Noname manuscript No.
(will be inserted by the editor)

Efficiently Learning the Preferences of People

Adriana Birlutiu · Perry Groot · Tom Heskes

Received: date / Accepted: date

Abstract This paper presents a framework for optimizing the preference learning process. In many real-world applications in which preference learning is involved the available training data is scarce and obtaining labeled training data is expensive. However, in most of the preference learning situations data is available from multiple subjects. We use the multi-task formalism to increase the individual training data by making use of the preference information learned from other subjects. Furthermore, since obtaining labels is expensive, we optimally choose which data to ask a subject for labelling such that to obtain maximum of information about his/her preferences. This paradigm — called active learning— has hardly been studied in a multi-task formalism. We propose an alternative for the standard criteria in active learning which actively chooses queries by making use of the available preference data from other subjects. The advantage of this alternative is the reduced compuation costs and time subjects are involved.

Keywords learning preferences, active learning, experimental design, multi-task learning, hierarchical modeling

A.Birlutiu
Institute for Computing and Information Sciences (iCIS)
Radboud University Nijmegen
Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands
Tel.: +31 24 3652354
E-mail: A.Birlutiu@science.ru.nl

P.Groot
Institute for Computing and Information Sciences (iCIS)
Radboud University Nijmegen
Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands
Tel.: +31 24 3652354
E-mail: Perry.Groot@science.ru.nl

T.Heskes
Institute for Computing and Information Sciences (iCIS)
Radboud University Nijmegen
Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands
Tel.: +31 24 3652696
E-mail: T.Heskes@science.ru.nl

1 Introduction

Recently, there has been a wide interest in learning the preferences of people within artificial intelligence research (Doyle, 2004). Preference learning is a crucial aspect in modern applications such as decision support systems (Chajewska et al, 2000), recommender systems (Blythe, 2002; Blei et al, 2003), and personalized devices (Clyde et al, 1993; Heskes and de Vries, 2005). The preferences of a person are learned in order to recommend/take decisions on his/her behalf. For example, a movie recommender system needs to learn your preferences about movies before making a recommendation about a movie you might prefer.

Since preference learning is a cumbersome process it is important to make it as efficient as possible in order to reduce the costs and time involved. In practice one has to minimize the time subjects are involved. The preferences should be learned accurately from a minimum number of training data.

In most of the situations in which preference learning is involved data is available from multiple subjects. Thus, even though individual data is scarce and difficult to obtain, we can optimize the learning of preferences of a new subject by making use of the available data from other subjects. Learning in this setting is well-known as multi-task or hierarchical learning and has been studied extensively in recent years in machine learning. By using the multi-task formalism, the preference data collected for other subjects can be gathered and used as a prior information when learning the preference of a new subject. Furthermore, to deal with the fact that obtaining labeled data is expensive we can speed up learning by optimally choosing the examples to be queried. At each learning step we can decide which example gives the most information about the subject's preferences. This paradigm, called active learning in the machine learning literature and related to sequential experimental design in statistics, has been studied extensively, but hardly in the multi-task setting.

The *aim of this article* is to introduce *a framework for optimizing the preference learning process*. This framework considers the *combination between active learning and multi-task learning in the preference learning context*. The *contribution of this paper* is *a criterion for active learning designed for the multi-task setting*.

1.1 Structure of this Article

The structure of this article is as follows.

In Section 2, we describe the learning framework. We consider learning from qualitative preference observations which can be modeled using the probabilistic choice models introduced in Section 2.1. Learning a utility function representing the preferences of a subject from this type of preference observations is described in Section 2.2. Learning the utility function in a multi-task setting by making use of the data available from other subjects is considered in Section 2.3.

In Section 3 we present several criteria for selecting the most informative experiments with respect to a subject's preferences. After reviewing some of the standard criteria from experimental design, we propose an alternative criterion which makes use of the preference observations collected already from a community of subjects. We show that this alternative criterion is connected to the standard criteria from experimental design. Furthermore, it has several advantages due to its interpretation and simplicity.

In Section 4 we demonstrate experimentally the usefulness of our framework on an audiological data set. The framework's efficiency is measured in terms of: *i)* the accuracy with which the preferences are learned; and *ii)* the number of preference observations needed in order to get a good model of a person's preferences.

In Section 5 we give conclusions.

1.2 Related Work

In this section we review some studies from preference learning, multi-task learning, and active learning related to the work presented in this paper.

1.2.1 Preference Learning

The existing approaches to preference learning can be divided into two categories. The first category aims at obtaining a ranking of instances from a set of pairwise preferences by solving an augmented binary classification problem (Herbrich et al, 1998; Har-peled et al, 2002; Fürnkranz and Hüllermeier, 2003; Aiolli and Sperduti, 2004). The second category uses regression to map instances to target valuations for direct ranking (Caruana et al, 1996; Crammer and Singer, 2001; Chu and Ghahramani, 2005b; Brochu et al, 2007). We focus on the latter and use a *utility function* in order to model the subject's preferences. The utility function is learned in a Bayesian framework. By formulating the preference elicitation problem as a probabilistic Bayesian learning problem, one can deal with inconsistencies in subjects responses as well as learn biases the subjects may have.

1.2.2 Multi-Task Learning

The basic idea in multi-task learning is that models learned for different tasks have parts in common. In a Bayesian framework this often boils down to the sharing of a hierarchical prior (Blei et al, 2003; Bakker and Heskes, 2003; Evgeniou et al, 2005; Xue et al, 2007). Learning preferences from qualitative preference statements with Gaussian processes in the context of multi-task learning has been considered in (Birlutiu et al, 2009).

1.2.3 Active Learning

Methods for active learning can be roughly divided into two categories: those with and without an explicitly defined objective function. Uncertainty sampling (Lewis and Gale, 1994), Query-by-Committee (Seung et al, 1992; Freund et al, 1997) and variants thereof belong to the latter category. They are based on the idea of selecting the most uncertain data given the previously trained models. The methods with an explicit objective function are often motivated by the theory of experimental design (Fedorov, 1972; Chaloner and Verdinelli, 1995; Schein and Ungar, 2007; Lewi et al, 2009; Dror and Steinberg, 2008). The objective function then quantifies the expected gain of labeling a particular input, for example in terms of the expected reduction in the entropy of the model parameters (Mackay, 1992; Cohn et al, 1996). The statistical methods perform better (Schein and Ungar, 2007), yet are computationally more demanding.

A sub-trend is to improve the performance of the active learning methods by combining them with heuristics designed either for the context in which they are applied or by the models they use, e.g., making use of the unlabeled data available (McCallum and Nigam, 1998; Yu et al, 2006), exploiting the clusters in the data (Dasgupta and Hsu, 2008), diversifying the set of hypotheses (Melville and Mooney, 2004), active learning for collaborative filtering (Jin and Si, 2004; Harpale and Yang, 2008; Boutilier et al, 2003), and active learning with Gaussian processes (Chu and Ghahramani, 2005a; Brochu et al, 2007).

In this paper we propose an alternative to the standard active learning criteria which makes use of the preference observations collected already from a community of subjects. This criterion, which we call the *Committee criterion*, is thus particularly designed for the multi-task setting that we consider in this paper. The idea behind the Committee criterion is related to the *Query-by-Committee* method from active learning which selects those queries that have maximum disagreement amongst an ensemble of hypotheses. The difference in our case is that the group of subjects, for which the preferences were already learned, plays the role of the ensemble of hypotheses.

1.3 Notation

Boldface notation is used for vectors and matrices and normal fonts for the components of vectors and matrices or scalars. Upperscripts are used to distinguish between different vectors or matrices and lowerscripts to address their components. The notation $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is used for a multivariate Gaussian with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. The transpose of a matrix $\boldsymbol{M}$ is denoted by $\boldsymbol{M}^T$. Capital letters are used for constants and small letters for indices, e.g., $i = 1, \dots, I$.

2 Learning Framework

The idea of using preference observations from other subjects in order to optimize the process of learning the preferences of a new subject can be basically applied in any preference learning context. In this paper, we restrict to qualitative preference observations which can be modeled using the probabilistic choice models described in this section.

