Spring 5-8-2019

Differential Estimation of Audiograms using Gaussian Process Active Model Selection

Trevor Larsen
Washington University in St. Louis

Washington University in St. Louis

McKelvey School of Engineering

Department of Computer Science

Thesis Examination Committee:

Dennis Barbour, Chair

Roman Garnett

Marion Neumann

Differential Estimation of Audiograms using Gaussian Process Active Model Selection

By

Trevor Jonathan Larsen

A thesis presented to the McKelvey School of Engineering of Washington University in St. Louis in partial fulfillment of the requirements for the degree of Master of Science

May 2019

St. Louis, Missouri

To my parents, for their unwavering support.

Acknowledgements

I would like to thank my research advisor and lab director, Dr. Barbour. He brought me into his lab as a junior, with little prior experience in machine learning. Under his mentorship and guidance, I've grown so much as an engineer, and internalized the power of perspective when trying to push the boundaries of the unknown.

Second, I would like to thank Dr. Garnett. Taking his Bayesian Methods in Machine Learning class completely transformed the way I viewed and understood probability, data and machine learning. His mentorship was critical in helping me derive the formulation of Bayesian Active Differential Selection in this thesis.

Third, I'd like to thank Dr. Neumann. Her CSE 200 class completely changed my life trajectory, when I decided to drop biomedical engineering and pre-med for computer science to pursue machine learning, and it was her that put me in touch with Dr. Barbour after taking her machine learning class. She has been incredibly supportive of me ever since.

I would also like to thank everyone else in the Barbour lab, from the time I joined to the present. James and David, you both inspired me to continue the awesome work you guys had started in the lab. Steven and Braham, you both taught me so much, and helped me grow as a developer. Jonathan, working with you this past year has been a pleasure. Thank you for a great final year.

Finally, I'd like to thank my parents for your unwavering support. I love you both.

Table of Contents

List of Tables

List of Figures

Abstract

Differential Estimation of Audiograms using Gaussian Process Active Model Selection

By

Trevor Jonathan Larsen

Master of Science in Computer Science

Washington University in St. Louis, 2019

Research Advisor: Professor Dennis Barbour

Classical methods for psychometric function estimation either require excessive resources to perform, as in the method of constants, or produce only a low resolution approximation of the target psychometric function, as in adaptive staircase or up-down procedures. This thesis makes two primary contributions to the estimation of the audiogram, a clinically relevant psychometric function estimated by querying a patient's for audibility of a collection of tones. First, it covers the implementation of a Gaussian process model for learning an audiogram using another audiogram as a prior belief to speed up the learning procedure. Second, it implements a use case of Bayesian active model selection to determine whether two audiograms differ. Both algorithms were tested using audiometric data from the National Institute for Occupational Safety and Health (NIOSH).

1. Introduction

Psychometric Functions

A psychometric function is an inferential model applied to a detection or discrimination task. It models the relationship between a physical stimulus and a response from a human or animal subject. Unidimensional psychometric functions, known as psychometric curves (PCs), have received much attention in the literature. One of the first and most widespread methods for modelling PCs is the method of constant stimuli, developed by Gustav Fechner, and described in Elemente der Psychophysik (Fechner, 1860). The method samples a fixed number of stimuli from the input domain, often equally spaced. While accurate, the main drawback of this method is that it requires many stimuli. Newer methods have attempted adaptive approaches to overcome this inefficiency, by using prior subject responses to influence future stimulus delivery. Developments in this direction include up-down methods (Levitt, 1971), and parameter estimation by sequential testing (Taylor & Creelman, 1967).

A psychometric function can either be parametric or nonparametric, though the vast majority of historical models are parametric. A parametric model uses a function that can be uniquely identified by a set of parameters, such as α, the threshold intensity at which a specific fraction of stimuli are observed, and β the reciprocal of the derivative of the PC with respect to stimulus intensity at α. A nonparametric model is defined only by the input data. Examples of nonparametric models include splines, K-nearest neighbor methods, and Gaussian processes (GP).

