Optimal experimental design in a hierarchical setting for probabilistic choice models

Adriana Birlutiu
Radboud University Nijmegen
20 PUBLICATIONS 63 CITATIONS
SEE PROFILE

Tom Heskes
Radboud University Nijmegen
217 PUBLICATIONS 2,602 CITATIONS
SEE PROFILE

Optimal experimental design in a hierarchical setting for probabilistic choice models

Adriana Birlutiu and Tom Heskes
Institute for Computing and Information Sciences
Radboud University Nijmegen
Nijmegen, The Netherlands
{a.birlutiu,t.heskes}@science.ru.nl

Abstract

We propose a new criterion for experimental design in the context of preference learning. This new criterion makes direct use of the data available from a group of subjects for which the preferences were already learned. Furthermore, we show the connections between this criterion and the standard criteria used in experimental design. Empirical results on a real audiological data set, show a factor of two speed-up for learning user preferences relative to random selection.

1 Introduction

Learning user preferences appears in many contexts. Consider, for example, the case in which the parameters of a (medical) device have to be tuned such as to adapt them optimally to a user's preferences. In order to do this, we learn user's preferences by means of experiments. However, in many cases, especially the one mentioned above, this is a tedious process. To reduce the costs, in terms of time invested and user burden, we would like to present to the user those experiments which give the most information about his/her preferences; we are thus in the context of optimal experimental design [11].

The goal of optimal design is to select experiments such that their outcomes give information for making a model that maximizes some criterion of accuracy. One criterion is the accuracy with which the parameters of the model can be estimated, which, in the Bayesian context, is equivalent to the reduction in the entropy of the posterior distribution (over the model parameters) that results from the outcome of the experiment, the so-called *D-optimal criterion* [2].

Active learning [4, 8, 5] is the equivalent of experimental design in the context of supervised learning. In this scenario, the learning algorithm selectively samples the unlabeled data to achieve high performance with relatively small training data. *Query-by-Committee* [8] is a method for active learning, which selects examples that have maximum disagreement amongst an ensemble of hypotheses.

Coming back to the problem of learning preferences, assume that we have available preference responses to some experiments from a group of people. We want to efficiently learn the preferences of a new person in as few experiments as possible, possibly by making use of the available data from the other subjects. One way to do this is to select those experiments for which the other subjects disagree the most. This is related to the *Query-by-Committee* method mentioned above, with the difference that the group of subjects, for which we already learned preferences, plays the role of the ensemble of hypotheses. Using this idea, we developed a criterion for optimal experimental design that makes use of the judgements of other subjects. We show that this new criterion is connected to the standard *D-optimal criterion* and, furthermore, it has several advantages due to its interpretation and simplicity.

The paper is organized as follows. Section 2 is about criteria for optimal selection of experiments; we start with a short presentation of the probabilistic choice models used; furthermore we present a way to gather the data from the other subjects and how to make use of it in a new criterion for experimental design; we show the connection with the standard D-optimal criterion and other approximations of it. Experimental results on a real audiological data set are shown in Section 3. Conclusions and directions for future research are presented in Section 4.

2 Criteria for optimal experimental design

2.1 Probabilistic choice models

Humans are very good in comparing options and expressing a preference for one of them. Therefore, in many settings, preferences are learned from experiments in which the person expresses a choice for one of the presented options. Let us consider probabilistic choice models of the form

$$p(k;\boldsymbol{x},\boldsymbol{\theta}) = \frac{\exp\left[\sum_{j=1}^{n} A_{kj}(\boldsymbol{x})\theta_j\right]}{Z(\boldsymbol{\theta})}\ , \tag{1}$$

with

$$Z(\boldsymbol{\theta}) \equiv \sum_{k=1}^{m} \exp\left[\sum_{j} A_{kj}(\boldsymbol{x})\theta_j\right]\ .$$

