol{\alpha}$ can again be done separately from optimizing $\boldsymbol{\zeta}$; however, the former two variables are coupled. We solve this optimization via an inner alternating maximization as well. In particular, problem (16b) amounts to solving problems of the following form:

$$\boldsymbol{\xi}_l^{(k+1)}, \alpha_l^{(k+1)} = \mathbf{argmax}_{\boldsymbol{\xi}_l, \alpha_l} \mathbb{E}_{q(\boldsymbol{\theta})}[Q_l], \quad l \in \mathcal{L}_{>2}, \tag{20}$$

where Q_l are given by (15). We solve these problems via alternating maximization for $n \in \mathbb{N}$ (Ahmed and Campbell, 2010):

$$\alpha_l^{(n+1)} = \mathbf{argmax}_{\alpha_l} \; f_l(\boldsymbol{\xi}_l^{(n)}, \alpha_l), \;\; l \in \mathcal{L}_{>2}, \tag{21a}$$

$$\boldsymbol{\xi}_l^{(n+1)} = \mathbf{argmax}_{\boldsymbol{\xi}_l} \; f_l(\boldsymbol{\xi}_l, \alpha_l^{(n+1)}), \;\; l \in \mathcal{L}_{>2}, \tag{21b}$$

where $f_l(\boldsymbol{\xi}_l, \alpha_l) = \mathbb{E}_{q(\boldsymbol{\theta})}[Q_l]$. By decoupling this computation via alternating maximization, we can again obtain closed-form formulas for (21):

Lemma 5 *Given $\boldsymbol{S}^{(k)}, \boldsymbol{\mu}^{(k)}$, the solution of (21a) w.r.t α_l has a closed form:*

$$\alpha_l^{(n+1)} = \big((m_l - 2)/4 + \textstyle\sum_{j \in A_l} \lambda(\xi_{lj}^{(n)}) \boldsymbol{x}_j^T \boldsymbol{\mu}^{(k)}\big) / \sum_{j \in A_l} \lambda(\xi_{lj}^{(n)}), \qquad l \in \mathcal{L}_{>2}, \tag{22}$$

where m_l is the size of set A_l, and the solution of (21b) w.r.t $\boldsymbol{\xi_l}$ also has a closed form:

$$\xi_{lj}^{(n+1)} = \sqrt{\boldsymbol{x}_j^T \boldsymbol{S}^{(k)} \boldsymbol{x}_j + \big(\boldsymbol{x}_j^T \boldsymbol{\mu}^{(k)} - \alpha_l^{(n+1)}\big)^2}, \;\; l \in \mathcal{L}_{>2}, \; j \in A_l. \tag{23}$$

The proof is in Appendix E of the supplement.

4.3. Tracking the Lower Bound

The monotonicity of steps (16a) and (16b) implies that:

$$\begin{aligned} \mathbf{L}(\boldsymbol{\zeta}^{(k)}, \boldsymbol{\xi}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{S}^{(k)}) \leq & \mathbf{L}(\boldsymbol{\zeta}^{(k+1)}, \boldsymbol{\xi}^{(k+1)}, \boldsymbol{\alpha}^{(k+1)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{S}^{(k)}) \\ \leq & \mathbf{L}(\boldsymbol{\zeta}^{(k+1)}, \boldsymbol{\xi}^{(k+1)}, \boldsymbol{\alpha}^{(k+1)}, \boldsymbol{\mu}^{(k+1)}, \boldsymbol{S}^{(k+1)}). \end{aligned} \tag{24}$$

We can thus see that function $\mathbf{L}(\boldsymbol{\zeta}^{(k)}, \boldsymbol{\xi}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{S}^{(k)})$ is monotone w. r. t k. As noted in Bishop (2006), the intermediate value in (24)–i.e., the lower bound after (16a) step–has a simple form, and can keep track of the lower bound. In our setting, this is given by:

$$\begin{aligned} \overline{\mathbf{L}}_k = & \mathcal{L}(\boldsymbol{\zeta}^{(k+1)}, \boldsymbol{\xi}^{(k+1)}, \boldsymbol{\alpha}^{(k+1)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{S}^{(k)}) \\ = & \textstyle\sum_{l \in \mathcal{L}} M_l^{(k)} + (\log(|\boldsymbol{S}^{(k)}|/|\boldsymbol{S}_0|))/2 + \boldsymbol{\mu}^{(k)T} [\boldsymbol{S}^{(k)}]^{-1} \boldsymbol{\mu}^{(k)}/2 - \boldsymbol{\mu}_0^T \boldsymbol{S}_0^{-1} \boldsymbol{\mu}_0/2, \end{aligned} \tag{25}$$

where $M_l^{(k)} = \begin{cases} \sum_{j \in A_l} \big(\log \sigma(\xi_{lj}^{(k)}) - \xi_{lj}^{(k)}/2 + \lambda(\xi_{lj}^{(k)})([\xi_{lj}^{(k)}]^2 - [\alpha_l^{(k)}]^2)\big) + (m_l - 2)\alpha_l^{(k)}/2, & l \in \mathcal{L}_{>2}, \\ \log \sigma(\zeta_l^{(k)}) - \zeta_l^{(k)}/2 + \lambda(\zeta_l^{(k)})[\zeta_l^{(k)}]^2, & l \in \mathcal{L}_2. \end{cases}$

The proof is in Appendix F of the supplement.

5. Evaluation

5.1. Datasets

ROP Dataset. The ROP dataset (K-Cramer et al., 2016) is a comparison dataset, i.e., it is of the form given by Eq. (1) with $m_l = 2$. It consists of $N = 100$ images of retinas, labeled by experts w.r.t. the presence of a disease called Retinopathy of Prematurity (ROP). We represent each image through a vector $\boldsymbol{x}_i \in \mathbb{R}^d$ where $d = 156$, using the feature extraction procedure of K-Cramer et al. (2016), comprising statistics of several indices such as blood vessel curvature, dilation, and tortuosity. Five

Name	Labelers $\lvert E\rvert$	Repetitions	N_e	d	m_l	$\lvert\mathcal{L}_e\rvert$	sample	Type	ϵ
ROP	5	6	100	156	2	4950	100	Comparison	$1e^{-6}$
Netflix-Com	30	1	1079-1198	30	2	6000	1000	Comparison	$1e^{-7}$
CAMRa-Com	30	1	1028-3300	10	2	6000	1000	Comparison	$1e^{-7}$
MLSR-Com	30	1	100-400	134	2	6000	1000	Comparison	$1e^{-6}$
Netflix-Rank	30	1	1079-1198	30	5	1500	300	Ranking	$1e^{-7}$
CAMRa-Rank	30	1	1028-3300	10	5	1500	300	Ranking	$1e^{-7}$
MLSR-Rank	30	1	100-400	134	3	1500	250	Ranking	$1e^{-6}$
Sushi	1(General)	30	100	20	10	500	100	Ranking	$1e^{-6}$

Table 2: Dataset Summary

experts provide binary comparison labels indicating severity between image pairs. Each expert e provides $|\mathcal{L}_e| = 4950$ pairwise comparisons between these 100 images[1].