Audiometry

One application of psychometric functions is audiometry. In pure-tone audiometry, subjects are presented with tone stimuli delivered at varying frequencies and intensities. This two-dimensional input domains makes approaches such as the method of constant stimuli particularly inefficient as it represents a two-dimensional grid search: an effective but inefficient algorithm. In 1944, the Hughson-Westlake (HW) algorithm was designed to assist in diagnosing hearing loss in soldiers who fought in World War II, due to the increased rates of noise-induced hearing loss in veterans caused by the war. A modified version is today used in the clinic for diagnosis. Due to the fact that the PC for hearing is sigmoid shaped, with tones having a high probability of being heard above some threshold intensity and a low probability of being heard below the threshold intensity, it is useful to find the threshold, which can be thought of as the middle of the sigmoid. The threshold-seeking algorithm proceeds along frequency by octaves, presenting tones in decreasing 10 dB increments or increasing 5 dB increments to find the threshold intensity for a given frequency. Once a tone is missed, the intensity is increased 5 dB. The algorithm terminates for each frequency after a set number of reversals (Carhart Raymond & Jerger James F., 1959; Hughson & Westlake, 1944). This method is therefore adaptive in intensity, though grid search in frequency. It is still in wide use today.

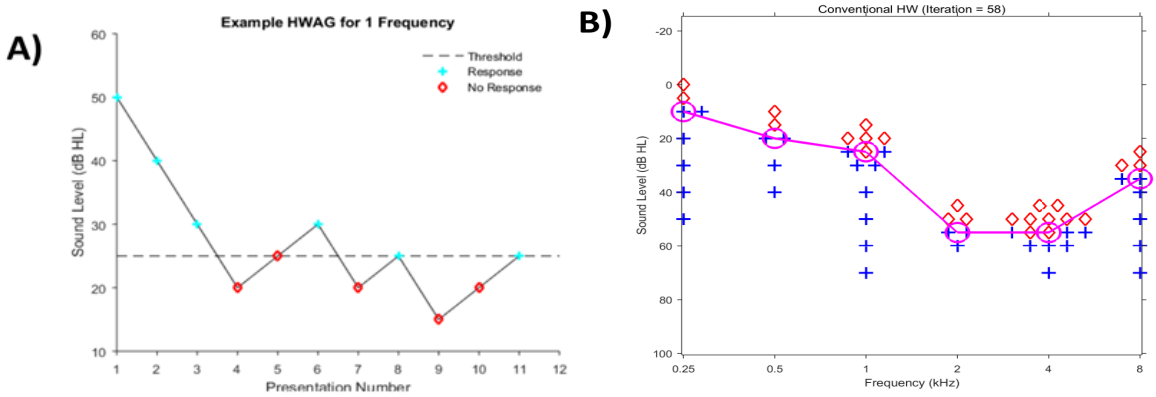


Figure 1. Visualization of HW procedure.

Left: threshold is found at a given frequency by decreasing intensity until a tone is missed, then increasing until it is detected, for 3 reversals. Right: This process is repeated at each octave frequency, with resulting thresholds connected using linear interpolation

The HW algorithm for audiometry has several issues that make it inefficient. First, the algorithm treats each octave independently, though thresholds are correlated across frequency. The algorithm begins each octave by delivering a tone at the same intensity. The algorithm could find the threshold more quickly by selecting initial tones for each octave that are closer to the threshold of nearby octaves. Second, the HW algorithm only gives the clinician data on where the threshold is located at octave frequencies. Because no data is collected between octave frequencies, the clinician can only guess at the shape of the threshold between octaves. Third, unlike the method of constant stimuli, the HW algorithm is designed to only give information about the location of the threshold at each frequency. It doesn't give information about the spread of the psychometric curve at individual frequencies as a function of intensity. Each of these problems is addressed by the Active Machine Learning Audiogram (AMLAG).

Prior Work

Prior work on using GPs for audiogram estimation began in 2015, using a variant of uncertainty sampling that selected points with maximum variance (Song et al., 2015). This work dramatically increased both the speed and accuracy of threshold audiogram estimation by at least an order of magnitude. A set of example plots from this work is included in Figure 2.