$\boldsymbol{\theta}$ is an n-dimensional vector of parameters, A is a function which extracts features of the input $\boldsymbol{x}$ related to option k, and there are m different options. $p(k;\boldsymbol{x},\boldsymbol{\theta})$ stands for the probability that a subject with parameters $\boldsymbol{\theta}$ prefers option k when given input $\boldsymbol{x}$. For future reference we define the derivatives of the log probabilities

$$\boldsymbol{g}(k;\boldsymbol{x},\boldsymbol{\theta}) \equiv \frac{\partial \log p(k;\boldsymbol{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\ , \qquad H(k;\boldsymbol{x},\boldsymbol{\theta}) \equiv \frac{\partial^2 \log p(k;\boldsymbol{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\ . \tag{2}$$

For $m = 2$ we have paired-comparison experiments and the model reduces to a logistic sigmoid function, also known as the Bradley-Terry model [3]. For $m > 2$ we have multiclass classification experiments and the model is a softmax function.

In order to learn a subject's preferences, we treat the vector of parameters $\boldsymbol{\theta}$ as a random variable. Before taking into account the information from the actual experiments performed with the subject, information available from other sources is incorporated in a prior distribution. After performing an experiment and observing its outcome, we compute, using Bayes' rule, the posterior distribution over $\boldsymbol{\theta}$. To keep things simple, we start with a Gaussian prior. Because the product between a Gaussian prior and the likelihood defined in Equation (1) is not a Gaussian, we approximate the posterior distribution to a Gaussian. There are several alternatives for doing this approximation: Laplace's method [12], Assumed Density Filtering [15], Expectation Propagation [15]. The choice of the approximation technique does not have too much influence on what follows.

2.2 Hierarchical modeling

Suppose that we have available preference responses to some experiments from a group of people (assume that we have M subjects, each of them with his/her own set of experiments and responses). We want to make use of this data, when learning the preferences of a new person. For this, we use hierarchical modeling [10, 9] to derive a method for gathering data from previous subjects in a prior for a new subject. The inference problems for each subject are coupled by giving them the same prior, i.e., we set $P(\boldsymbol{\theta}_i) = G(\boldsymbol{\theta}_i;\boldsymbol{\mu},\Sigma)$ a Gaussian prior with the same $\boldsymbol{\mu}$ and Σ for all subjects. The posterior for each subject is assumed to be (close to) a Gaussian with mean $\boldsymbol{\theta}_i^*$ and variance V_i. We would like to find the prior mean and variance that maximize the likelihood of all data.

The level II maximum likelihood values for the prior mean $\boldsymbol{\mu}$ and the prior variance Σ can be found by applying Expectation-Maximization algorithm (see e.g., [9]), which reduces in this case to the iteration, till convergence, of the following equations:

$$\boldsymbol{\mu} = \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{\theta}_i^*$$

$$\Sigma = \frac{1}{M}\sum_{i=1}^{M}(\boldsymbol{\theta}_i^* - \boldsymbol{\mu})(\boldsymbol{\theta}_i^* - \boldsymbol{\mu})^T + \frac{1}{M}\sum_{i=1}^{M} V_i \tag{3}$$

where $\boldsymbol{\theta}_i^*$ and V_i are the posterior mean and variance for subject i computed based on the previous prior mean and variance. The first term in the righthand side of Equation (3) measures the variance between the most probable estimates for different subjects, the second term the variance of the probabilities $P(\boldsymbol{\theta}_i)$ around these most probable estimates, averaged over all the subjects. Thus, we can make use of the available data from the group of other subjects, by taking as the prior of a new subject, the Gaussian $G(\boldsymbol{\theta}_i; \boldsymbol{\mu}, \Sigma)$.