Netflix Dataset. The Netflix dataset has 600 users and 17770 movies. We select 30 users who have rated more than $N = 1079$ movies. Each movie has a feature vector $\boldsymbol{x}_i \in \mathbb{R}^d$, with $d = 30$, obtained via matrix factorization (Koren et al., 2009) over the entire dataset. In the original data, each movie has a score between 1 to 5. We generate 2 synthetic datasets, one containing rankings (Netflix-Rank) and one containing pairwise comparisons (Netflix-Com). For each user, we can generate synthetic ranking data in the form (2) with $m_l = 5$ and pairwise comparison data in the form (1) with $m_l = 2$. For Netflix-Rank data, we select sets A_l uniformly at random among all sets of 5 movies with different ratings, then rank the five movies from highest to lowest. We set $|\mathcal{L}_e| = 6000$ per user for Netflix-Rank. For Netflix-Com data, each pairwise comparison $\{c_l, \bar{c}_l\}$ is selected by uniformly at random among a pair of samples with different scores; in this data, we generate $|\mathcal{L}_e| = 1500$ comparisons per user e.

CAMRa Dataset. The CAMRa dataset (Bento et al., 2011) has 640 users and 23893 movies. We select 30 users who have rated more than $N = 1028$ movies. Each movie has a feature vector $\boldsymbol{x}_i \in \mathbb{R}^d$, with $d = 10$, obtained via matrix factorization over the entire dataset. Each movie has a score ranging from 1 to 5. As for Netflix, we generate 2 synthetic datasets, one containing rankings (CAMRa-Rank) and one containing pairwise comparisons (CAMRa-Com). We set $|\mathcal{L}_e| = 6000$ rankings per user e for CAMRa-Rank and $|\mathcal{L}_e| = 1500$ comparisons per user e in CAMRa-Com.

MSLR Dataset. The MSLR-WEB10K dataset (Qin and Liu, 2013) has 10000 queries. The dataset consists of 134-dimensional features such as covered query term number, covered query term ratio, stream length, inverse document frequency (IDF), etc. Relevance judgments are obtained from a labeling set of a commercial web search engine (Microsoft Bing) queries, which take 5 values from 0 (irrelevant) to 4 (perfectly relevant). As we did for users in the Netflix and CAMRa datasets, for each query ID e, we generate synthetic ranking data in the form (2) with $m_l = 3$ (the sample with relevance judgment from 0 to 2) and pairwise comparison data in the form (1) with $m_l = 2$. We generate $|\mathcal{L}_e| = 6000$ per query ID e for MLSR-Rank and $|\mathcal{L}_e| = 1500$ comparisons per query ID e for MSLR-Com.

SUSHI Dataset. The SUSHI Preference dataset (Kamishima et al., 2009) consists of rankings of $N = 100$ sushi food items by $|\mathcal{L}| = 5000$ customers. Each customer ranks $m_l = 10$ items according to her preferences, hence the data can be expressed in form (2) in which $m_l = 10$. Each sushi item is associated with a feature vector $\boldsymbol{x}_i \in \mathbb{R}^d$ where $d = 20$, consisting of features such as style, group, heaviness/oiliness in taste, frequency, and normalized price.

1. Some experts observe and rate the same pair more than once.

5.2. Experiment Setup

Cross Validation. We perform two types of 3-fold cross-validation: over the dataset $\mathcal{L}$ and over samples $\mathcal{N}$; we describe both in more detail below.

$\mathcal{L}$ Partition: For the dataset D defined in either form (1) or (2), we perform standard cross-validation over $\mathcal{L}$: that is, we split $\mathcal{L}$ into a training set $\mathcal{L}_{train}$ and test set $\mathcal{L}_{test}$. This corresponds to a standard partitioning of set D into $D_{train} = \{A_l | l : l \in \mathcal{L}_{train}\}$ and $D_{test} = \{A_l | l : l \in \mathcal{L}_{test}\}$. Note that, under this partitioning, samples $\mathcal{N}$ appear in both the training and test set.

$\mathcal{N}$ Partition: We also partition the dataset D by $\mathcal{N}$. To do so, we first partition $\mathcal{N}$ into $\mathcal{N}_{train}$ and $\mathcal{N}_{test}$, and then set the training set to $D_{train} = \{A_l | l : A_l \subseteq \mathcal{N}_{train}\}$ and the test set to $D_{test} = \{A_l | l : A_l \subseteq \mathcal{N}_{test}\}$. Note that, for every $l \in \mathcal{L}$, sets A_l that contain samples in both $\mathcal{N}_{train}$ *and* $\mathcal{N}_{test}$ are dropped.

For ROP, Netflix, CAMRa and MLSR we cross validate each dataset D_e separately across experts/users/queries $e \in E$. For ROP (in which we only have 5 experts), we repeat the 3-fold partitioning 6 times. For Sushi, we treat all rankings as generated by one labeller, and repeat the random 3-fold partitioning 30 times. The size $|E|$ and the number of repetitions for each dataset are summarized in Table 2: note that, as a result, we perform exactly 30 cross-validations in total per dataset. To speed up training, we train only on a sample of the entire training set $D_{train}(\mathcal{L}_e)$; the sample size for each dataset is also in Table 2. In all cases, although we subsample the training set, we evaluate our models on entire test set D_{test}.

Algorithms. We use a Gaussian prior with mean $\mathbf{0}$ and covariance $\boldsymbol{S}_0 = 1/\eta \mathbf{I}_d \in \mathbb{R}^{d \times d}$, where η is a hyperparameter. We implement Alg. 1 to infer post parameter $\boldsymbol{\mu}, \boldsymbol{S}$. We set the stopping criterion to:

$$\sqrt{||\boldsymbol{\xi}^{(k+1)} - \boldsymbol{\xi}^{(k)}||_2^2 + ||\boldsymbol{\zeta}^{(k+1)} - \boldsymbol{\zeta}^{(k)}||_2^2} \leqslant \epsilon, \tag{26}$$

where $\boldsymbol{\zeta}$ and $\boldsymbol{\xi}$ are given by Eq. (19) and Eq. (23) respectively. We define $\overline{\mathbf{L}}_\infty$ as the variational lower bound $\overline{\mathbf{L}}_k$ given by (25), at the last iteration. We also implement maximum a posteriori estimation (MAP) to infer an estimate $\hat{\boldsymbol{\theta}}_{MAP}$, which is the $\boldsymbol{\theta} \in \mathbb{R}^d$ maximizing Eq. (6) with $s_i = \boldsymbol{x}_i^T \boldsymbol{\theta}$, $i \in \mathcal{N}$.