2.1 Probabilistic Choice Models

Since people are very good in making comparisons, in many real-world applications preferences are learned from experiments in which the subject makes a choice for one of the presented alternatives. Let $X = \{x_1, \dots, x_I\}$ be a set of inputs. Let $\mathcal{D}$ be a set of J observed preference comparisons over instances in X corresponding to a subject,

$$\mathcal{D} = \{(a_j, c_j) | 1 \leq j \leq J, c_j \in \{1, \dots, A\}\} \tag{1}$$

with $a_j = (x_{i_1(j)}, \dots, x_{i_A(j)})$ the alternatives presented and c_j the choice made, $i_1, \dots, i_A : \{1, \dots, J\} \to \{1, \dots, I\}$ index functions such that $i_1(j)$ represents the first input presented in the jth preference comparison and $c_j = c$ means that $x_{i_c(j)}$ is chosen from the A alternatives presented in the jth comparison. For $A = 2$ this setup reduces to pairwise comparison between two alternatives.

The main idea behind probabilistic choice models is to assume a latent utility function value $U(x)$ associated with each input x which captures the individual preference of a subject for x. In the ideal case the latent function values are consistent with the preference observations, which in probabilistic terms can be written as $p(c_j = c|a_j, U) = 1$ $U(x_{i_c(j)}) > U(x_{i_{c'}(j)})$, $c' \neq c$, which means that alternative c is preferred over the other alternatives c' in the jth comparison whenever the utility for c exceeds the utilities for the other alternatives c'. In practice, however, subjects are often inconsistent in their responses. A very inconsistent subject will have a high uncertainty associated with the utility function; this uncertainty is directly taken into account in the probabilistic framework. We define this probabilistic framework by making a standard modeling assumption (Bradley and Terry, 1952; Kanninen, 2002; Glickman and Jensen, 2005) that the subject's decision in such a forced-choice comparison follows a multinomial logistic model, which is defined as

$$p(c_j = c|a_j, U) = \frac{\exp\left[U(x_{i_c(j)})\right]}{\sum_{c'=1}^{A} \exp\left[U(x_{i_{c'}(j)})\right]} \; . \tag{2}$$

Efficiently learning preferences reduces to learning the unknown utility function U as accurately and with as few comparisons as possible.

2.2 Learning the Utility Function

The utility function U is a real-valued function, $U : X \to \mathbb{R}$, which associates with every input $x \in X$ a real number $U(x)$. Each input $x \in X$ is characterized by a set of features, $\boldsymbol{\phi}(x) \in \mathbb{R}^T$. One possible choice for the utility function is to write it as a linear combination of the features,

$$U(x) = \sum_{i=1}^{T} \alpha_i \phi_i(x) \; , \tag{3}$$

where $\boldsymbol{\alpha}$ is a vector of weights specific to a subject which captures the importance of each feature of x when evaluating the utility U. The preferences are thus encoded in the vector $\boldsymbol{\alpha}$ and learning the utility function reduces to learning $\boldsymbol{\alpha}$.

To allow more flexibility for U we can use a semiparametric model where the utility function is defined as a linear combination of basis functions defined by a kernel κ centered on the data points,

$$U(x) = \sum_{i=1}^{I} \alpha_i \kappa(x, x_i) \; , \tag{4}$$

where the vector $\boldsymbol{\alpha}$ with dimension I—the number of data points—captures the preferences of the subject.

In order to learn the utility function from Equation (4) we treat the vector of parameters $\boldsymbol{\alpha}$ as a random variable. The same can be done when using the linear utility model from Equation (3). We consider a Gaussian prior distribution over $\boldsymbol{\alpha}$,

$p(\boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, which is updated based on the observations from the preference comparisons using Bayes' rule,

$$p(\boldsymbol{\alpha}|\mathcal{D}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto p(\boldsymbol{\alpha}) \prod_{j=1}^{J} p(c_j|a_j, \boldsymbol{\alpha}) \,, \tag{5}$$

with the likelihood terms of the form given in Equation (2). The posterior distribution obtained is approximated to a Gaussian. The Gaussian approximation of the posterior is a good approximation because with few data points the posterior is close to the prior which is a Gaussian, and with many data points the posterior approaches again a Gaussian as a consequence of the central limit theorem (Bishop, 2006). To perform the approximation of the posterior we use deterministic methods (e.g., Laplace's method (Mackay, 2002), Expectation Propagation (Minka, 2001)) since they are computationally cheaper than the non-deterministic ones (sampling) and because they are known to be accurate for these types of models (Glickman and Jensen, 2005).

2.3 Multi-task Preference Learning

One property which distinguishes the preference learning from other learning settings is that in most of the cases preference observations are available from multiple subjects. We make use of this property when learning the preferences of a new subject. We use Bayesian hierarchical modeling which assumes that the parameters for individual models are drawn from the same hierarchical prior distribution. Let us assume that we already have preference data available from a group of M subjects. We make the common assumption of a Gaussian prior distribution, $p(\boldsymbol{\alpha}^m) = \mathcal{N}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}),\ m = 1, \ldots, M$ with the same $\bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{\Sigma}}$ for the preference models of all subjects. This prior is updated using Bayes' rule based on the observations from each subject, resulting in a posterior distribution for each individual subject. The hierarchical prior is obtained by maximizing the log-likelihood of all data in a so-called type-II maximum likelihood approach. This optimization is performed by applying the EM algorithm (Gelman et al, 2003), which reduces to the iteration (until convergence) of the following two steps:

E-step: Estimate the sufficient statistics (mean $\boldsymbol{\mu}^m$ and covariance matrix $\boldsymbol{\Sigma}^m$) of the posterior distribution corresponding to each subject m, given the current estimates ($\bar{\boldsymbol{\mu}}^{(t)}$ and $\bar{\boldsymbol{\Sigma}}^{(t)}$) of the hierarchical prior.

M-step: Re-estimate the parameters of the hierarchical prior:

$$\bar{\boldsymbol{\mu}}^{(t+1)} = \frac{1}{M} \sum_{j=1}^{M} \boldsymbol{\mu}^m \,,$$

$$\bar{\boldsymbol{\Sigma}}^{(t+1)} = \frac{1}{M} \sum_{j=1}^{M} (\boldsymbol{\mu}^m - \bar{\boldsymbol{\mu}}^{(t+1)})(\boldsymbol{\mu}^m - \bar{\boldsymbol{\mu}}^{(t+1)})^T + \frac{1}{M} \sum_{j=1}^{M} \boldsymbol{\Sigma}^m \,.$$

Once we have learned the hierarchical prior, $\mathcal{N}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$, we can use it as an informative prior for the preference model of a new subject in Equation (5).

3 Active Preference Learning

Active learning, also known in the statistics literature as sequential experimental design, is suitable for the situations in which labeling points is difficult, time-consuming, and expensive. This is also the case in most of the preference learning settings in which the labels are given by people in an explicit way. The idea behind active learning is that by optimal selection of the training points a greater accuracy can be obtained than by random selection. The scenarios in which active learning can be applied belong to one of the following three categories: *i)* generating *de novo* points for labeling; *ii)* stream-based active learning where the learner decides whether to request the label of a given instance or not; *iii)* pool-based active learning where queries are selected from a large pool of unlabeled data. In this paper we consider the pool based active learning.

We assume that the current model of the preference data observed from a subject is a Gaussian distribution $\mathcal{M} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $p(c|a, \mathcal{M})$ be the probability of an observation given the data seen so far and let the new model obtained after incorporating an observation (a, c) be $\mathcal{M}_{(a,c)} \approx \mathcal{N}(\boldsymbol{\mu}_{(a,c)}, \boldsymbol{\Sigma}_{(a,c)})$. There are several strategies for active learning, all being concerned with evaluating the *informativeness* of the unlabeled points. In the following we briefly review these strategies and the way they can be implemented in the learning framework considered here.