2.3 Whose side you're on?

Furthermore, we want to make use of the data available from other subjects also when selecting the experiments to perform with a new subject i. Subject i starts off with a prior learned from the other subjects as explained in Section 2.2. The goal is to come up with a criterion for experiment selection that makes direct use of the judgements of the other subjects. A very simple and straightforward option would be to take those experiments for which the other subjects disagree the most, according to their responses given to the experiments. However, we will see further in Section 3, that this is not good enough. We propose instead the following criterion:

$$I_{\text{committee}}(\boldsymbol{x}) = \frac{1}{M-1}\sum_{j\neq i}\sum_{k}\bar{p}_{\backslash i}(k;\boldsymbol{x})\log\left[\frac{\bar{p}_{\backslash i}(k;\boldsymbol{x})}{p_j(k;\boldsymbol{x})}\right] - \sum_{k}\bar{p}_{\backslash i}(k;\boldsymbol{x})\log\left[\frac{\bar{p}_{\backslash i}(k;\boldsymbol{x})}{p_i(k;\boldsymbol{x})}\right] \,, \tag{4}$$

with $p_j(k;\boldsymbol{x}) = p(k;\boldsymbol{x},\boldsymbol{\theta}_j^*)$, where $\boldsymbol{\theta}_j^*$ is the maximum posterior solution for subject j. $\bar{p}_{\backslash i}(k;\boldsymbol{x})$ is the logarithmic average over all $j \neq i$ and is defined as follows:

$$\bar{p}(k;\boldsymbol{x}) = \frac{1}{\Gamma(\boldsymbol{x})}\exp\left[\int d\boldsymbol{\theta}\, P(\boldsymbol{\theta})\log p(k;\boldsymbol{x},\boldsymbol{\theta})\right] \,,$$

with

$$\Gamma(\boldsymbol{x}) = \sum_k \exp\left[\int d\boldsymbol{\theta}\, P(\boldsymbol{\theta})\log p(k;\boldsymbol{x},\boldsymbol{\theta})\right] \,. \tag{5}$$

The criterion proposed in Equation (4), is similar to the one proposed in [14, 13], with the difference that in our case we are in a different setting, preference learning, and that we have real subjects as members of the committee.

Because of the log linear form in Equation (1), we immediately find

$$\bar{p}(k;\boldsymbol{x}) = p(k;\boldsymbol{x},\boldsymbol{\mu}) \;\text{ with }\; \boldsymbol{\mu} = \int d\boldsymbol{\theta}\, P(\boldsymbol{\theta})\boldsymbol{\theta} \,.$$

In words, the criterion is the disagreement between the other subjects (as measured through the average Kullback-Leibler divergence) minus the disagreement between the current user and the (geometric) average of all the other subjects. The first term favors experiments on which the other subjects disagree. The intuition behind the negative term is that it makes less sense to present experiments on which the subject already formed an opinion different from that of the other subjects. In other words, the most interesting experiments are those on which the other subjects disagree, with the current subject (still) in the middle. Hence the title of the section.

2.4 Connection with D-optimal criterion

A popular criterion in experimental design is the expected log determinant of the variance of the approximation. Define $V(k,\boldsymbol{x})$ to be the new variance after presenting $\boldsymbol{x}$ and observing response k. The so-called D-optimal criterion then reads

$$I_{\text{det}}(\boldsymbol{x}) = -\sum_k p(k;\boldsymbol{x})\log\det V(k,\boldsymbol{x}) + \log\det V \,,$$

with $p(k;\boldsymbol{x})$ the probability that the subject indeed gives response k when presented $\boldsymbol{x}$, and where we subtracted the log determinant of the current variance. The best experiment is the one that maximizes $I_{\text{det}}(\boldsymbol{x})$. To evaluate this, in principle we would have to recompute the variances as well as the probabilities $p(k;\boldsymbol{x})$ for all combinations of presentations $\boldsymbol{x}$ and observations k.