Metrics. We measure two evaluation metrics in the test set for both ranking and comparison data. First, we measure comparison AUC, by treating comparisons as binary variables; to measure AUC over rankings, we break rankings of length M to the corresponding $M(M-1)$ pairwise comparisons and treat them as binary variables to be predicted. We also measure a *top-K* metric, defined below; we use this metric as a means of evaluating performance over a task that would be more sensitive to correct posterior distribution estimation than AUC. Intuitively, the top-K metric measures the "value" of an estimate of the K top ranked items in the test set. Formally, for each item $i \in \mathcal{S}$, where S is the set of items in the test set, we define i's ground truth "value" as:

$$w_i = \frac{\text{\#comparison pairs in } D_{test} \text{ which } i \text{ is the winner}}{\text{\#comparison pairs in } D_{test} \text{ containing } i} = \frac{\sum_{A_l \subseteq D_{test}} \sum_{(i,j) \in A_l} \mathbb{1}_{(c=i|i,j)}}{\sum_{A_l \subseteq D_{test}} \sum_{(i,j) \in A_l} 1}. \tag{27}$$

We estimate the value of item $i \in \mathcal{S}$ as:

$$\hat{w}_i = \Big(\sum_{A_l \subseteq D_{test}} \sum_{(i,j) \in A_l} \mathbf{P}(c = i | i, j) \Big) / \Big(\sum_{A_l \subseteq D_{test}} \sum_{(i,j) \in A_l} 1 \Big). \tag{28}$$

where $\mathbf{P}(c{=}i|i,j){=}1/(1{+}\exp(-\boldsymbol{x}_{i,j}^T \hat{\boldsymbol{\theta}}))$ for MAP, and $\mathbf{P}(c{=}i|i,j){=}\mathbb{E}_{\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{S})}[1/(1{+}\exp(-\boldsymbol{x}_{i,j}^T \boldsymbol{\theta}))]$ for variational inference. We define the *top-K* metric as the ratio $T_K{=}W(\hat{\mathcal{S}}^+)/(\max_{\mathcal{S}:|\mathcal{S}|=K} W(\mathcal{S}))$, where $W(\mathcal{S}){=}\sum_{i \in \mathcal{S}} w_i$, and $\hat{\mathcal{S}}^+{=}\arg\max_{\mathcal{S}:|\mathcal{S}|=K} \hat{W}(\mathcal{S})$ for $\hat{W}(\mathcal{S}){=}\sum_{i \in \mathcal{S}} \hat{w}_i$. Intuitively, this is

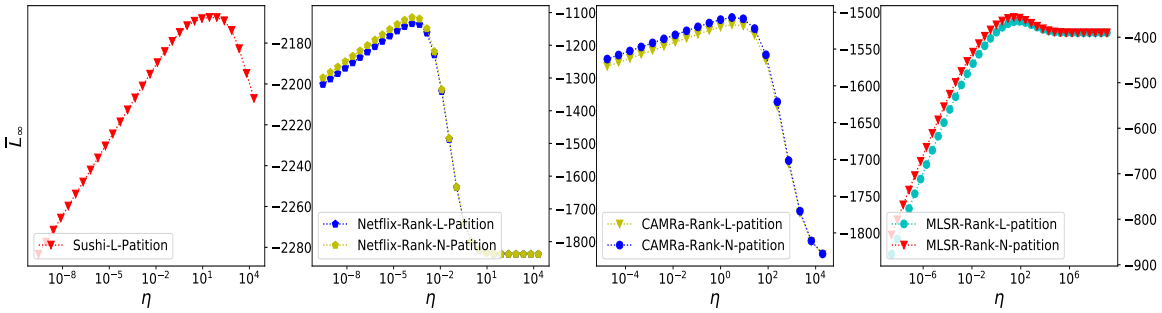


Figure 1: Variational lower bound ceiling for different datasets

the ration of the (ground truth) value of the estimated top-K set, over the (ground truth) value of the (true) optimal set.

6. Experiment Results

Variational Lower Bound Ceiling. For each dataset, we plot the mean variational lower bound ceiling $\overline{\mathbf{L}}_\infty$ across different experts or different splits as a function of the hyperparameter η in Fig. 1. This lower bound can be used to determine hyperparameter η, *without cross-validation* (i.e., without observing the test-set labels): this is one of the advantages of using variational inference. We denote by η_L the value of η that maximizes $\overline{\mathbf{L}}_\infty$. We observe that η_L is similar for $\mathcal{N}$ and $\mathcal{L}$ partitions.[2]

Top-K Metric. Table 3 includes top-K metric for both M-VI (variational inference) and MAP estimation. In particular, we show performance for both algorithms with hyperparameter η determined by cross validation (η_V^* and η_M^* respectively) and as well as for η_L determined by maximizing the lower bound as in Fig. 1. Note that MAP estimation for η_L is only provided for comparison purposes, as variational inference is needed to compute η_L. We provide results under both $\mathcal{L}$ and $\mathcal{N}$ partitions. Comparing M-VI and MAP under $\mathcal{L}$ partition and cross-validated η, we see that, with the exception of the MLSR dataset, M-VI is better than MAP. Using η_L rather than the parameter determined by cross validation slightly decreases performance. On the same vein, the $\mathcal{N}$ partition setting is slightly more difficult than $\mathcal{L}$ partition setting; performance is again slightly lower.

Fig. 2 shows the full effect of η for the CAMRa-Rank dataset. Overall, M-VI is better than MAP, while η_L acquired by Fig. 1 is close to (but not exactly equal to) the η that maximizes the top-K value across cross validation.

AUC Metric. Table 4 shows AUC metric performance for both M-VI and MAP estimation; we note that AUC is a metric towards which MAP estimation is best suited to, so we expect MAP to perform well. Again, the table shows AUC performance for both algorithms with hyperparameter η determined by cross validation (η_V^* and η_M^* respectively) and as well as for η_L determined by maximizing the lower bound as in Fig. 1; we again also provide results under both $\mathcal{L}$ and $\mathcal{N}$ partitions. Comparing M-VI and MAP under $\mathcal{L}$ partition and $\mathcal{N}$ partition we see that, except for MLSR-Com and Netflix-Rank datasets, M-VI is almost identical as MAP. Using η_L rather than the parameter determined by cross validation slightly decreases performance. Again, the $\mathcal{N}$ partition setting is slightly more difficult than $\mathcal{L}$ partition setting except for Netflix-Rank and CAMRa-Rank dataset.