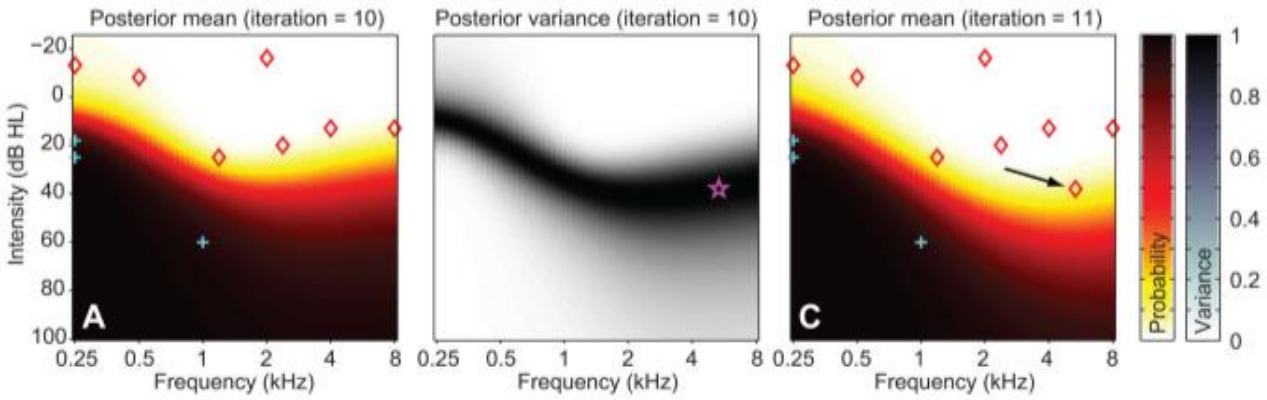
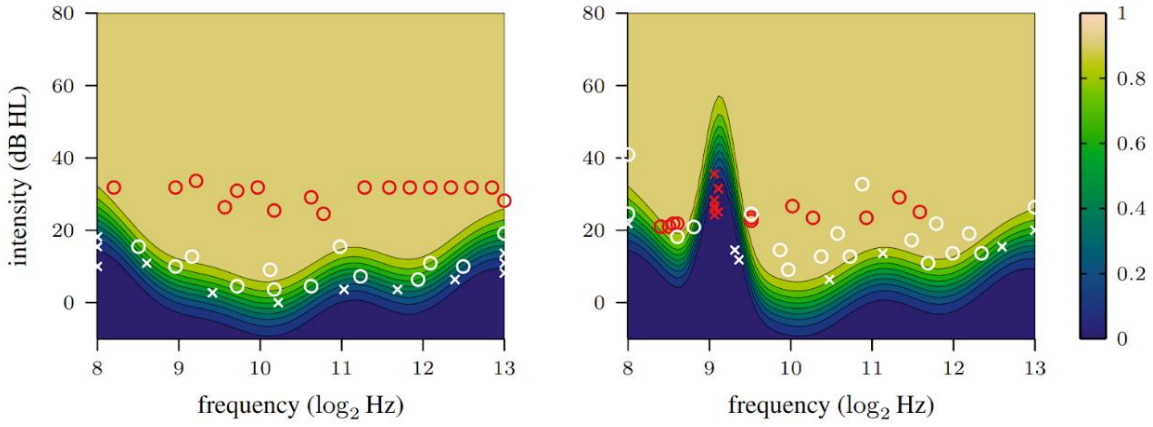


Figure 2. BALV Audiogram

Figure taken from (Song et al., 2015). Plots A and C show the posterior mean function at iterations 10 and 11, while plot B, the middle plot, shows the posterior variance. The star in plot B is the next point to be sampled, and occurs at the maximum variance value. Its effect on the posterior mean can be observed by comparing plots A and C.

Bayesian Active Model Selection (BAMS) uses active learning to distinguish which of a number of predetermined models best explains a function being actively observed, and was introduced with an application for automated notch-shaped hearing loss detection in machine learning audiograms (J. Gardner et al., 2015). A depiction of BAMS can be seen in Figure 3.



(a) Normal hearing model ground truth. (b) Notch model ground truth.

Figure 3. BAMS Example

Figure taken from (J. Gardner et al., 2015). Circles are heard tones, crosses are unheard tones. Red tones are points delivered using BAMS, while white tones are delivered using the GP audiogram method described in (J. R. Gardner, Song, Weinberger, Barbour, & Cunningham, 2015). In plot a, the red tones are spread out evenly over the frequency domain to search for a notch, whereas they are clustered together at the notch in plot b.