Lemma 1. *In a first order approximation, assuming that $V(k,\boldsymbol{x})$ is close to V, we can simplify*

$$I_{det}(\boldsymbol{x}) \approx \sum_k p(k;\boldsymbol{x},\boldsymbol{\theta}^*)\boldsymbol{g}(k;\boldsymbol{x},\boldsymbol{\theta}^*)^T V \boldsymbol{g}(k;\boldsymbol{x},\boldsymbol{\theta}^*) \; ,$$

where $\boldsymbol{\theta}^*$ is the maximum posterior solution. (Proof in the appendix.)

The criterion from Equation (4) (the disagreement in the committee making statements about $\boldsymbol{x}$ induced by the uncertainty in the posterior) gives us another interpretation of the criterion $I_{\text{det}}(\boldsymbol{x})$.

Lemma 2. *In a lowest order approximation we can make the connection with the standard D-optimal criterion I_{det}*

$$I_{committee}(\boldsymbol{x}) = \frac{1}{2}\sum_k p(k;\boldsymbol{x},\boldsymbol{\mu})\boldsymbol{g}(k;\boldsymbol{x},\boldsymbol{\mu})^T \tilde{V} \boldsymbol{g}(k;\boldsymbol{x},\boldsymbol{\mu}) \; , \tag{6}$$

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

$$\tilde{V} \equiv \frac{1}{M-1}\sum_{j\neq i}(\boldsymbol{\theta}_j^* - \boldsymbol{\mu})(\boldsymbol{\theta}_j^* - \boldsymbol{\mu})^T - (\boldsymbol{\theta}_i^* - \boldsymbol{\mu})(\boldsymbol{\theta}_i^* - \boldsymbol{\mu})^T \; .$$

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

$$\begin{aligned}\sum_k p(k;\boldsymbol{x},\boldsymbol{\mu}) \log\left[\frac{p(k;\boldsymbol{x},\boldsymbol{\mu})}{p(k;\boldsymbol{x},\boldsymbol{\theta}^*)}\right] &\approx -\frac{1}{2}\sum_k p(k;\boldsymbol{x},\boldsymbol{\mu})(\boldsymbol{\theta}^* - \boldsymbol{\mu})^T H(k;\boldsymbol{x},\boldsymbol{\mu})(\boldsymbol{\theta}^* - \boldsymbol{\mu}) \\ &= \frac{1}{2}\sum_k p(k;\boldsymbol{x},\boldsymbol{\mu}) \left[(\boldsymbol{\theta}^* - \boldsymbol{\mu})^T \boldsymbol{g}(k;\boldsymbol{x},\boldsymbol{\mu})\right]^2 \; , \end{aligned} \tag{7}$$

the first order term canceled since (see Equation (9) from appendix) $\sum_k p(k;\boldsymbol{x},\boldsymbol{\theta})\boldsymbol{g}(k;\boldsymbol{x},\boldsymbol{\theta}) = 0$. Using Equation (7) we obtain the result stated in the lemma.

□

So $I_{\text{committee}}(\boldsymbol{x})$ is somehow reminiscent of the “standard” D-optimal criterion with the following main differences.

1. The gradients $\boldsymbol{g}(k;\boldsymbol{x})$ are evaluated at the prior mean $\boldsymbol{\mu}$ instead of at the current posterior mean $\boldsymbol{\theta}^*$. This effect could be small since $\boldsymbol{\theta}^*$ is still close enough to $\boldsymbol{\mu}$ for a sufficiently accurate approximation of the gradients, in particular at the start when selecting the right experiments is most important.
2. The current posterior variance V is replaced by $\tilde{V}$. The idea here is that the effect of the precise weighting of the gradients is not tremendously important. For example, in practice I_{trace} from Equation (8), which corresponds to a weighting V^2 (Lemma 3), works about as well as I_{det}, which corresponds to a weighting V. And again, at the start $\tilde{V}$ is pretty close to V, since $\boldsymbol{\theta}^*$ is then still close to $\boldsymbol{\mu}$ and V to the prior variance Σ.