1. **Uncertainty Sampling** (Lewis and Gale, 1994). In this strategy an active learner chooses for labeling the example for which the model's predictions are most uncertain. The uncertainty of the predictions can be measured, for example, using Shannon entropy
$$\text{Uncertainty}(a) = \sum_c p(c|a, \mathcal{M}) \log p(c|a, \mathcal{M}) \; . \tag{6}$$
For a binary classifier this strategy reduces to querying points whose prediction probabilities are close to 0.5. Intuitively this strategy aims at finding as fast as possible the decision boundary since this is indicated by the regions where the model is most uncertain.
2. **Query-by-Committee (QBC)** (Seung et al, 1992) is an effective active learning approach that has successfully been applied to many problems. In each iteration QBC *i)* constructs a committee of models based on the current training set, and *ii)* ranks the unlabeled examples according to some measure of disagreement among the committee members. The input with the highest disagreement score is then selected for labeling and added to the training data. We will discuss how to adapt QBC to our preference learning setting in Section 3.1.
3. **Variance Reduction** (Mackay, 1992). This strategy, also known in experimental design as D-optimality (Fedorov, 1972; Chaloner and Verdinelli, 1995; Berger, 1994; Ford and Silvey, 1980), chooses as the most informative experiments the ones that give the most reduction in the model's variance. The motivation behind this strategy is a result of (Geman et al, 1992) which shows that the generalization error can be decomposed into three components: *i)* noise (which is independent of the model or training data); *ii)* bias (due to the model); *iii)* model's variance. Since the model cannot influence the noise and the bias components, the future generalization error can only be influenced via the model's variance. Formally, this criterion can be written as
$$\text{Variance}(a) = \sum_c p(c|a, \mathcal{M}) \text{variance}[\mathcal{M}_{(a,c)}] \; . \tag{7}$$

In the setting considered in this paper the model is a Gaussian distribution, thus the variance of the model is expressed in the covariance matrix. In order to use Equation (7) we need to choose a measure for the variance. We can consider, for example, the determinant of the covariance matrix

$$\text{Variance-logdet}(a) = \sum_c p(c|a, \mathcal{M}) \log \det(\boldsymbol{\Sigma}_{(a,c)}) \ , \tag{8}$$

or the trace of the covariance matrix

$$\text{Variance-trace}(a) = \sum_c p(c|a, \mathcal{M}) \operatorname{Tr}(\boldsymbol{\Sigma}_{(a,c)}) \ . \tag{9}$$

4. **Expected Model Change** (Cohn et al, 1996). This strategy chooses as the most informative query the one which when added to the training set would yield the greatest model change. Quantifying the model change depends on the learning framework. For gradient-based optimization the change can be measured via the training gradient, i.e., the vector used to re-estimate parameter values (Settles and Craven, 2008). In the Bayesian framework, the model change can be quantified via a distance measure between the current distribution and the posterior distribution obtained after incorporating the candidate point

$$\text{Change}(a) = \sum_c p(c|a, \mathcal{M}) \text{distance} \left[\mathcal{M}, \mathcal{M}_{(a,c)}\right] \ .$$

A suitable distance for our setting is the Kullback-Leibler divergence between distributions, which for two Gaussians has a closed form solution and can be written as follows

$$\begin{aligned} \text{Change-KL}(a) =& \sum_c p(c|a, \mathcal{M}) KL \left[\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \mathcal{N}(\boldsymbol{\mu}_{(a,c)}, \boldsymbol{\Sigma}_{(a,c)})\right] \\ =& \sum_c p(c|a, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \left[\log \left(\frac{\det \boldsymbol{\Sigma}_{(a,c)}}{\det \boldsymbol{\Sigma}} \right) + \operatorname{Tr} \left(\boldsymbol{\Sigma}_{(a,c)}^{-1} \boldsymbol{\Sigma} \right) + \right. \\ & \left. + \left(\boldsymbol{\mu}_{(a,c)} - \boldsymbol{\mu} \right)^T \boldsymbol{\Sigma}_{(a,c)}^{-1} (\boldsymbol{\mu}_{(a,c)} - \boldsymbol{\mu}) - n \right] . \end{aligned} \tag{10}$$

Uncertainty sampling, QBC and its variants are attractive due to their applicability in various machine learning settings. Variance reduction and expected model change are robust and in many situations they have proved to be the best one can do (Schein and Ungar, 2007). Although more robust, the variance reduction and expected model change strategies are computationally more demanding since for each candidate comparison query and each possible label the posterior distribution induced has to be computed. The posterior distribution cannot be computed analytically and approximations are needed; these approximations are usually costly. In the following we discuss several variants of QBC designed for the setting in which there is available data from multiple subjects (Section 3.1). Furthermore, we show how these variants of QBC can be naturally linked to the hierarchical Bayesian modeling for reducing the computations (Section 3.2). Finally, we show connections between the active learning criteria mentioned above by reducing them to a similar form (Section 3.3).

3.1 QBC for Preference Learning

3.1.1 The Committee Members

For the QBC approach to be effective it is important that the committee be made up of consistent and representative models. The main idea in this paper is to exploit the preference learning setting with multiple subjects and use the learned models of other subjects $\mathcal{M}_1, \ldots, \mathcal{M}_M$ as the committee members when learning the preferences of a new subject.

After choosing the committee we still have to decide upon a suitable criterion for selecting the next examples. Some measures of disagreement among the committee members appear to be most obvious, and in the following we will consider several alternatives.

3.1.2 Vote Criterion

A simple and straightforward way is to consider the labels assigned by the other subjects, e.g., through the Vote criterion defined as

$$\text{Vote}(a) = \max_c \sum_{m=1}^{M} \delta(a, c; m) \, , \tag{11}$$

where $\delta(a, c; m) = 1$ if $(a, c) \in \mathcal{D}_m$, and $\delta(a, c; m) = 0$ otherwise. The score Vote(a) is minimal when the labels assigned by the committee members are equally distributed (total disagreement), and maximal when all members fully agree. There are two problems with this criterion. First, a comparison a may not be labeled by a subject m. This can be overcome if we consider the prediction for a computed based on the learned model of subject m and allow each committee member to 'vote' for its winning class. This same idea is implemented in the so-called vote entropy method (Dagan and Engelson, 1995). Second, as we will also confirm in our own experiments, in practical applications just scoring votes turns out to be suboptimal. The reason, as also suggested in (McCallum and Nigam, 1998), is that the vote criterion does not take into account the confidences of the committee members' predictions.

3.1.3 Disagreement Criterion

We will use the following notation for the predictive probability corresponding to a subject $m = 1, \ldots, M$

$$p_m(c|a) \equiv p(c|a, \mathcal{M}_m) \, .$$

The predictive probability can be computed either by taking into account the entire distribution $\mathcal{M}_m = \mathcal{N}(\boldsymbol{\mu}^m, \boldsymbol{\Sigma}^m)$

$$p_m(c|a) = \int p(c|a, \boldsymbol{\alpha}) \mathcal{N}(\boldsymbol{\alpha}|\boldsymbol{\mu}^m, \boldsymbol{\Sigma}^m) d\boldsymbol{\alpha} \, ,$$

or, for computational reasons, we can consider only a point estimate for $\mathcal{M}_m$, for example, the mean of the Gaussian distribution, and use it in Equation (2)

$$p_m(c|a) = p(c|a, \boldsymbol{\mu}^m) \, . \tag{12}$$

Inspired by (McCallum and Nigam, 1998), we propose to measure disagreement by taking the average prediction of the entire committee and computing the average Kullback-Leibler (KL) divergence of the individual predictions from the average:

$$\text{Disagreement}(a) = \sum_{m=1}^{M} \frac{1}{M} \text{KL}[\bar{p}(\cdot|a)||p_m(\cdot|a)] \, , \tag{13}$$

with $\bar{p}(\cdot|x)$ the average predictive probability of the entire committee, which will be more precisely defined in Section 3.2, and the KL divergence for discrete probabilities defined as

$$\text{KL}[p_1(\cdot|a)||p_2(\cdot|a)] = \sum_c p_1(c|a) \log\left(\frac{p_1(c|a)}{p_2(c|a)}\right) \, .$$

In (McCallum and Nigam, 1998), the Disagreement is computed between committee members constructed based on the current model, i.e., the committee changes with every update and the criterion has to be recomputed with every update. A committee of models learned on different scenarios is fixed and thus selecting examples solely based on the Disagreement criterion leads to a fixed instead of an active design: all examples can be ranked beforehand (the same applies to the Vote criterion defined above).

3.1.4 Committee Criterion

To arrive at an active design and take into account the current model, we propose a small modification, based on the following intuition. Querying examples on which the committee members disagree makes sense, because it will force the current model to make a choice between options that, according to the committee, are reasonably plausible. However, when the current model on a particular example already "made up its mind", i.e., deviates substantially from the average prediction based on what it learned from other input/output pairs, it makes no sense to still query that example, even though the committee members might disagree. Taking into account this consideration, we propose the Committee criterion which assigns a score to a candidate query comparison a through

$$\text{Committee}(a) = \frac{1}{M} \sum_{m=1}^{M} \text{KL}[\bar{p}(\cdot|a)||p_m(\cdot|a)] - \text{KL}[\bar{p}(\cdot|a)||p(\cdot|a)] \, , \tag{14}$$

with $p(\cdot|a)$ the current model's predictive probability based on the data seen so far. According to this Committee criterion, the most interesting experiments are those on which the other models disagree (the first term on the righthand side of Equation (14)), with the current model (still) undecided (the second term on the righthand side of Equation (14)).

An advantage of the Committee criterion is its computational efficiency: the first term on the righthand side of Equation (14), the Disagreement, as well as the average predictive probability can be computed beforehand. The Committee criterion does require computation of the predictive probabilities corresponding to the current model, but this is the least one could expect from an active design. This is to be compared with the QBC criterion, which requires constructing new committee members with each update, and D-optimal experimental design, which calls for keeping track of variances.