While both of the aforementioned papers explored the use of machine learning audiometry in one ear, this approach was extended to exploit the shared variance between ears using a conjoint audiogram (Barbour et al., 2018; DiLorenzo, 2017). This approach was a dramatic improvement, learning the audiogram for both ears in just as much time as, or faster than learning a single ear individually. Performance of conjoint audiometry can be observed in Figure 4.

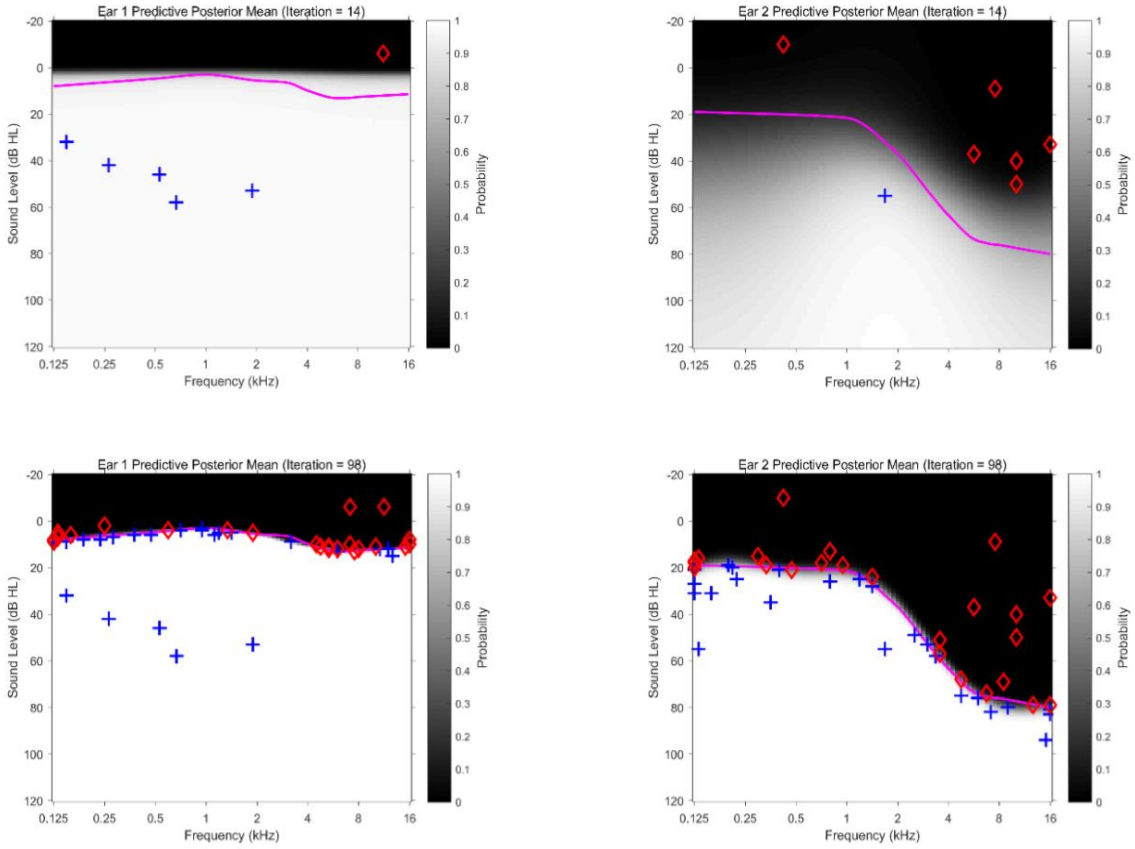


Figure 4. Conjoint Example

Figure taken from (DiLorenzo, 2017). Blue plusses represent heard tones, while red diamonds represent unheard tones. The purple line represents the ground truth threshold. Left plots represent ear 1, while right plots represent ear 2. Top plots are the posterior mean estimation at iteration 14, while bottom plots are the posterior estimation at iteration 98. After 14 iterations, ear 1 has been approximated to within a few dB of the threshold, while ear 2, with hearing loss, is getting close but still needs more tones to converge. After 98 iterations, the estimate matches ground truth almost completely.

Current methods for audiogram estimation begin without using prior information from the patient. The underlying psychometric function that the audiogram is meant to estimate is correlated with previous audiometric tests that the patient has taken, as well as other audiometric tests more generally. In the same way that conjoint audiogram estimation utilizes covariance between ears to reduce the number of tones needed to estimate the underlying function, we propose an algorithm for active differential estimation that is able to utilize the covariance across audiometric tests. We also propose to use BAMS to determine whether two audiograms are sampled from the same distribution.