2.5 Other design criteria

Alternatively, we can consider other criteria, which are basically approximations of the standard D-optimal criterion $I_{\text{det}}(\boldsymbol{x})$. An option is the weighted Kullback-Leibler divergence between the current Gaussian approximation and the one after presenting $\boldsymbol{x}$ and observing k, I_{kl}. Again, we would like to maximize $I_{\text{kl}}(\boldsymbol{x})$ to find the “best” experiment. In a first order approximation, assuming that $V(k,\boldsymbol{x})$ is close to V, it can be proved that

$$I_{\text{kl}} \approx I_{\text{det}}$$

i.e., the two criteria are indistinguishable.

Instead of the log determinant, we can also take the trace of the variance, the so-called A-optimal criterion, as our criterion for selecting the best experiment. We define

$$I_{\text{trace}}(\boldsymbol{x}) = -\sum_k p(k;\boldsymbol{x})\,\mathrm{Tr}\,V(k,\boldsymbol{x})\,. \tag{8}$$

Lemma 3. *In a first order approximation, the trace criterion boils down to*

$$I_{trace}(\boldsymbol{x}) = \sum_k p(k;\boldsymbol{x},\boldsymbol{\theta}^*)\boldsymbol{g}(k;\boldsymbol{x},\boldsymbol{\theta}^*)^T V^2 \boldsymbol{g}(k;\boldsymbol{x},\boldsymbol{\theta}^*)\,.$$

3 Experiments

We evaluate different criteria on a data set of audiological experiments described in [1]. The data set consists of predictions of sound quality of 14 normal hearing and 18 hearing impaired persons. Each person was subjected to 576 paired-comparison tests of the form $(\boldsymbol{x}_1, \boldsymbol{x}_2, k)$, where $k = \{1, 2\}$ denotes whether sound sample $\boldsymbol{x}_1$ or $\boldsymbol{x}_2$ was preferred by the patient, respectively. We used this data set to address the following two questions:

1. Can we use the already learned preferences of other subjects to better learn the preferences of the current subject?
2. Can we learn faster by optimally selecting the experiments to present to a subject?

In a simulation, one subject was left out, and the hierarchical method, described in Section 2.2, was used to gather data from the rest of the subjects in a probability distribution, which was used as the starting prior for the left-out subject. The data set for the left-out subject, was split into training (used for learning preferences) and testing (the accuracy of the predictions on the test data was used as a measure of how much we learned about subject's preferences). For each subject, we averaged the results across several splits using cross-validation. Furthermore, the results were averaged over all subjects.

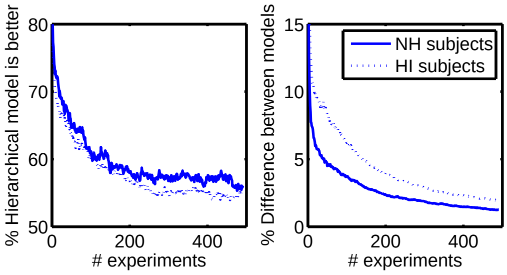

Figure 1: Left: percentage of the number of times the prediction accuracy using the learned prior is better than the prediction accuracy with a flat prior. Right: percentage of the number of predictions on which the two models (with the learned and with a flat prior) disagree. The NH and HI labels refer to the simulations on the data set from normal-hearing and hearing-impaired subjects, respectively.

In order to answer the first question, we compared the performances obtained using the hierarchical prior versus a flat prior which assumes no information about subject's preferences. We made predictions for the outcomes of the experiments from the test data, using the model from Equation (1); where $\boldsymbol{\theta}$ is the mean of the posterior distribution of a subject which resulted either by starting with the hierarchical prior or with a flat prior. The righthand side of Figure 1, gives the percentage of predictions on which the two models (the one with the hierarchical and the one with flat prior) disagree, with respect to the total number of predictions made. The lefthand side of Figure 1, shows the percentage of correct predictions made using the hierarchical prior, with respect to the number of predictions on which the two models disagree. Especially in the beginning of the learning process, with few experiments, the model with a prior learned from the community of other subjects outperforms the model with a flat prior. Thus, we can affirmatively answer the first question.