Note that we have not made any restriction so far with respect to the probabilistic models used in the active learning design. In the following we restrict ourselves to the log-linear models introduced in Section 2. They have some nice properties, which simplify the computation of the committee criterion (Section 3.2), and provide a natural link to hierarchical Bayesian modeling (Section 2.3). The general idea, of using the already learned models from the other scenarios as the committee members in a QBC-like approach, is of course also applicable to other models.

3.2 Average Probability

For linear form of the utility functions the likelihood function defined in Equation (2) is a log-linear model (Christensen, 1997). The log-odds of the model are linear in the parameter.

Let $\boldsymbol{\alpha}^m$ be a point estimate of the model $\mathcal{M}_m$ learned from the observations related to subject m. Let $p_m(c|a)$ be the predictive probability defined in Equation (12). We define the average predictive probability of the committee, $\bar{p}(c|a)$, as the prediction probability that is closest to the prediction probabilities of the members:

$$\bar{p}(c|a) \equiv \operatorname*{argmin}_{p(c|a)} \sum_{m=1}^{M} \frac{1}{M} \mathrm{KL}[p(c|a)||p_m(c|a)] \,. \tag{15}$$

The solution is the so-called logarithmic opinion pool (Bordley, 1982)

$$\bar{p}(c|a) = \frac{1}{Z(a)} \prod_{m=1}^{M} [p_m(c|a)]^{\frac{1}{M}} = \frac{1}{Z(a)} \exp\left(\frac{1}{M} \sum_{m=1}^{M} \log p_m(c|a) \right) \,, \tag{16}$$

with $Z(a)$ a normalization constant

$$Z(a) = \sum_c \prod_{m=1}^{M} [p_m(c|a)]^{\frac{1}{M}} \,.$$

For log-linear models, the logarithmic opinion pool boils down to a simple averaging of model parameters:

$$\bar{p}(c|a) = p(c|a, \bar{\boldsymbol{\mu}}) \quad \text{with} \quad \bar{\boldsymbol{\mu}} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\mu}^m \,. \tag{17}$$

This natural combination between log-linear model and logarithm opinion pool makes that we have a preference for the logarithmic opinion pool over the linear opinion pool used in (McCallum and Nigam, 1998).

The average $\bar{\boldsymbol{\mu}}$ in the logarithmic opinion pool is then precisely the mean of the learned hierarchical prior. Summarizing, once we have learned a hierarchical prior from the data available for subjects 1 through M using the EM algorithm, we can start off the new model $M + 1$ from this prior (as is normally done in hierarchical Bayesian learning). On top of this, the same EM algorithm gives us the information we need to compute the Committee criterion that can be used subsequently to select new inputs to label.

3.3 Connections between Criteria

Assuming the updates of the posterior distribution for each (a, c) lead to small changes in the model $\mathcal{M}$ we can approximate the active learning criteria to the following form

$$\sum_c p(c|a,\cdot)\boldsymbol{g}(c|a,\cdot)^T \cdot \boldsymbol{g}(c|a,\cdot) \; .$$

More precisely,

$$\text{Variance-logdet}(a) \approx \sum_c p(c|a,\boldsymbol{\mu})\boldsymbol{g}(c|a,\boldsymbol{\mu})^T \, \boldsymbol{\Sigma} \, \boldsymbol{g}(c|a,\boldsymbol{\mu}) \; , \tag{18}$$

$$\text{Committee}(a) \approx \frac{1}{2}\sum_c p(c|a,\bar{\boldsymbol{\mu}})\boldsymbol{g}(c|a,\bar{\boldsymbol{\mu}}) \, \tilde{\boldsymbol{\Sigma}} \, \boldsymbol{g}(c|a,\bar{\boldsymbol{\mu}}) \; , \tag{19}$$

with $\bar{\boldsymbol{\mu}}$ the hierarchical prior mean and

$$\tilde{\boldsymbol{\Sigma}} \equiv \frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{\mu}^m - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}^m - \bar{\boldsymbol{\mu}})^T - (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^T \; .$$

Furthermore,

$$\text{Variance-trace}(a) = \sum_c p(c|a,\boldsymbol{\mu})\boldsymbol{g}(c|a,\boldsymbol{\mu})^T V^2 \boldsymbol{g}(c|a,\boldsymbol{\mu}) \; ,$$

$$\frac{1}{2}\text{Change-KL}(\boldsymbol{x}) \approx \text{Variance-logdet}(\boldsymbol{x}) \; ,$$

The proofs of the above approximations are given in the Appendix.

We will focus on the differences between the Variance-logdet criterion (considered as the reference) and the Committee criterion. The differences between their approximations are:

1. The gradients $\boldsymbol{g}(c|a,\cdot)$ are evaluated at different points: the prior hierarchical mean $\bar{\boldsymbol{\mu}}$ and the current posterior mean $\boldsymbol{\mu}$. This effect is small since $\boldsymbol{\mu}$ is still close enough to $\bar{\boldsymbol{\mu}}$ for a sufficiently accurate approximation of the gradients, in particular at the start of the learning when selecting the right points to label is more important.
2. The current posterior variance $\boldsymbol{\Sigma}$ is replaced by $\tilde{\boldsymbol{\Sigma}}$. The effect of the precise weighting of the gradients is not so important, and again, at the beginning of learning $\tilde{\boldsymbol{\Sigma}}$ is close to $\boldsymbol{\Sigma}$.

In practice, the effect of the approximations from above is not very important, since we will see in the experimental evaluation that the Committee criterion and the Variance-logdet criterion work about the same. This happens because the way in which experiments are selected, is more important at the beginning of the learning process, when $\boldsymbol{\mu}$ is still close to the prior mean $\bar{\boldsymbol{\mu}}$, and $\tilde{\boldsymbol{\Sigma}}$ to $\boldsymbol{\Sigma}$.

4 Experimental Evaluation

In this section we discuss the results of the experimental evaluation on an artificially simulated data set and on a real data set.

4.1 Artificial Data

We generated artificial data for 10 settings. Each setting consists of a set of inputs ($x \in X$, x is a 2-dimensional vector) and $M = 50$ scenarios (each scenario is characterized by a vector $\boldsymbol{\theta}$ which generates a binary response for $x \in X$ according to Equation (2)). The goal of the simulations on the artificial data set is to validate empirically the similarities between the active learning criteria derived in Section 3.3.

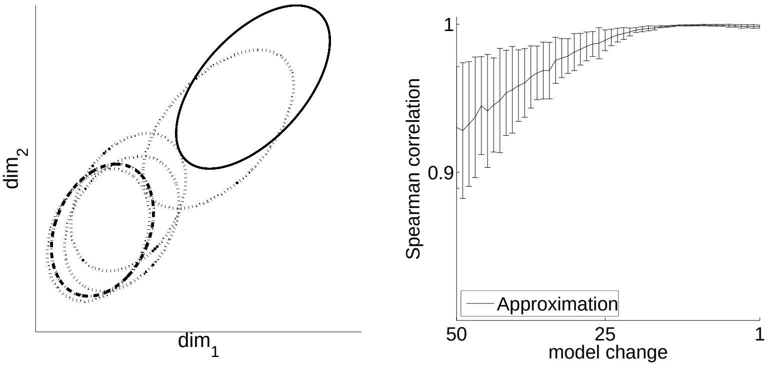
First, we quantified the effect of the approximation of the Variance-logdet criterion stated in Equation (18). The Variance-logdet criterion needs to compute the variance of the posterior induced by a candidate point, which is in most of the cases computationally expensive. We give an approximation scheme of this criterion which only needs to compute gradients. The derivation of the approximation is given in Lemma 2 in Appendix. The approximation is based on Equations (22) and (23) which assumes a small model change after incorporating a new data point, i.e., the difference between the prior Gaussian distribution before incorporating a candidate point and the posterior distribution obtained after including the point. The left-hand side of Figure 1 illustrates the model change for Gaussian distributions. The solid-line ellipse is the contour plot of the Gaussian representing a prior distribution. This prior is updated by incorporating training data points. After each update the posterior distribution obtained is considered as the prior in the next update. The dotted-line ellipses are the contour plots of the intermediate Gaussians obtained after $1, 2, 3$, and 4 updates. The dashed-line ellipse is the contour plot of the model obtained after 5 updates from the prior. The magnitude of the change decreases as more data points are incorporated into the training. The plot on the right-hand side of Figure 1 is the Spearman rank correlation coefficient between the scores assigned by the Variance-logdet criterion and its approximation to 20 randomly selected points in each setting from each scenario. The error bars show the standard deviation of the correlations averaged over 10 settings and in each setting over 50 scenarios. The model change on the x-axis is estimated by the larger eigenvalue of the covariance matrix. The accuracy of the approximation of the Variance-logdet is better for smaller changes from the prior to posterior.