2. General Methods

This thesis has two aims. The first aim is to introduce a framework that we are calling Bayesian Active Differential Estimation (BADE) for estimating a psychometric function using a prior test for that function as an input. Specifically, we will do this for the case of the audiogram. This allows for faster estimation of the new audiogram by exploiting correlation between audiograms, and is done by expanding upon the conjoint estimation framework. The second aim of this thesis is to develop a framework we are calling Bayesian Active Differential Selection (BADS). The goal of this framework is to determine whether or not a new estimated model differs from a prior estimated model, using BAMS (J. Gardner et al., 2015). While we are using the audiogram as a use case, this methodology is general and can be expanded to other domains as well. The BADE and BADS algorithms were tested using ground truth audiograms generated from the National Institute for Occupational Safety and Health (NIOSH) occupational hearing database (Masterson et al., 2013).

NIOSH Database Median Audiogram Generation

To test the BADE and BADS algorithms, we generated ground truth data from the NIOSH audiometric testing (Masterson et al., 2013). Each entry in the NIOSH database includes 7 threshold intensity values per ear at 500 Hz, 1000 Hz, 2000 Hz, 3000 Hz, 4000 Hz, 6000 Hz and 8000 Hz. We classified each ear into one of seven categories of hearing loss, based on the pure-tone average (PTA) of each ear, calculated by taking the mean of the threshold values at 500 Hz, 1000 Hz, and 2000 Hz. These categories are indicated in Table 1.

Table 1

Hearing loss classification using pure-tone average (Clark, 1981)

Degree of hearing loss	Hearing loss range (dB HL)
Normal	-10 to 15
Slight	16 to 25
Mild	26 to 40
Moderate	41 to 55
Moderately severe	56 to 70
Severe	71 to 90
Profound	91+

Within each category, we generated a canonical audiogram by taking the median threshold value at each frequency. Ground truths were then extrapolated from these threshold values using a cubic spline interpolation first in the frequency domain, followed by creating a sigmoid in the intensity domain. The resulting ground truth audiograms are presented in Figure 5.