In order to answer the second question, we compared the performances obtained by random versus optimal selection of experiments. For each subject, we started with the prior learned from the other subjects, and updated this prior based on the information from experiments which were selected either random or optimal. The optimal selection was implemented using $I_{\text{committee}}$ criterion. In practice, the committee criterion performs about the same as the D-optimal criterion, or any of its approximations. We computed the number of experiments needed by random selection to get the same accuracy on the test data as with the optimal selection. Figure 2 shows that, indeed by optimally selecting experiments, the preferences can be learned faster. We implemented a variant of optimal selection where we choose experiments according to the difference between the number of subjects which preferred the first alternative and the number of subjects which preferred the second alternative. The experiments for which this difference is small are considered hard to predict. The plots show that just presenting those experiments which are hard to predict according to the responses given by the other subjects, does not perform much better than random selection.

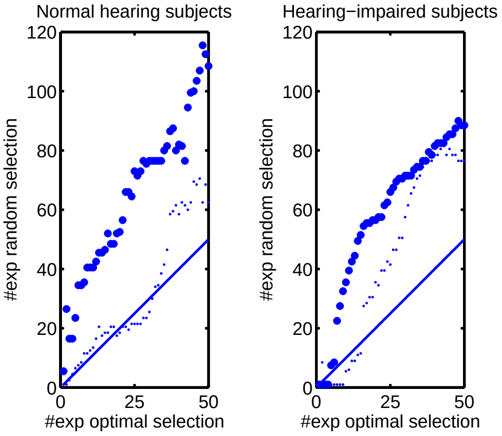

Figure 2: The number of listening experiments needed using random selection (on the y-axis) to get the same accuracy as with the optimal selection (on the x-axis). The optimal experiment selection is implemented by presenting those experiment which are hard to predict according to other subjects responses (small dots) and actively using the pool criterion (large dots).

Figure 3 shows the connections between the criteria for experiment selection discussed in this paper. The plots are shown for one normal hearing subject. We started with the prior learned from the group of other normal hearing subjects, and made updates of this prior by taking into account the information from 10 randomly selected listening experiments. At this point, we computed the scores of each of the criteria for one randomly chosen listening experiment. We made scatter plots of the ranks of these scores; in the title of each plot, we wrote the Spearman correlation coefficient between the two criteria for which the plots are displayed. From the first plot from the left, we see that the approximation of the D-optimal criterion, as stated in Lemma 1, is very accurate. The fifth plot from the left, shows the scores computed using the I_{det} criterion and the criterion which selects the experiments which are hard to predict; as expected, these two are not connected. Furthermore, in the rest of the plots, we can see that the different criteria are indeed strongly correlated, as predicted from the theory in Section 2.

4 Conclusions and discussions

We discussed and analyzed criteria for optimal experimental design, and showed that for the probabilistic choice model introduced in section 2.1 they are all connected to the standard D-optimal criterion. A direction for future work is to extend this analysis to other types of models. We proposed a new criterion that makes direct use of the judgements of other subjects. The new criterion, in practice, works about as well as (any other sensible approximation of) D-optimal experimental design. The advantage of this new criterion could be in the interpretation and the (relative) simplicity: it is only based on probabilities computed from maximum a posteriori solutions, i.e., there is no

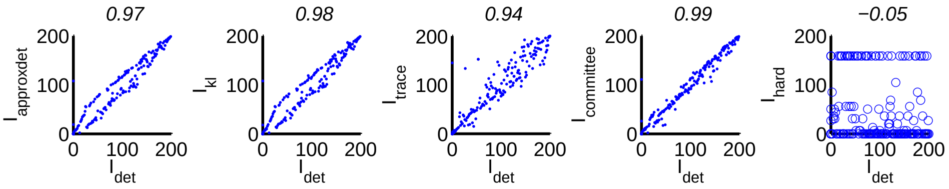

Figure 3: Scatter plots for different criteria for experiment selection. Title of each plot: the Spearman correlation coefficient between the two criteria for which the plot is displayed. See the main text for further details.

need to keep track of variances or to compute gradients. Furthermore, it is efficient to compute since the first term as well as the average can be computed beforehand.