2. Our code is publicly available at: https://github.com/neu-spiral/VariationalPlackettLuce

		Size(K)	M-VI (η_V^*)	MAP (η_M^*)	M-VI (η_L)	MAP (η_L)
L Partition	Camra	15	0.967	0.961	0.966	0.961
		20	0.958	0.952	0.957	0.952
	Netflix	15	0.993	0.993	0.991	0.992
		20	0.993	0.993	0.993	0.993
	Sushi	15	0.805	0.805	0.79	0.803
		20	0.824	0.824	0.809	0.823
	MLSR	15	0.936	0.937	0.926	0.928
		20	0.922	0.924	0.911	0.916
N Partition	Camra	15	0.839	0.838	0.839	0.837
		20	0.821	0.821	0.82	0.819
	Netflix	15	0.914	0.91	0.914	0.908
		20	0.903	0.903	0.901	0.903
	MLSR	15	0.607	0.611	0.529	0.521
		20	0.602	0.603	0.538	0.528

CAMRa: CAMRa-Rank; Netflix: Netflix-Rank; MLSR: MLSR-Rank.

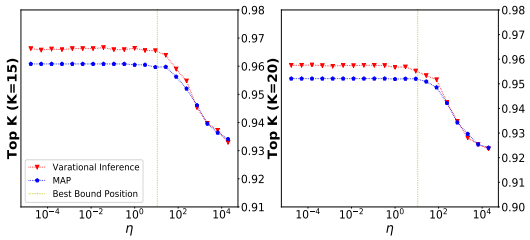
(a) Top-K metric for ranking datasets

		Size(K)	M-VI (η_V^*)	MAP (η_M^*)	M-VI (η_L)	MAP (η_L)
L Partition	Camra	15	0.899	0.898	0.899	0.897
		20	0.89	0.89	0.888	0.887
	Netflix	15	0.972	0.971	0.971	0.97
		20	0.967	0.967	0.967	0.967
	ROP	15	0.925	0.924	0.922	0.919
		20	0.925	0.924	0.919	0.918
	MLSR	15	0.86	0.853	0.828	0.83
		20	0.85	0.841	0.817	0.817
N Partition	Camra	15	0.841	0.84	0.837	0.837
		20	0.836	0.836	0.835	0.836
	Netflix	15	0.932	0.931	0.929	0.929
		20	0.928	0.928	0.925	0.925
	ROP	15	0.899	0.899	0.899	0.898
		20	0.914	0.91	0.914	0.909
	MLSR	15	0.641	0.641	0.628	0.627
		20	0.636	0.634	0.628	0.625

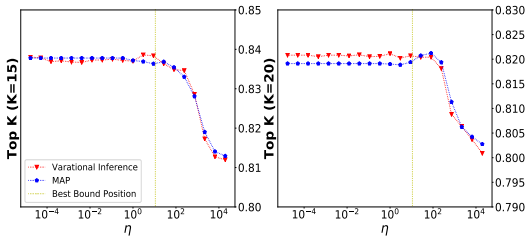
CAMRa: CAMRa-Com; Netflix: Netflix-Com; MLSR: MLSR-Com.

(b) Top-K metric for comparison datasets

Table 3: Top-K metric for ranking and comparison datasets.



(a) CAMRa-Rank (L partition)



(b) CAMRa-Rank (N partition)

Figure 2: The top-K metric result: left two figures are CAMRa-Rank dataset for $\mathcal{L}$ partition and right two figures are CAMRa-Rank dataset for $\mathcal{N}$ partition. Red curve: variational inference; blue curve: MAP; best bound position: the η_L that can maximize the ceiling of variational lower bound.

The reason is that, in these datasets, under the $\mathcal{N}$ partition, the total number of pairwise comparisons in the test set is much smaller than the number under the $\mathcal{L}$ partition.

Posterior Distribution. We also conduct additional experiments to assess the quality of the inferred distribution. Recall that we reduce rankings in the test set to comparison pairs. As our variational inference method gives us a posterior Gaussian distribution $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S})$, based on the Bradley Terry model, the score difference $s_{i,j} = s_i - s_j$, that governs a pairwise outcome, also has a Gaussian distribution $s_{i,j} \sim \mathcal{N}(\mu, \varrho)$, where $\mu = \boldsymbol{x}_{i,j}^T \boldsymbol{\mu}$, $\varrho = \sqrt{\boldsymbol{x}_{i,j}^T \boldsymbol{S} \boldsymbol{x}_{i,j}}$. Intuitively, high μ should indicate high probability of $i \succ j$, and low μ should indicate high probability of $i \prec j$. In contrast, high variance, as captured by ϱ, should ameliorate this effect. In Fig. 4, we bin comparisons in the test set with respect to values of mean μ and ϱ, and plot the positive ratio (PR) capturing the number of pairs in a bin for which $i \succ j$, over the total number of pairs in the bin. Behavior is exactly as expected: there is an increase of PR as μ increases overall; however, higher variance ϱ worsens the accuracy of this prediction: PR tends towards 0.5 as ϱ increases in both the high and low μ regime.

	Name	M-VI (η_V^*)	MAP (η_M^*)	M-VI (η_L)	MAP (η_L)
L partition	CAMRa	0.731	0.732	0.731	0.732
	Netflix	0.872	0.877	0.871	0.876
	Sushi	0.591	0.592	0.579	0.589
	MLSR	0.723	0.728	0.698	0.702
N partition	CAMRa	0.821	0.821	0.821	0.821
	Netflix	0.902	0.902	0.901	0.901
	MLSR	0.64	0.64	0.592	0.586

CAMRa: CAMRa-Rank; Netflix: Netflix-Rank; MLSR: MLSR-Rank.