Second, we quantified the similarity between the Variance-logdet and the Committee criteria. They can be approximated to the same form (Equations (18) and (19)) but with the difference of evaluating the gradients and the predictive probabilities at different points, and having a different weighting matrix. In each scenario we considered one model as the current one and learned a hierarchical prior from the other models. We computed the scores assigned by Variance-logdet and Committee criteria to 100 randomly selected points after a certain number of updates from the prior. Figure 3 shows the Spearman rank correlations between theses scores as a function of the number of updates from the prior. The correlations decreases with more updates from the prior.

4.2 Real Data

4.2.1 Preferences for Sound Quality

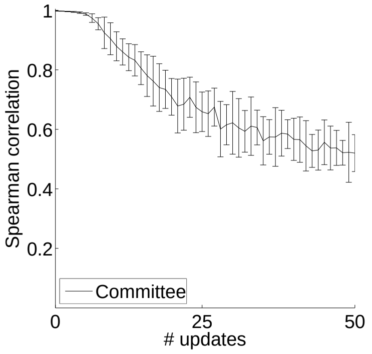
We performed a set of simulations on a data set data set of audiological experiments. The data set, which was described in Arehart et al (2007), consists of evaluations of sound quality from 32 subjects, 14 normal hearing and 18 hearing impaired subjects.



(a) Gaussian update

(b) Approximation

Fig. 1: Left: Update of the Gaussian prior (the solid line) to new posteriors. Right: Spearman rank correlation coefficient between the scores assigned by the Variance-logdet criterion and its approximation to 20 randomly selected points. The error bars show the standard deviation of the correlations averaged over 10 settings and in each setting over 50 scenarios. The model change on the x-axis is estimated by the larger eigenvalue of the covariance matrix.



(a) Committee

Fig. 2: Spearman correlation coefficient between the scores assigned by the Variance-logdet and Committee criteria as a function of the number of updates from the prior. The error bars show the standard deviation of the correlations averaged over 10 settings and in each setting over 50 scenarios.

Each person was subjected to 576 paired-comparison listening experiments of the form (a, c), where $a = (x_1, x_2)$ and x_1 and x_2 represent one sound sample processed with two different settings of the hearing-aid parameters, and $c = \{1, 2\}$ denotes which of the two options was preferred by the user.

Multi-task learning. Our first optimization proposed to the preference learning is multi-task learning. We implemented the multi-task formalism by learning a hierarchical prior which is used as the starting model when learning the preferences of a new subject. In each simulation, one subject was left out and the hierarchical prior was learned from the rest of the subjects. The data set for the left-out subject, was split into 5 folds, 1 fold was used for training and and the others for testing. The hierarchical prior was updated with the points from the training (the learning phase). After every update predictions were made on the testing set (the accuracy of the predictions on the test data was used as a measure of how much we learned about the subject preferences).

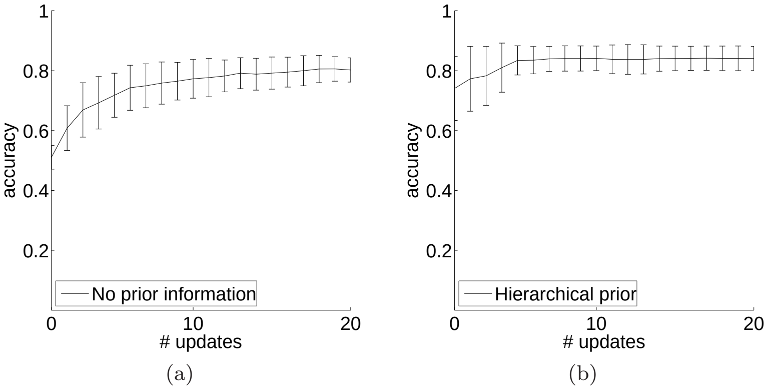


Fig. 3: Left: prediction accuracy obtained by a model which assumes no prior information about preferences. Right: prediction accuracy obtained in the multi-task setting.

Active learning. We used the same setup as above. The training data was used as a pool out of which points were chosen for labeling using the active learning criteria from Section 3. A hierarchical prior was learned from the group of subjects other than the current one. We compared the performances obtained by random and active learning criteria. For each subject, we updated the hierarchical prior based on the information from experiments which were selected either randomly or optimally using one of the active learning criteria. The plots in Figure 4 show the prediction error obtained on the test data.

5 Conclusions

Our approach to active learning in a multi-scenario setting combines a hierarchical Bayesian prior (to learn from related scenarios) with active learning (to learn efficiently by selecting informative examples). Our new Committee criterion inspired by the Query-by-Committee method, is very similar to the standard criteria from experimental design, in particular in the early stages of active learning, but computationally much more efficient,

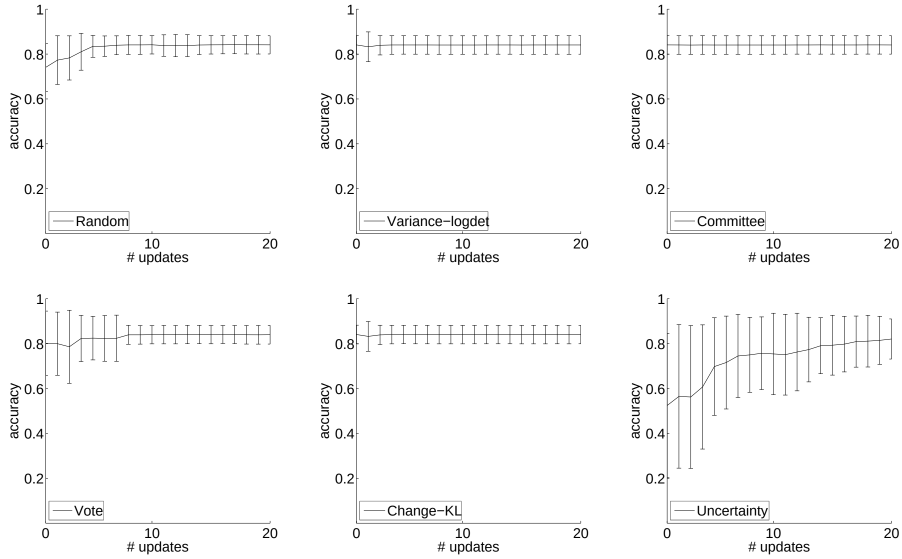


Fig. 4: Prediction error.

Appendix

In this appendix we prove the equivalences between the active learning criteria stated in Section 3. We show that these criteria can be approximated to the same form, namely

$$\sum_c p(c|a,\cdot)\boldsymbol{g}(c|a,\cdot)^T \cdot \boldsymbol{g}(c|a,\cdot)\ .$$

The difference between the approximations for different criteria is the point in which the gradients and the probabilities are evaluated and the weighting matrix of the gradients. We assume that the current model of the data is a Gaussian distribution $\mathcal{M} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathcal{M}_{(a,c)} = \mathcal{N}(\boldsymbol{\mu}_{(a,c)}, \boldsymbol{\Sigma}_{(a,c)})$ represents the Gaussian approximation of posterior distribution obtained after incorporating (a, c). We consider probabilistic choice models of the from given in Equation (2). We simplify the notation omitting the references to comparison j and to the index function.

$$p(c|a,U) = \frac{\exp\left[U(x_c)\right]}{\sum_{c'=1}^{A} \exp\left[U(x_{c'})\right]}$$

Following the expression of the utility function from Section 2.2 we rewrite the equation from above as

$$p(c|a,\boldsymbol{\alpha}) = \frac{\exp\left[\sum_{i=1}^{T} \phi_{ci}(x_c)\alpha_i\right]}{Z(\boldsymbol{\alpha},a)} \quad \text{with} \quad Z(\boldsymbol{\alpha},a) \equiv \sum_{c'} \exp\left[\sum_i \phi_{c'i}(x_{c'})\alpha_i\right]$$

We define the derivatives of the log probabilities

$$\boldsymbol{g}(c|a,\boldsymbol{\alpha}) \equiv \frac{\partial \log p(c|a,\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \,,$$
$$\boldsymbol{H}(c|a,\boldsymbol{\alpha}) \equiv \frac{\partial^2 \log p(c|a,\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} \,.$$

We first state and proof a lemma which will be be used in further proofs.