References

[1] K.H. Arehart, J.M. Kates, M.C. Anderson, and L.O. Harvey. Effects of noise and distortion on speech quality judgments in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 122:1150–1164, 2007.

[2] M.P.F. Berger. D-optimal sequential sampling designs for item response theory models. *Journal of Educational and Behavioral Statistics*, 19:43–56, 1994.

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

[4] D.A. Cohn, Z. Ghahramani, and M.I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[5] S. Dasgupta and D.J. Hsu. Hierarchical sampling for active learning. *Twenty-Fifth International Conference on Machine Learning (ICML)*, 2008.

[6] V.V. Fedorov. *Theory of Optimal Experiments*. Academic Press, 1972.

[7] I. Ford and S.D. Silvey. A sequentially constructed design for estimating a nonlinear parametric function. *Biometrika*, 67:381–388, 1980.

[8] Y. Freund, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. In *Machine Learning*, pages 133–168, 1997.

[9] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC, July 2003.

[10] D. V. Lindley and A. F. M. Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(1):1–41, 1972.

[11] D.V. Lindley. *Bayesian Statistics - A Review*. SIAM, Philadelphia, 1972.

[12] D.J.C. Mackay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA, 2002.

[13] Andrew McCallum and Kamal Nigam. Employing em and pool-based active learning for text classification. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, pages 350–358, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[14] Prem Melville, Stewart M. Yang, Maytal Saar-tsechansky, and Raymond Mooney. Active learning for probability estimation using jensen-shannon divergence. In *In Proceedings of the European Conference on Machine Learning (ECML-05*, pages 268–279, 2005.

[15] T. P. Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, M.I.T., 2001.

[16] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, New York, NY, USA, 1992. ACM.

[17] K. Yu, J. Bi, and V. Tresp. Active learning via transductive experimental design. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 1081–1088, New York, 2006.

Appendix

Lemma 4. *For any $\boldsymbol{x}$ and $\boldsymbol{\theta}$ we have the following relation between second derivatives and first derivatives.*

$$\sum_k p(k;\boldsymbol{x},\boldsymbol{\theta})H(k;\boldsymbol{x},\boldsymbol{\theta}) = -\sum_k p(k;\boldsymbol{x},\boldsymbol{\theta})\boldsymbol{g}(k;\boldsymbol{x},\boldsymbol{\theta})\boldsymbol{g}(k;\boldsymbol{x},\boldsymbol{\theta})^T \ ,$$

with $\boldsymbol{g}$ and H the first and second derivatives defined in Equation (2).

Proof. We use shorthand notation $p_k = p(k;\boldsymbol{x},\boldsymbol{\theta})$, $g_{kj} = g_j(k;\boldsymbol{x},\boldsymbol{\theta})$, etc., omitting the dependencies on $\boldsymbol{x}$ and $\boldsymbol{\theta}$. Write $u_k \equiv \sum_j A_{kj}\theta_j$ and thus $\log p_k = u_k - \log Z$. Then it is easy to see that

$$g_{kj} = A_{kj} - \frac{\partial \log Z}{\partial \theta_j} \ , \qquad h_{k,ij} = -\frac{\partial^2 \log Z}{\partial \theta_i \partial \theta_j} = h_{ij} \ ,$$