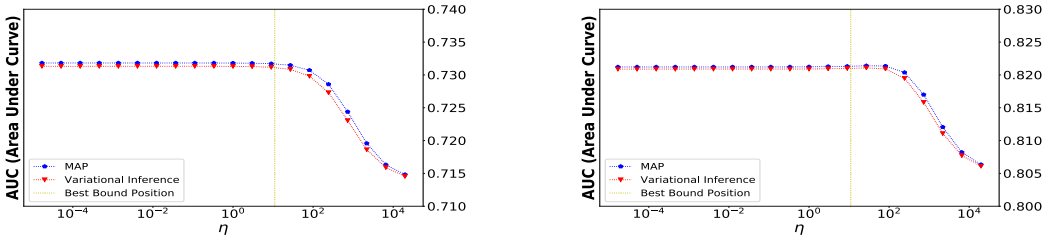
(a) AUC for ranking dataset

	Name	M-VI (η_V^*)	MAP (η_M^*)	M-VI (η_L)	MAP (η_L)
L partition	CAMRa	0.822	0.822	0.822	0.822
	Netflix	0.884	0.884	0.884	0.884
	ROP	0.87	0.875	0.865	0.865
	MLSR	0.789	0.779	0.738	0.738
N partition	CAMRa	0.813	0.812	0.813	0.812
	Netflix	0.867	0.867	0.867	0.867
	ROP	0.835	0.832	0.835	0.832
	MLSR	0.635	0.635	0.629	0.628

CAMRa: CAMRa-Com; Netflix: Netflix-Com; MLSR: MLSR-Com.

(a) AUC for comparison dataset

Table 4: AUC metric for ranking and comparison datasets.



(a) CAMRa-Rank (L partition)

(b) CAMRa-Rank (N partition)

Figure 3: The AUC result: left figure is CAMRa-Rank dataset for $\mathcal{L}$ partition and right figure is CAMRa-Rank dataset for $\mathcal{N}$ partition. Red curve: variational inference; blue curve: MAP; best bound position: the η_L that can maximize the ceiling of variational lower bound.

For CAMRa-Rank and Netflix-Rank dataset, as we have many experts, so we check the positive ratio for a specific mean interval described in Sec. 5.2. We find that the expert with low variance can have a more stable prediction result. For a positive range, if the standard deviation is small, the positive ratio would be higher.

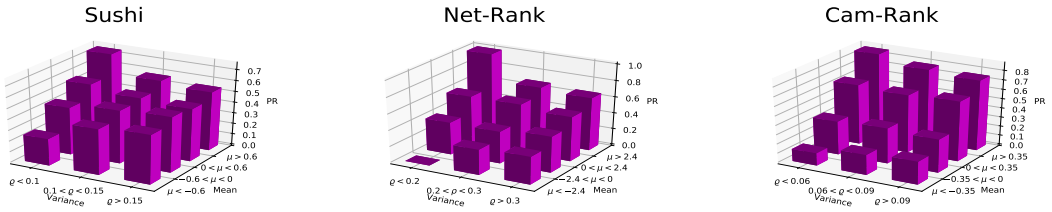
In Fig. 5, we plot the positive ratio (PR) for each expert as a function of the inferred standard deviation of predictions for this expert, for two multi-expert datasets. We again observe a downwards trend, with PR decreasing as the inferred standard-deviation of experts increases.

7. Conclusion

Variational inference can be used to learn posterior distributions under the Plackett-Luce model in both the ranking and top-choice settings. We have demonstrated the suitability of this approach for tasks beyond prediction; an additional use of a posterior would be to perform, e.g., active learning or experimental design, as in, e.g., Guo et al. (2018) and Guo et al. (2019); investigating this is an interesting future direction.

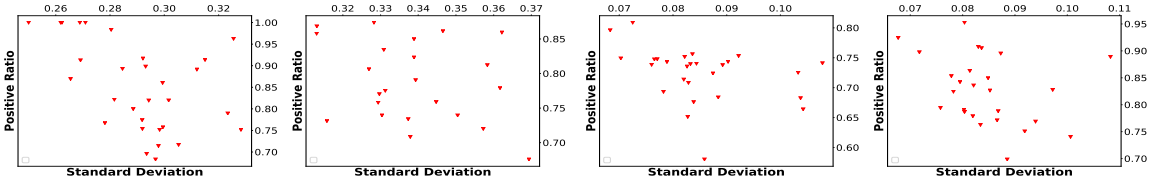
Acknowledgement

Our work is supported by NIH (R01EY019474, P30EY10572, K12EY27720), NSF (SCH-1622542 at MGH; SCH-1622536 and CCF-1750539 at Northeastern; SCH-1622679 at OHSU), and by unrestricted departmental funding from Research to Prevent Blindness (OHSU) .



(a) Sushi (b) Netflix-Rank L Partition (c) CAMRa-Rank L Partition

Figure 4: The positive ratio histogram for one expert: the left one is for Sushi dataset, the middle one is L partition for Netflix-Rank dataset, the right one is L partition for Netflix-Rank dataset.



(a) Netflix-Rank L (b) Netflix-Rank N (c) CAMRa-Rank L (d) CAMRa-Rank N

Figure 5: The positive ratio among different experts with different variance: the left one is L partition for CAMRa-Rank dataset, the right one is N partition for CAMRa-Rank dataset.

References

A. Agarwal, P. Patil, and S. Agarwal. Accelerated spectral ranking. In *ICML*, 2018.

N. Ahmed and M. Campbell. Variational bayesian data fusion of multi-class discrete observations with applications to cooperative human-robot estimation. In *ICRA*, 2010.

P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of gibbs posteriors. *JMLR*, 2016.

J. Bento, N. Fawaz, A. Montanari, and S. Ioannidis. Identifying users from their rating patterns. In *CAMRa*, 2011.

C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

G. Bouchard. Efficient bounds for the softmax function, applications to inference in hybrid models. In *Presentation at the Workshop at NIPS*, 2007.

R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 1952.

A. Brun, A. Hamad, O. Buffet, and A. Boyer. Towards preference relations in recommender systems. In *ECML/PKDD*, 2010.

F. Caron and A. Doucet. Efficient bayesian inference for generalized bradley terry models. *JCGS*, 2012.

F. Caron and Y. W. Teh. Bayesian nonparametric models for ranked data. In *NIPS*, 2012.

F. Caron, Y. W. Teh, T. B. Murphy, et al. Bayesian nonparametric plackett–luce models for the analysis of preferences for college degree programmes. *AOAS*, 2014.

W. Cheng, E. Hüllermeier, and K. J. Dembczynski. Label ranking methods based on the plackett-luce model. In *ICML*, 2010.

M. S. Desarkar, S. Sarkar, and P. Mitra. Aggregating preference graphs for collaborative rating prediction. In *Recsys*, 2010.

I. C. Gormley and T. B. Murphy. Exploring voting blocs within the irish electorate: A mixture modeling approach. *JASA*, 2008.

J. Guiver and E. Snelson. Bayesian inference for plackett-luce ranking models. In *ICML*, 2009.

Y. Guo, P. Tian, J. K-Cramer, S. Ostmo, J. P. Campbell, M. F. Chiang, D. Erdogmus, J. G Dy, and S. Ioannidis. Experimental design under the bradley-terry model. In *IJCAI*, 2018.