Lemma 1 *For any a and $\boldsymbol{\alpha}$ we have the following relation between second and first derivatives:*

$$\sum_c p(c|a,\boldsymbol{\alpha})\boldsymbol{H}(c|a,\boldsymbol{\alpha}) = -\sum_c p(c|a,\boldsymbol{\alpha})\boldsymbol{g}(c|a,\boldsymbol{\alpha})\boldsymbol{g}(c|a,\boldsymbol{\alpha})^T \,,$$

Proof. We use shorthand notation $p_c = p(c|a,\boldsymbol{\alpha})$, $g_{cj} = g_j(c|a,\boldsymbol{\alpha})$, etc, omitting the dependencies on a and $\boldsymbol{\alpha}$.

From $\log p_c = \sum_j \phi_{ci}\alpha_j - \log Z$, it is easy to see that

$$g_{cj} = \phi_{cj} - \frac{\partial \log Z}{\partial \alpha_j}$$
$$H_{c,ij} = -\frac{\partial^2 \log Z}{\partial \alpha_i \partial \alpha_j}$$

Furthermore,

$$\begin{aligned}
\frac{\partial \log Z}{\partial \alpha_j} &= \frac{1}{Z}\frac{\partial Z}{\partial \alpha_j} = \frac{1}{Z}\sum_c \exp\left[\sum_{j'} \phi_{cj'}\alpha_{j'}\right]\phi_{cj} = \sum_c p_c \phi_{cj} \\
\frac{\partial^2 \log Z}{\partial \alpha_i \partial \alpha_j} &= \sum_c \phi_{cj}\frac{\partial p_c}{\partial \alpha_i} \\
&= \sum_c \phi_{cj}\left[\frac{\exp(\sum_{j'}\phi_{cj'}\alpha_{j'})\phi_{ci}Z - \frac{\partial Z}{\partial \alpha_i}\exp(\sum_{j'}\phi_{cj'}\alpha_{j'})}{Z^2}\right] \\
&= \sum_c \phi_{cj}[\phi_{ci}p_c - \underbrace{\frac{1}{Z}\frac{\partial Z}{\partial \alpha_i}} p_c] \\
&= \sum_c \phi_{cj}[\phi_{ci}p_c - \sum_{c'}(p_{c'}\phi_{c'i})p_c] \\
&= \sum_c p_c\phi_{cj}\phi_{ci} - \sum_c p_c\phi_{cj}\sum_{c'} p_{c'}\phi_{c'i},
\end{aligned}$$

and thus

$$g_{cj} = \phi_{cj} - \sum_c p_c\phi_{cj} \tag{20}$$
$$H_{c,ij} = -\sum_{c'} p_{c'}\phi_{c'j}\phi_{c'i} + \sum_{c'} p_{c'}\phi_{c'j}\sum_{c''} p_{c''}\phi_{c''i} = H_{ij} \,.$$

Note that the second derivative is in fact independent of c.

We then have

$$\begin{aligned}\sum_c p_c H_{c,ij} = \sum_c p_c H_{ij} = H_{ij} &= -\sum_c p_c \phi_{ci}\phi_{cj} + \sum_c p_c \phi_{ci} \sum_{c'} p_{c'}\phi_{c'j} \\ &= -\sum_c p_c \phi_{ci}\phi_{cj} + \sum_c p_c \phi_{ci} \sum_{c'} p_{c'}\phi_{c'j} - \sum_c p_c \phi_{ci} \sum_{c'} p_{c'}\phi_{c'j} + \sum_c p_c \phi_{ci} \sum_{c'} p_{c'}\phi_{c'j} \\ &= -\sum_c p_c \left(\phi_{ci} - \sum_{c'} p_{c'}\phi_{c'i}\right)\left(\phi_{cj} - \sum_{c'} p_{c'}\phi_{c'j}\right) \\ &= -\sum_c p_c g_{ci} g_{cj}\end{aligned}$$

which proofs the result stated in the lemma. □

Lemma 2 *In a first order approximation, assuming that $\boldsymbol{\Sigma}_{(a,c)}$ is close to $\boldsymbol{\Sigma}$, we can simplify*

$$\textit{Variance-logdet}(a) \approx \sum_c p(c|a,\boldsymbol{\mu})\boldsymbol{g}(c|a,\boldsymbol{\mu})^T \boldsymbol{\Sigma}\, \boldsymbol{g}(c|a,\boldsymbol{\mu}) \,.$$

Proof.

In a first order approximation we have

$$\boldsymbol{\Sigma}^{-1}_{(a,c)} \approx \boldsymbol{\Sigma}^{-1} - \frac{\partial^2 \log p(c|a,\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T}\Big|_{\boldsymbol{\alpha}=\boldsymbol{\mu}} \tag{21}$$

where we ignored the change from the old $\boldsymbol{\alpha}$ to a new MAP solution depending on c and a.

For a matrix A and ϵ small compared to A, the following holds

$$\log\det(A+\epsilon) = \log\det(A) - \mathrm{Tr}[A^{-1}\epsilon] \tag{22}$$

Assuming that $\boldsymbol{\Sigma}^{-1}_{(a,c)}$ is close to $\boldsymbol{\Sigma}^{-1}$ we can plug them in the equation from above to get

$$\log\det\boldsymbol{\Sigma}^{-1}_{(a,c)} \approx \log\det\boldsymbol{\Sigma}^{-1} - \mathrm{Tr}[\boldsymbol{\Sigma}\,\boldsymbol{H}(c|a,\boldsymbol{\mu})] \,. \tag{23}$$

The probability that the subject indeed gives the response c when presented a follows by integrating $p(c|a,\boldsymbol{\alpha})$ over the current posterior. We make a second order Taylor expansion of $p(c|a,\boldsymbol{\alpha})$ around the point $\boldsymbol{\mu}$:

$$\begin{aligned}p(c|a) &= \int d\boldsymbol{\alpha} p(c|a,\boldsymbol{\alpha})\mathcal{N}(\boldsymbol{\alpha}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \\ &\approx \int d\boldsymbol{\alpha}\left[p(c|a,\boldsymbol{\alpha}) + (\boldsymbol{\alpha}-\boldsymbol{\mu})^T\frac{\partial p(c|a,\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}\Big|_{\boldsymbol{\alpha}=\boldsymbol{\mu}} + \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\mu})^T\frac{\partial^2 p(c|a,\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T}\Big|_{\boldsymbol{\alpha}=\boldsymbol{\mu}}(\boldsymbol{\alpha}-\boldsymbol{\mu})\right]\mathcal{N}(\boldsymbol{\alpha}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \\ &= p((c|a,\boldsymbol{\mu}) + \frac{1}{2}\mathrm{Tr}\left[\boldsymbol{\Sigma}\frac{\partial^2 p(c|a,\boldsymbol{\alpha})}{\partial\alpha\partial\alpha^T}\Big|_{\boldsymbol{\alpha}=\boldsymbol{\mu}}\right] \,.\end{aligned}$$

The first order term cancels since the gradient is zero at the maximum solution $\boldsymbol{\alpha}$. Again in lowest order we can ignore the correction upon $p(c|a, \boldsymbol{\alpha})$ and arrive at

$$
\begin{aligned}
\text{Variance-logdet}(a) &= -\sum_c p(c|a, \boldsymbol{\mu}) \log\det \boldsymbol{\Sigma}_{(a,c)} + \log\det \boldsymbol{\Sigma} \\
&= -\sum_c p(c|a, \boldsymbol{\mu}) \left[\log\det \boldsymbol{\Sigma}_{(a,c)} - \log\det \boldsymbol{\Sigma}\right] \\
&= -\sum_c p(c|a, \boldsymbol{\mu})[\log\det(\boldsymbol{\Sigma}^{-1}_{(a,c)})^{-1} - \log\det(\boldsymbol{\Sigma}^{-1})^{-1}] \\
&= -\sum_c p(c|a, \boldsymbol{\mu})[-\log\det(\boldsymbol{\Sigma}^{-1}_{(a,c)}) + \log\det(\boldsymbol{\Sigma}^{-1})] \\
&= -\sum_c p(c|a, \boldsymbol{\mu}) \mathrm{Tr}[\boldsymbol{\Sigma} \boldsymbol{H}(c|a, \boldsymbol{\mu})]
\end{aligned}
$$