i.e., the second derivative is in fact independent of k. Furthermore

$$\begin{aligned} \frac{\partial \log Z}{\partial \theta_j} &= \frac{1}{Z}\frac{\partial Z}{\partial \theta_j} = \sum_k p_k A_{kj} \\ \frac{\partial^2 \log Z}{\partial \theta_i \partial \theta_j} &= \sum_k A_{kj}\frac{\partial p_k}{\partial \theta_i} = \sum_k A_{kj} p_k g_{ki} = \sum_k p_k A_{ki} A_{kj} - \sum_k p_k A_{ki} \sum_{k'} p_{k'} A_{k'j} \ , \end{aligned}$$

and thus

$$\begin{aligned} g_{kj} &= A_{kj} - \sum_{k'} p_k' A_{k'j} \\ h_{k,ij} &= -\sum_{k'} p_{k'} A_{k'i} A_{k'j} + \sum_{k'} p_{k'} A_{k'i} \sum_{k''} p_{k''} A_{k''j} \ . \end{aligned} \tag{9}$$

We then have

$$\begin{aligned} \sum_k p_k h_{k,ij} &= -\sum_k p_k A_{ki} A_{kj} + \sum_k p_k A_{ki} \sum_{k'} p_{k'} A_{k'j} \\ &= -\sum_k p_k \left(A_{ki} - \sum_{k'} p_{k'} A_{k'i}\right)\left(A_{kj} - \sum_{k'} p_{k'} A_{k'j}\right) = -\sum_k p_k g_{ki} g_{kj} \ . \end{aligned}$$

□

Lemma 1. *In a first order approximation, assuming that $V(k,\boldsymbol{x})$ is close to V, we can simplify*

$$I_{\textit{det}}(\boldsymbol{x}) \approx \sum_k p(k|\boldsymbol{x},\boldsymbol{\theta}^*)\boldsymbol{g}(k;\boldsymbol{x},\boldsymbol{\theta}^*)^T V \boldsymbol{g}(k;\boldsymbol{x},\boldsymbol{\theta}^*) \ .$$

Proof. In a first order approximation we have

$$V(k,\boldsymbol{x})^{-1} \approx V^{-1} - \left.\frac{\partial^2 \log p(k|\boldsymbol{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \ ,$$

where we ignored the change from the old $\boldsymbol{\theta}^*$ to a new maximum a posteriori solution depending on k and $\boldsymbol{x}$. Again making a first order expansion, assuming that $V(k,\boldsymbol{x})$ is close to V, we have

$$\log\det V(k,\boldsymbol{x})^{-1} \approx \log\det V^{-1} - \mathrm{Tr}\left[VH(k;\boldsymbol{x},\boldsymbol{\theta}^*)\right] \ .$$

The probability that the subject indeed gives the response k when presented $\boldsymbol{x}$ follows by integrating $p(k|\boldsymbol{x},\boldsymbol{\theta})$ over the current posterior:

$$p(k|\boldsymbol{x}) \approx \int d\boldsymbol{\theta}\, p(k;\boldsymbol{x},\boldsymbol{\theta})G(\boldsymbol{\theta};\boldsymbol{\theta}^*,V) \approx p(k;\boldsymbol{x},\boldsymbol{\theta}^*) + \frac{1}{2}\,\mathrm{Tr}\left[V \left.\frac{\partial^2 p(k;\boldsymbol{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right] \ ,$$

with $\boldsymbol{\theta}^*$ the maximum a posteriori solution and V the corresponding variance. Again in lowest order we can ignore the correction upon $p(k;\boldsymbol{x},\boldsymbol{\theta}^*)$ and arrive at

$$I_{\text{det}}(\boldsymbol{x}) \approx -\sum_k p(k|\boldsymbol{x},\boldsymbol{\theta}^*)\,\mathrm{Tr}\left[VH(k;\boldsymbol{x},\boldsymbol{\theta}^*)\right] \ .$$

Lemma 4 then gives the result. □