Y. Guo, J. Dy, D. Erdogmus, J. Kalpathy-Cramer, S. Ostmo, J. P. Campbell, M. F. Chiang, and S. Ioannidis. Accelerated experimental design for pairwise comparisons. In *SDM*, 2019.

D. R. Hunter et al. Mm algorithms for generalized bradley-terry models. *Ann. Stat*, 2004.

T. Jaakkola and M. Jordan. A variational approach to bayesian logistic regression models and their extensions. In *Sixth International Workshop on Artificial Intelligence and Statistics*, 1997.

J. K-Cramer, J.P. Campbell, D. Erdogmus, P. Tian, D. Kedarisetti, C. Moleta, J.D. Reynolds, K. Hutcheson, M.J. Shapiro, M.X. Repka, et al. Plus disease in retinopathy of prematurity: improving diagnosis by ranking disease severity and using quantitative image analysis. *Ophthalmology*, 2016.

T. Kamishima and S. Akaho. Efficient clustering for orders. In *Mining complex data*. Springer, 2009.

T. Kamishima, M. Hamasaki, and S. Akaho. Trbagg: A simple transfer learning method and its application to personalization in collaborative tagging. In *ICDM*, 2009.

M. Khan and W. Lin. Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. In *AISTATS*, 2017.

M. Khan, S. Mohamed, B. Marlin, and K. Murphy. A stick-breaking likelihood for categorical data analysis with latent gaussian models. In *AISTATS*, 2012.

A. Khetan and S. Oh. Data-driven rank breaking for efficient rank aggregation. *JMLR*, 2016.

Y. Koren and J. Sill. Ordrec: an ordinal model for predicting personalized item rating distributions. In *Recsys*, 2011.

Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 2009.

R. Kumar, A. Tomkins, S. Vassilvitskii, and E. Vee. Inverting a steady-state. In *WSDM*, 2015.

R. D. Luce. Individual choice behavior. 1959.

L. Maystre and M. Grossglauser. Fast and accurate inference of plackett–luce models. In *NIPS*, 2015.

K. P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

S. Negahban, S. Oh, and D. Shah. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 2016.

S. Negahban, S. Oh, K. K. Thekumparampil, and J. Xu. Learning from comparisons and choices. *JMLR*, 2018.

S. Park and S. Choi. Bayesian aggregation of binary classifiers. In *ICDM*, 2010.

R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society*, 1975.

T. Qin and Tie-Yan Liu. Introducing LETOR 4.0 datasets. *CoRR*, abs/1306.2597, 2013. URL http://arxiv.org/abs/1306.2597.

S. Ragain, A. Peysakhovich, and J. Ugander. Improving pairwise comparison models using empirical bayes shrinkage. *arXiv preprint arXiv:1807.09236*, 2018.

A. Rajkumar and S. Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *ICML*, 2014.

D. Sculley. Combined regression and ranking. In *SIGKDD*, 2010.

N. Stewart, G. D. Brown, and N. Chater. Absolute identification by relative judgment. *Psy. Review*, 2005.

W. T. Sun, T. H. Chao, Y. H. Kuo, and W. H. Hsu. Photo filter recommendation by category-aware aesthetic learning. *IEEE T MULTIMEDIA*, 2017.

P. Tian, Y. Guo, J. K-Cramer, S. Ostmo, J. P. Campbell, M. F. Chiang, D. Erdogmus, J. Dy, and S. Ioannidis. A severity score for retinopathy of prematurity. In *KDD*. ACM, 2019.

Y. S. Wang, R. L. Matsueda, E. A. Erosheva, et al. A variational em method for mixed membership models with multivariate rank data: An analysis of public policy preferences. *AOAS*, 2017.

İ. Yıldız, P. Tian, J. Dy, D. Erdoğmuş, J. Brown, J. K-Cramer, S. Ostmo, J. P. Campbell, M. F. Chiang, and S. Ioannidis. Classification and comparison via neural networks. *Neural Networks*, 2019.

Y. Zheng, L. Zhang, X. Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. In *WWW*. ACM, 2009.

Appendix A. Proof of Lemma 1

We first use the standard quadratic bound (Jaakkola and Jordan, 1997; Bishop, 2006): for any $y \in \mathbb{R}$ and $\xi \in \mathbb{R}$,

$$\log(1+e^{-y}) \leqslant \lambda(\xi)(y^2-\xi^2) + (-y+\xi)/2 + \log(1+e^{-\xi}), \tag{29}$$

which implies:

$$\sigma(y) \geqslant \sigma(\xi) e^{(y-\xi)/2 - \lambda(\xi)(y^2-\xi^2)}. \tag{30}$$

When $A = \{c_l, \bar{c}_l\}$, we have:

$$\begin{aligned} p(c=c_l|A) &\overset{(3)}{=} \frac{s_{c_l}}{s_{c_l}+s_{\bar{c}_l}} \overset{(9)}{=} \frac{1}{1+e^{-\boldsymbol{\theta}^T(\boldsymbol{x}_{c_l}-\boldsymbol{x}_{\bar{c}_l})}} = \sigma(\boldsymbol{x}_{c_l,\bar{c}_l}^T\boldsymbol{\theta}) \\ &\overset{(30)}{\geqslant} \sigma(\xi)\exp\Big((\boldsymbol{x}_{c_l,\bar{c}_l}^T\boldsymbol{\theta}-\xi)/2 - \lambda(\xi)((\boldsymbol{x}_{c_l,\bar{c}_l}^T\boldsymbol{\theta})^2-\xi^2)\Big). \end{aligned} \tag{31}$$

To prove Eq. (13), we use Bouchard's inequality (Bouchard, 2007), which states that for all $\boldsymbol{y} = [y_i]_i \in \mathbb{R}^K$ and all $\alpha \in \mathbb{R}$:

$$\log(\sum_{k=1}^K e^{y_k}) \leqslant \alpha + \sum_{k=1}^K \log(1+e^{y_k-\alpha}). \tag{32}$$

Combining Eq. (32) with Eq. (29), for every $\boldsymbol{\xi} = [\xi_i]_{i=1,\cdots,K} \in \mathbb{R}_+^K$ we get

$$\sum_{k=1}^K e^{y_k} \leqslant e^{\alpha} \prod_{k=1}^K \Big((1+e^{-\xi_k}) e^{(y_k-\alpha+\xi_k)/2 + \lambda(\xi_k)((y_k-\alpha)^2-\xi_k^2)}\Big). \tag{33}$$