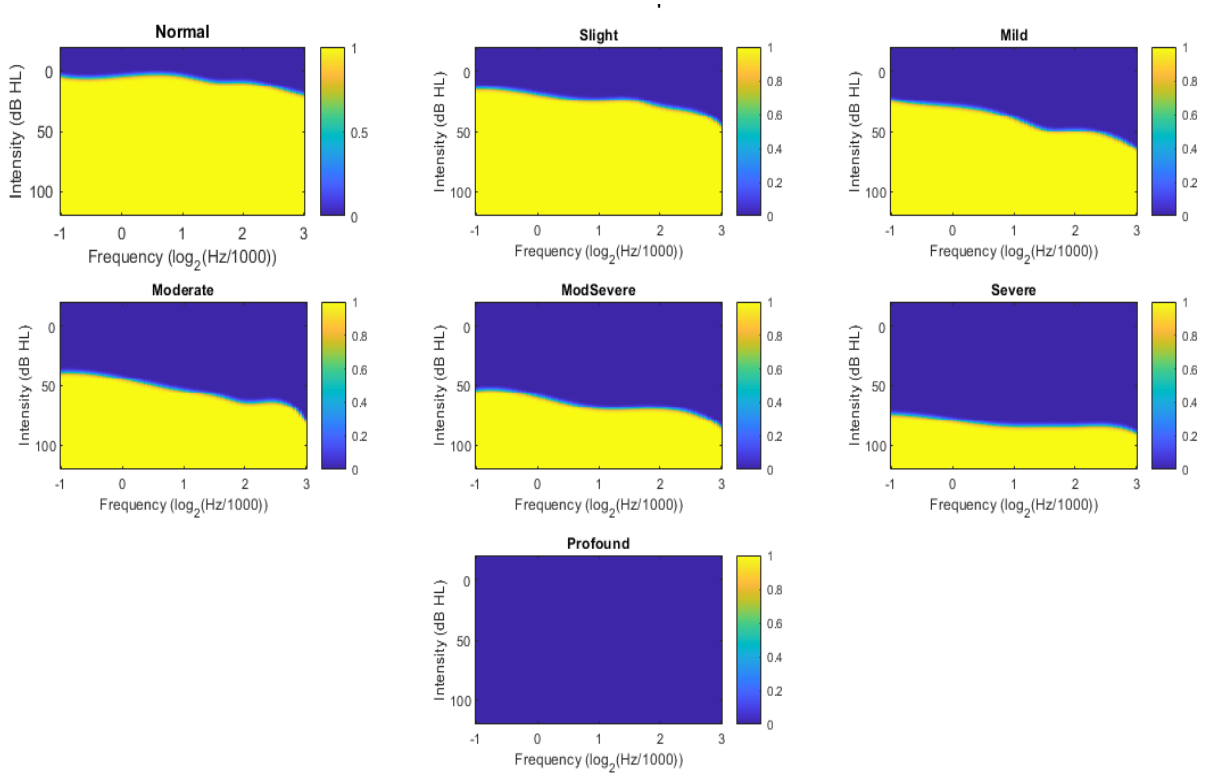


Figure 5. Ground truth audiograms of each hearing loss class.

Aim 1: Bayesian Active Differential Estimation

In this section, we extend the AMLAG framework to include information from prior audiometric tests. This extra prior information is used to exploit the covariance between the prior estimated audiogram and the new audiogram in order to estimate the new audiogram in a smaller number of new tones.

Feature Space

We begin by taking the data from the prior audiogram X_1, which is made up of individual observations $x_{1i} = (f_i, I_i)$, and augment this data matrix with a new feature column representing which test the data point come from $t_i = 1$ for all prior observations such that $x_{1i} = (f_i, I_i, t_i = 1)$. All new observations will be constrained such that $t_i = 2$.

Mean Function

In this case, we use a constant mean function $\mu(x_i) = c \forall x_i \in X$. While audiograms do not necessarily have a constant mean, prior research has shown that a constant mean function is sufficient for audiograms, as the covariance function captures the shape of the audiogram in the posterior distribution (Barbour et al., 2018; DiLorenzo, 2017; J. Gardner et al., 2015; Song et al., 2015).

Covariance Function or Kernel

For our covariance function, we use a composite function made up of three parts. Recall that the domain of our input X has three features: frequency, intensity, and the binary variable 'test,' denoting whether the given data point is from the old audiogram or the new audiogram. The first dimension, frequency, uses an isotropic squared exponential kernel:

$$K_f(x_f, x'_f) = \lambda^2 e^{-\frac{(x_f - x'_f)^T (x_f - x'_f)}{2\gamma^2}},$$

where λ^2 is a scale factor, and ℓ^2 is a length scale, parameterizing how close two values need to be in order to covary (i.e. closer points in frequency space will covary more than points that are farther apart). This enforces the idea that an audiometric function should be continuous and smooth across the frequency domain.

The second dimension, intensity, uses an isotropic linear kernel:

$$K_I(x_I, x'_I) = \frac{1}{\ell^2} x_I^T x_I$$

The probability of detecting stimuli should be low at low intensities, and scale to near 100% probability at high intensities. At any given frequency, we expect to observe sigmoidal behavior for the likelihood function. This is achieved using the above linear kernel with a cumulative

Gaussian function for our likelihood. This likelihood function is explained in more detail in the next section.

The final dimension, the test dimension, is a binary variable for a categorical domain. Thus, for the test dimension we use a discrete covariance kernel.

$$K_t(x_t, x_t') = \begin{cases} s_{11} & \text{if } x_t \equiv x_t' = 1 \\ s_{22} & \text{if } x_t \equiv x_t' = 2 \\ s_{12} & \text{if } x_t \neq x_t' \end{cases},$$

where s_{11} and s_{22} can be interpreted as variance parameters for their respective subsets of the domain, while s_{12} can be interpreted as the covariance between the two audiograms. Thus the test kernel, the conjoining kernel, essentially acts as a covariance matrix between the two functions.