Lemma 1 then gives the result. □

Lemma 3 *In a lowest order approximation the committee criterion can be written as*

$$
\frac{1}{2} \mathit{Committee}(a) = \frac{1}{2} \sum_c p(c|a, \bar{\boldsymbol{\mu}}) \boldsymbol{g}(c|a, \bar{\boldsymbol{\mu}})^T \tilde{\boldsymbol{\Sigma}} \, \boldsymbol{g}(c|a, \bar{\boldsymbol{\mu}}) \, ,
$$

where $\bar{\boldsymbol{\mu}}$ is the prior mean learned from all other users and

$$
\tilde{\boldsymbol{\Sigma}} \equiv \frac{1}{M} \sum_{m=1}^{M} (\boldsymbol{\mu}^m - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}^m - \bar{\boldsymbol{\mu}})^T - (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^T \, .
$$

Making a second order Taylor expansion, the Kullback-Leibler divergence between probabilities based on $\bar{\boldsymbol{\mu}}$ and $\boldsymbol{\mu}$ when these are close together is:

$$
\begin{aligned}
&\sum_c p(c|a, \bar{\boldsymbol{\mu}}) \log \left[\frac{p(c|a, \bar{\boldsymbol{\mu}})}{p(c|a, \boldsymbol{\mu})}\right] \\
&\quad \approx -\frac{1}{2} \sum_c p(c|a, \bar{\boldsymbol{\mu}}) (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^T \boldsymbol{H}(c|a, \bar{\boldsymbol{\mu}}) (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}) \\
&\quad = \frac{1}{2} \sum_c p(c|a, \bar{\boldsymbol{\mu}}) \left[(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^T \boldsymbol{g}(c|a, \bar{\boldsymbol{\mu}})\right]^2 \, ,
\end{aligned}
\tag{24}
$$

where the first order term cancels since, from Equation (20),

$$
\sum_c p(c|a, \boldsymbol{\alpha}) \boldsymbol{g}(c|a, \boldsymbol{\alpha}) = 0 \, .
$$

Using Equation (24) to approximate the KL divergences between the predictive probabilities from the definition of the Committee criterion (Equation (14)) and computing the sums, we obtain the result stated in the lemma. □

Lemma 4 *In a first order approximation, the Variance-trace criterion boils down to*

$$
\mathit{Variance\text{-}trace}(a) = \sum_c p(c|a, \boldsymbol{\mu}) \boldsymbol{g}(c|a, \boldsymbol{\mu})^T \boldsymbol{\Sigma}^2 \boldsymbol{g}(c|a, \boldsymbol{\mu}) \, .
$$

Proof. We have

$$\boldsymbol{\Sigma}_{(a,c)} = \left(\boldsymbol{\Sigma}^{-1}_{(a,c)}\right)^{-1} = \left(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}_{(a,c)} - \boldsymbol{\Sigma}^{-1}\right)^{-1} \approx -\boldsymbol{\Sigma}\left[\boldsymbol{\Sigma}^{-1}_{(a,c)} - \boldsymbol{\Sigma}^{-1}\right]\boldsymbol{\Sigma}\,.$$

Use of Equation (21) and Lemma 1 give the result. □

Suppose we have a Gaussian prior with mean α and variance Σ and have observed N data samples $\{x_i, c_i\}$. In the Laplace approximation, the posterior is then approximated by another Gaussian with mean equal to the maximum a posteriori solution

$$\boldsymbol{\mu} \equiv \operatorname*{argmin}_{\alpha} E(\boldsymbol{\alpha})\,,$$

with

$$E(\boldsymbol{\alpha}) = -\sum_{i=1}^{N} \log p(c_i|x_i, \boldsymbol{\alpha}) + \frac{1}{2}(\boldsymbol{\alpha} - \alpha)^T \Sigma^{-1}(\boldsymbol{\alpha} - \alpha)$$

and variance equal to the inverse of the Hessian, the second derivative of $E(\boldsymbol{\alpha})$,

$$V^{-1} = \left.\frac{\partial^2 E(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T}\right|_{\boldsymbol{\alpha}=\boldsymbol{\mu}}\,. \tag{25}$$

Lemma 5 *In a first order approximation, assuming that $\boldsymbol{\Sigma}_{(a,c)}$ is close to $\boldsymbol{\Sigma}$, we have*

$$\textit{Change-KL}(a) \approx \textit{Variance-logdet}(a)\,,$$

i.e., the two criteria are indistinguishable.

Proof. We evaluate the terms of Change-KL criterion one by one,

$$\text{Change-KL}(a) = \sum_c p(c|a)\left[\underbrace{\log\left(\frac{\det\boldsymbol{\Sigma}_{(a,c)}}{\det\boldsymbol{\Sigma}}\right)}_{} + \underbrace{\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}_{(a,c)}\boldsymbol{\Sigma}\right)}_{} + \underbrace{(\boldsymbol{\mu}_{(a,c)} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}_{(a,c)}(\boldsymbol{\mu}_{(a,c)} - \boldsymbol{\mu})}_{} - n\right]$$

The first term gives

$$\sum_c p(c|a)\log\left(\frac{\det\boldsymbol{\Sigma}_{(a,c)}}{\det\boldsymbol{\Sigma}}\right) = -\text{Variance-logdet}(a)\,.$$

The second term

$$\begin{aligned}\sum_c p(c|a)\,\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}_{(a,c)}\boldsymbol{\Sigma}\right] &= \sum_c p(c|a)\,\mathrm{Tr}[(\boldsymbol{\Sigma}^{-1} - \boldsymbol{H}(c|a,\boldsymbol{\mu}))\boldsymbol{\Sigma}] \\ &= n - \sum_c p(c|a)\,\mathrm{Tr}[H(c|a,\boldsymbol{\mu})\boldsymbol{\Sigma}] = n - \sum_c p(c|a,\boldsymbol{\mu})\,\mathrm{Tr}[\boldsymbol{\Sigma}\,\boldsymbol{H}(c|a,\boldsymbol{\mu})] \\ &= n + \text{Variance-logdet}(a)\end{aligned}$$

In the Laplace approximation, the new posterior mean corresponds to the minimum of the error with the log probability of the new observation added:

$$\boldsymbol{\alpha}(a,c) = \operatorname*{argmin}_{\boldsymbol{\alpha}}\left[E(\boldsymbol{\alpha}) + \log p(c|a,\boldsymbol{\alpha})\right]\,.$$

In a lowest order approximation we assume $\boldsymbol{\alpha}(a,c)$ to be close to $\boldsymbol{\alpha}$ and make a Taylor expansion of the term between brackets:

$$\begin{aligned} E(\boldsymbol{\alpha}) + \log p(k;\boldsymbol{x},\boldsymbol{\alpha}) \approx\ & E(\boldsymbol{\mu}) + \log p(k;\boldsymbol{x},\boldsymbol{\mu}) \\ & + (\boldsymbol{\alpha}-\boldsymbol{\mu})^T \left[\frac{\partial E(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\Big|_{\boldsymbol{\alpha}=\boldsymbol{\mu}} + \frac{\partial \log p(k,\boldsymbol{x},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\Big|_{\boldsymbol{\alpha}=\boldsymbol{\mu}} \right] \\ & + \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\mu})^T \left[\frac{\partial^2 E(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T}\Big|_{\boldsymbol{\alpha}=\boldsymbol{\mu}} + \frac{\partial^2 \log p(k,\boldsymbol{x},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T}\Big|_{\boldsymbol{\alpha}=\boldsymbol{\mu}} \right] (\boldsymbol{\alpha}-\boldsymbol{\mu}) \end{aligned}$$

$\frac{\partial E(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\Big|_{\boldsymbol{\alpha}=\boldsymbol{\mu}} = 0$ since it represents the gradient of the posterior evaluated at the maximum point $\boldsymbol{\mu}$. $\frac{\partial^2 E(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T} = V^{-1}$ as defined in Equation (25). $\frac{\partial^2 \log p(k,\boldsymbol{x},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T}\Big|_{\boldsymbol{\alpha}=\boldsymbol{\mu}}$ can be ignored, since it corresponds to just one likelihood term and therefore is small in comparison with $\frac{\partial^2 E(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T}$ which corresponds to all the other likelihood terms. We have

$$E(\boldsymbol{\alpha}) + \log p(k;\boldsymbol{x},\boldsymbol{\alpha}) \approx E(\boldsymbol{\mu}) + \log p(k;\boldsymbol{x},\boldsymbol{\mu}) + \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\mu})^T V^{-1} (\boldsymbol{\alpha}-\boldsymbol{\mu}) + (\boldsymbol{\alpha}-\boldsymbol{\mu})^T \boldsymbol{g}(k;\boldsymbol{x},\boldsymbol{\mu})$$