Hence, the top query probability under the Plackett Luce model satisfies:

$$\begin{aligned} p(c=c_l|A) &= \frac{\exp(\boldsymbol{x}_{c_l}^T\boldsymbol{\theta})}{\sum_{j\in A_l}\exp(\boldsymbol{x}_j^T\boldsymbol{\theta})} \\ &\overset{(3)(33)}{\geqslant} \frac{\exp(\boldsymbol{x}_{c_l}^T\boldsymbol{\theta})}{e^{\alpha}\prod_{j\in A}\Big((1+e^{-\xi_j})e^{(\boldsymbol{x}_j^T\boldsymbol{\theta}-\alpha+\xi_j)/2+\lambda(\xi_j)((\boldsymbol{x}_j^T\boldsymbol{\theta}-\alpha)^2-\xi_j^2)}\Big)} \\ &= \exp(\boldsymbol{x}_{c_l}^T\boldsymbol{\theta}-\alpha)\prod_{j\in A}\Big(\sigma(\xi_j)e^{(-\boldsymbol{x}_j^T\boldsymbol{\theta}+\alpha-\xi_j)/2-\lambda(\xi_j)((\boldsymbol{x}_j^T\boldsymbol{\theta}-\alpha)^2-\xi_j^2)}\Big) \end{aligned} \tag{34}$$

□

Appendix B. Proof of Lemma 2

The posterior $q(\boldsymbol{\theta})$ has the Gaussian form, i.e.:

$$q(\boldsymbol{\theta}) = \frac{1}{B_q} e^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^T\boldsymbol{S}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})} \tag{35}$$

where $B_q = (2\pi)^{d/2}|\boldsymbol{S}|^{1/2}$, $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{S} \in \mathbb{R}^{d\times d}$ and the ELBO $\mathbf{L}(q)$ satisfies:

$$\begin{aligned}\mathbf{L}(q) \overset{(11)}{=}& \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log \frac{p_0(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right] + \mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{l\in\mathcal{L}} \log \frac{\exp(\boldsymbol{\theta}^T \boldsymbol{x}_{c_l})}{\sum_{j\in A_l}\exp(\boldsymbol{\theta}^T\boldsymbol{x}_j)}\right] \\ =& \mathbb{E}_{q(\boldsymbol{\theta})}\left[\frac{(\boldsymbol{\theta}-\boldsymbol{\mu})^T\boldsymbol{S}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})}{2} - \frac{(\boldsymbol{\theta}-\boldsymbol{\mu}_0)^T\boldsymbol{S}_0^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_0)}{2}\right] + \frac{1}{2}\log\frac{|\boldsymbol{S}|}{|\boldsymbol{S}_0|} \\ &+ \mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{l\in\mathcal{L}} \log \frac{\exp(\boldsymbol{\theta}^T \boldsymbol{x}_{c_l})}{\sum_{j\in A_l}\exp(\boldsymbol{\theta}^T\boldsymbol{x}_j)}\right].\end{aligned} \tag{36}$$

By Lemma 1, $\exp(\boldsymbol{\theta}^T\boldsymbol{x}_{c_l})/\sum_{j\in A_l}\exp(\boldsymbol{\theta}^T\boldsymbol{x}_j)$ is bounded by Q_l, and Lemma 2 follows. □

Appendix C. Proof of Lemma 3

The variational lower bound (14) is an expectation of a quadratic function of $\boldsymbol{\theta}$, and we can obtain the corresponding variational parameters $\boldsymbol{S}, \boldsymbol{\mu}$ by identifying the linear and quadratic terms in $\boldsymbol{\theta}$. To maximize the lower bound in Eq. (14), the quadratic term satisfies:

$$\boldsymbol{\theta}^T\boldsymbol{S}^{-1}\boldsymbol{\theta} = \boldsymbol{\theta}^T\boldsymbol{S}_0^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}^T\Big(2\sum_{l\in\mathcal{L}_2}\lambda(\zeta_l^{(k)})\boldsymbol{x}_{c_l,\bar{c}_l}\boldsymbol{x}_{c_l,\bar{c}_l}^T + 2\sum_{l\in\mathcal{L}_{>2}}\sum_{j\in A_l}\lambda(\xi_{lj}^{(k)})\boldsymbol{x}_j\boldsymbol{x}_j^T\Big)\boldsymbol{\theta}, \tag{37}$$

and the linear term satisfies:

$$\boldsymbol{\theta}^T\boldsymbol{S}^{-1}\boldsymbol{\mu} = \boldsymbol{\theta}^T\boldsymbol{S}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\theta}^T\left(\sum_{l\in\mathcal{L}_2}\frac{\boldsymbol{x}_{c_l,\bar{c}_l}}{2} + \sum_{l\in\mathcal{L}_{>2}}\left(\boldsymbol{x}_{c_l} + \sum_{j\in A_l}\big(2\lambda(\xi_{lj}^{(k)})\alpha_l^{(k)}\boldsymbol{x}_j - \frac{\boldsymbol{x}_j}{2}\big)\right)\right). \tag{38}$$

As the Eq. (37) and Eq. (38) hold for any $\boldsymbol{\theta}$, Lemma 3 follows.

□

Appendix D. Proof of Lemma 4

From (14), objective of (16b) is separable w.r.t. $\boldsymbol{\zeta}$ and $\boldsymbol{\xi}, \boldsymbol{\alpha}$, and

$$\boldsymbol{\zeta}_l^{(k+1)} = \underset{\zeta_l}{\mathbf{argmax}}\ \mathbb{E}_q[Q_l], \quad l \in \mathcal{L}_2, \tag{39}$$

where Q_l is given by (15). Hence, the optimal ζ_l is a stationary point:

$$\frac{\partial\mathbf{L}(\boldsymbol{\zeta},\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\mu}^{(k)},\boldsymbol{S}^{(k)})}{\partial\zeta_l} = -\lambda'(\zeta_l)\left(\mathbb{E}_{q(\boldsymbol{\theta})}\left[(\boldsymbol{\theta}^T\boldsymbol{x}_{c_l,\bar{c}_l})^2\right] - \zeta_l^2\right) = 0,\ l \in \mathcal{L}_2, \tag{40}$$

As $\lambda'(\zeta_l) > 0$, stationary points satisfy $\mathbb{E}_{q(\boldsymbol{\theta})}\left[(\boldsymbol{\theta}^T\boldsymbol{x}_{c_l,\bar{c}_l})^2\right] = \zeta_l^2$, which yields (19) for $\zeta_l \geq 0$. □

Appendix E. Proof of Lemma 5

By (15), we have that:

$$f_l(\boldsymbol{\xi}_l, \alpha_l) = -\alpha_l + \sum_{j\in A_l}\Big(\log\sigma(\xi_{lj}) + \frac{\alpha_l - \xi_{lj}}{2} - \lambda(\xi_{lj})\big(\mathbb{E}_{q(\boldsymbol{\theta})}[(\boldsymbol{x}_j^T\boldsymbol{\theta}-\alpha_l)^2] - \xi_{lj}^2\big)\Big). \tag{41}$$

This is a quadratic function with respect to α_l. The solution for Eq. (21a) is a stationary point, so we have:

$$\frac{\partial f_l}{\partial \alpha_l} = \frac{(m_l - 2)}{2} - 2\sum_{j \in A_l} \lambda(\xi_{lj}^{(n)})\alpha_l + 2\sum_{j \in A_l} \lambda(\xi_{lj}) \mathbb{E}_{q(\boldsymbol{\theta})}[\boldsymbol{x}_j^T \boldsymbol{\theta}] = 0,\ l \in \mathcal{L}_{>2}, \tag{42}$$

which implies Eq. (22). For $\boldsymbol{\xi}_l$, the solution for Eq. (21b) should also be a stationary point:

$$\frac{\partial f_l}{\partial \xi_{lj}} = -\lambda'(\xi_{lj}) \left(\mathbb{E}_{q(\boldsymbol{\theta})} \left[(\boldsymbol{\theta}^T \boldsymbol{x}_j - \alpha_l^{(n+1)})^2 \right] - \xi_{lj}^2 \right) = 0,\ l \in \mathcal{L}_{>2}, j \in A_l, \tag{43}$$

As $\lambda'(\xi_{lj}) > 0$, stationary points satisfy $\mathbb{E}_{q(\boldsymbol{\theta})} \left[(\boldsymbol{\theta}^T \boldsymbol{x}_{lj} - \alpha_l^{(n+1)})^2 \right] = \xi_{lj}^2$, which implies (23) for $\xi_{lj} \geq 0$. □

Appendix F. Proof of Lemma 6

The derivation follows the same argument as in Jaakkola and Jordan (1997). Briefly, to see why Eq. (25) holds, note that in step (16a), covariance $\boldsymbol{S}$ and mean $\boldsymbol{\mu}$ are updated. Subsequently, $\mathbf{L}$ defined as in Eq. (14) is the sum of two terms: the KL divergence between two Gaussian distributions and a normalization factor. The optimum in step (16a) therefore occurs when the two Gaussian distributions are identical. Because of this, we can omit all quadratic and linear terms from (14) when computing (25): we only need to calculate the normalization factor by adding the constant items in Eq. (14).

In more details, quantity Q_l defined in Eq. (15) can be written as:

$$Q_l = Q_l^{'} + Q_l^{''} + \overline{Q}_l \tag{44}$$

where $Q_l^{'}, Q_l^{''}$ and $\overline{Q}_l$ are defined as:

$$Q_l^{'} = \begin{cases} \boldsymbol{x}_{c_l}^T \boldsymbol{\theta} + \sum\limits_{j \in A_l} \Big[-\boldsymbol{x}_j^T \boldsymbol{\theta}/2 + 2\alpha_l \lambda(\xi_{lj}) \big(\boldsymbol{x}_j^T \boldsymbol{\theta} \big) \Big], & l \in \mathcal{L}_{>2}, \\ \boldsymbol{x}_{c_l, \bar{c}_l}^T \boldsymbol{\theta}/2, & l \in \mathcal{L}_2, \end{cases} \tag{45a}$$

$$Q_l^{''} = \begin{cases} -\sum\limits_{j \in A_l} \lambda(\xi_{lj}) \big(\boldsymbol{x}_j^T \boldsymbol{\theta} \big)^2, & l \in \mathcal{L}_{>2}, \\ -\lambda(\zeta_l) (\boldsymbol{x}_{c_l, \bar{c}_l}^T \boldsymbol{\theta})^2, & l \in \mathcal{L}_2, \end{cases} \tag{45b}$$

$$\overline{Q}_l = \begin{cases} -\alpha_l + \sum\limits_{j \in A_l} \Big[\log \sigma(\xi_{lj}) + \frac{\alpha_l - \xi_{lj}}{2} - \lambda(\xi_{lj}) \big(\alpha_l^2 - \xi_{lj}^2 \big) \Big], & l \in \mathcal{L}_{>2}, \\ \log \sigma(\zeta_l) - \zeta_l/2 + \lambda(\zeta_l) \zeta_l^2, & l \in \mathcal{L}_2. \end{cases} \tag{45c}$$

After step (16a), the following equations hold:

$$\mathbb{E}_{q(\boldsymbol{\theta})} \left[\sum_{l \in \mathcal{L}} Q_l^{'} \right] + \mathbb{E}_{q(\boldsymbol{\theta})} \Big[-\boldsymbol{\theta}^T \boldsymbol{S}^{-1} \boldsymbol{\mu} + \boldsymbol{\theta}^T \boldsymbol{S}_0^{-1} \boldsymbol{\mu}_0 \Big] = 0, \tag{46a}$$

$$\mathbb{E}_{q(\boldsymbol{\theta})} \left[\sum_{l \in \mathcal{L}} Q_l^{''} \right] + \mathbb{E}_{q(\boldsymbol{\theta})} \Big[\boldsymbol{\theta}^T \boldsymbol{S}^{-1} \boldsymbol{\theta}/2 - \boldsymbol{\theta}^T \boldsymbol{S}_0^{-1} \boldsymbol{\theta}_0/2 \Big] = 0, \tag{46b}$$

which are equivalent to Eq. (38) and Eq. (37). Thus, Eq. (14) can be written as:

$$\begin{aligned}\mathbf{L}(\boldsymbol{\zeta}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{S}) =& \mathbb{E}_{q(\boldsymbol{\theta})}\left[\sum_{l\in\mathcal{L}} \overline{Q}_l\right] + \frac{1}{2}\log\frac{|\boldsymbol{S}|}{|\boldsymbol{S}_0|} + \mathbb{E}_{q(\boldsymbol{\theta})}\left[\frac{\boldsymbol{\mu}^T\boldsymbol{S}^{-1}\boldsymbol{\mu}}{2} - \frac{\boldsymbol{\mu}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{\mu}_0}{2}\right] \\ =& \sum_{l\in\mathcal{L}} \overline{Q}_l + \frac{1}{2}\log\frac{|\boldsymbol{S}|}{|\boldsymbol{S}_0|} + \frac{\boldsymbol{\mu}^T\boldsymbol{S}^{-1}\boldsymbol{\mu}}{2} - \frac{\boldsymbol{\mu}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{\mu}_0}{2}.\end{aligned} \tag{47}$$

This is equivalent to Eq. (25). □