Setting the derivative w.r.t. α to zero we obtain the approximate solution

$$\boldsymbol{\mu}_{(a,c)} \approx \boldsymbol{\mu} - \boldsymbol{\Sigma} \boldsymbol{g}(c|a,\boldsymbol{\mu}) .$$

In the same lowest order, we obtain for the third term

$$\begin{aligned} & \sum_c p(c|a) \left(\boldsymbol{\mu}_{(a,c)} - \boldsymbol{\mu} \right)^T \boldsymbol{\Sigma}^{-1}_{(a,c)} (\boldsymbol{\mu}_{(a,c)} - \boldsymbol{\mu}) \approx \\ & \qquad \approx \sum_c p(c|a,\boldsymbol{\mu}) \boldsymbol{g}(c|a,\boldsymbol{\mu})^T \boldsymbol{\Sigma}\, \boldsymbol{g}(c|a,\boldsymbol{\mu}) = \text{Variance-logdet}(a) . \end{aligned}$$

Collection of all the terms then gives the result. □

References

Aiolli F, Sperduti A (2004) Learning preferences for multiclass problems. In: Advances in Neural Information Processing Systems 17, MIT Press, pp 17–24

Arehart K, Kates J, Anderson C, Harvey Jr L (2007) Effects of noise and distortion on speech quality judgments in normal-hearing and hearing-impaired listeners. J Acoust Soc Am 122(2):1150–1164

Bakker B, Heskes T (2003) Task clustering and gating for Bayesian multitask learning. Journal of Machine Learning Research 4:83–99

Berger M (1994) D-optimal sequential sampling designs for item response theory models. Journal of Educational and Behavioral Statistics 19:43–56

Birlutiu A, Groot P, Heskes T (2009) Multi-Task Preference Learning with an Application to Hearing Aid Personalization. In: Neurocomputing, DOI 10.1016/j.neucom.2009.11.025

Bishop C (2006) Pattern Recognition and Machine Learning. Springer

Blei D, Ng A, Jordan M (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022, DOI 0.1162/jmlr.2003.3.4-5.993

Blythe J (2002) Visual exploration and incremental utility elicitation. In: Eighteenth national conference on Artificial intelligence, American Association for Artificial Intelligence, Menlo Park, CA, USA, pp 526–532

Bordley R (1982) A multiplicative formula for aggregating probability assessments. Management Science 28(10):1137–1148

Boutilier C, Zemel R, Marlin B (2003) Active collaborative filtering. In: Proceedings of the Nineteenth Annual Conference on Uncertainty in Artificial Intelligence, pp 98–106

Bradley R, Terry M (1952) Rank analysis of incomplete block designs, I. the method of paired comparisons. Biometrika 39:324–345

Brochu E, de Freitas N, Ghosh A (2007) Active preference learning with discrete choice data. In: Advances in Neural Information Processing Systems

Caruana R, Baluja S, Mitchell T (1996) Using the future to sort out the present: Rankprop and multitask learning for medical risk evaluation. In: Advances in Neural Information Processing Systems 8, pp 959–965

Chajewska U, Koller D, Parr R (2000) Making rational decisions using adaptive utility elicitation. In: Proceedings of the Seventeenth National Conference on Artificial Intelligence, pp 363–369

Chaloner K, Verdinelli I (1995) Bayesian experimental design: A review. Statistical Science 10:273–304

Christensen R (1997) Log-Linear Models and Logistic Regression, 2nd edn. Springer-Verlag

Chu W, Ghahramani Z (2005a) Extensions of Gaussian processes for ranking: semi-supervised and active learning. In: NIPS 2005 Workshop on Learning to Rank, Whistler, BC

Chu W, Ghahramani Z (2005b) Preference Learning with Gaussian Processes. In: Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany

Clyde M, Muller P, Parmigiani G (1993) Optimal designs for heart defibrillators

Cohn D, Ghahramani Z, Jordan M (1996) Active learning with statistical models. Journal of Artificial Intelligence Research 4:129–145

Crammer K, Singer Y (2001) Pranking with ranking. In: Advances in Neural Information Processing Systems 14, MIT Press, pp 641–647

Dagan I, Engelson SP (1995) Committee-based sampling for training probabilistic classifiers. In: In Proceedings of the Twelfth International Conference on Machine Learning, Morgan Kaufmann, pp 150–157

Dasgupta S, Hsu D (2008) Hierarchical sampling for active learning. In: ICML '08: Proceedings of the 25th international conference on Machine learning, ACM, New York, NY, USA, pp 208–215, DOI 10.1145/1390156.1390183

Doyle J (2004) Prospects for preferences. Computational Intelligence 20(2):111–136

Dror H, Steinberg D (2008) Sequential experimental designs for generalized linear models. Journal of the American Statistical Association (103):288–298

Evgeniou T, Micchelli C, Pontil M (2005) Learning multiple tasks with kernel methods. Journal of Machine Learning Research 6:615–637

Fedorov V (1972) Theory of Optimal Experiments. Academic Press

Ford I, Silvey S (1980) A sequentially constructed design for estimating a nonlinear parametric function. Biometrika 67:381–388

Freund Y, Shamir E, Tishby N (1997) Selective sampling using the Query by Committee algorithm. In: Machine Learning, pp 133–168

Fürnkranz J, Hüllermeier E (2003) Pairwise preference learning and ranking. In: Proceedings of the 14th European Conference on Machine Learning, Springer-Verlag, pp 145–156

Gelman A, Carlin J, Stern H, Rubin D (2003) Bayesian Data Analysis, Second Edition. Chapman & Hall/CRC

Geman S, Bienenstock E, Doursat R (1992) Neural networks and the bias/variance dilemma. Neural Computation 4(1):1–58, DOI 10.1162/neco.1992.4.1.1

Glickman M, Jensen S (2005) Adaptive paired comparison design. Journal of Statistical Planning and Inference 127:279–293

Har-peled S, Roth D, Zimak D (2002) Constraint classification: A new approach to multiclass classification and ranking. In: Advances in Neural Information Processing Systems 15, pp 365–379

Harpale A, Yang Y (2008) Personalized active learning for collaborative filtering. In: SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, pp 91–98, DOI 10.1145/1390334.1390352

Herbrich R, Graepel T, Bollmann-Sdorra P, Obermayer K (1998) Learning preference relations for information retrieval

Heskes T, de Vries B (2005) Incremental utility elicitation for adaptive personalization. In: Verbeeck K, Tuyls K, Nowé A, Manderick B, Kuijpers B (eds) BNAIC 2005, Proceedings of the Seventeenth Belgium-Netherlands Conference on Artificial Intelligence, Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten, Brussels, pp 127–134

Jin R, Si L (2004) A bayesian approach toward active learning for collaborative filtering. In: AUAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence, AUAI Press, Arlington, Virginia, United States, pp 278–285

Kanninen B (2002) Optimal design for multinomial choice experiments. Journal of Marketing Research 39:307–317

Lewi J, Butera R, Paninski L (2009) Sequential optimal design of neurophysiology experiments. Neural Computation 21(3):619–687

Lewis D, Gale W (1994) A sequential algorithm for training text classifiers. In: SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, Springer-Verlag New York, Inc., New York, NY, USA, pp 3–12

Mackay D (1992) Information-based objective functions for active data selection. Neural Computation 4:590–604

Mackay D (2002) Information Theory, Inference & Learning Algorithms. Cambridge University Press, New York, NY, USA

McCallum A, Nigam K (1998) Employing EM and pool-based active learning for text classification. In: International Conference on Machine Learning 1998, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp 350–358

Melville P, Mooney R (2004) Diverse ensembles for active learning. In: ICML '04: Proceedings of the twenty-first international conference on Machine learning, ACM, New York, NY, USA, p 74, DOI 10.1145/1015330.1015385

Minka T (2001) A family of approximation methods for approximate Bayesian inference. PhD thesis, MIT

Schein A, Ungar L (2007) Active learning for logistic regression: an evaluation. Machine Learning 68(3):235–265, DOI 10.1007/s10994-007-5019-5

Settles B, Craven M (2008) An analysis of active learning strategies for sequence labeling tasks. In: EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Morristown, NJ, USA, pp 1070–1079

Seung H, Opper M, Sompolinsky H (1992) Query by Committee. In: COLT, ACM, New York, NY, USA, pp 287–294

Xue Y, Liao X, Carin L, Krishnapuram B (2007) Multi-task learning for classification with Dirichlet process priors. Journal of Machine Learning Research 8:35–63

Yu K, Bi J, Tresp V (2006) Active learning via transductive experimental design. In: International Conference on Machine Learning, New York, pp 1081–1088