We combine these kernels into one composite kernel using the following equation:

$$K(x, x') = K_t(x_t, x_t')[K_I(x_I, x_I') + K_f(x_f, x_f')]$$

Likelihood Function

The likelihood function of a GP parametrizes the probability of observing the data $p(y|f, X)$. For our model, we use the cumulative Gaussian likelihood for binary classification, which is sometimes referred to as a probit likelihood. This is parameterized as $p(y_i = 1|f_i) = \Phi(f_i)$ and is one of the standard likelihoods for classification tasks (Rasmussen & Williams, 2006).

Inference Function

Computing the exact form of the posterior distribution

$$p(f|X,y) = \frac{1}{Z}p(f|X)\prod_{i=1}^{n} p(y_i|f_i)$$

is intractable because of the probit likelihood function (Rasmussen & Williams, 2006). Thus, the posterior must be approximated. For this model, this was done using expectation propagation (Minka, 2001; Rasmussen & Williams, 2006).

Model Estimation and Hyperparameter Selection

Let θ be defined as the set of all hyperparameters for the model. θ for the mean function and the frequency and intensity kernels are initialized to the ending values of the same hyperparameters of the prior audiogram. The hyperparameters for the discrete test covariance matrix are initialized to $exp(1)$ for s_{11} and $1 + exp(1)$ for s_{22}. Hyperparameter s_{12} is initialized Let as 1. Let be defined as $D = \{X, y\}$ the set of all observations and associated responses. Each iteration, new hyperparameters are selected to attempt to maximize $P(\theta|D)$. Due to the fact that the underlying distribution of $P(\theta|D)$ may be multimodal, we perform gradient descent twice. The first iteration of gradient descent is done by beginning with the hyperparameters returned from the previous iteration. The second iteration is done by beginning with a hyperparameter selected from a Gaussian distribution centered at the final hyperparameter values of the prior model. The θ with the higher marginal likelihood is saved, and used for computing the posterior.

Active Learning

Next, we use Bayesian Active Learning by Disagreement (BALD) to select the data point (x*,y*) that maximizes the difference in entropy between the posterior distribution of θ and the expected posterior distribution of θ given (x*, y*), according to the equation below:

$$\underset{x^*}{\operatorname{argmax}} H(\theta|D) - E_{y^*}[H(\theta|D, x^*, y^*)]$$

Since this equation can be intractable, it can be rewritten as

$$\underset{x^*}{\operatorname{argmax}} H(y^*|x^*, D) - E_{\theta}[H(y^*|D, x^*, \theta)]$$

This form of the equation is much easier to compute, being computable in O(1) time (Houlsby, Huszár, Ghahramani, & Lengyel, 2011). To calculate this, we can calculate $p = P(y^*|x^*, D)$ and pass this to the Bernoulli entropy equation $H(p) = plog_2(p) - (1-p)log_2(1-p)$. To calculate $E_{\theta}[H(y^*|D, x^*, \theta)]$, we use an approximation:

$$E_{f\sim(f|D)}[H(y^*|D, x^*, \theta)] = \frac{\sqrt{\frac{\pi ln2}{2}} e^{-\frac{\mu^2}{2(\sigma^2 + \frac{\pi ln2}{2})}}}{\sqrt{\sigma^2 + \frac{\pi ln2}{2}}}$$

In order to stabilize the acquisition and prevent sampling issues, the BALD values are normalized to between [0,1] and Gaussian noise is added. Finally, the point with the maximum of the modified BALD values is selected for observation.

Aim 2: Bayesian Active Differential Selection

In this section we introduce how BAMS can be used to detect whether psychometric functions have changed. In BAMS the ultimate goal is to determine which of two or more models has the highest probability of generating the observed data. In this case, we create two models for use in BADS, with the goal of answering the question: Is the underlying

psychometric function generating our new observations the same psychometric function, or a different psychometric function that has some nonunity correlation with the prior psychometric function? The first model is a model that treats the underlying distribution as being the same, and which we refer to as the "same model." The second model views the observed function as being a function correlated with, but not necessarily the same as, the prior function. We refer to this model as the "changed model." More details of these models are given below.

Feature Space

We begin by taking a prior audiogram, and augmenting it with all 1s for the test variable, as we did in BADE, such that $x_i = (f_i, I_i, t_i)$. Also as in BADE, all future observations will have $t_i = 2$.

Same Model

The same model is very similar to the model proposed for BADE. Like BADE, It uses a constant mean function. The frequency and intensity covariance kernels are also the same, i.e. an isotropic squared exponential kernel and an isotropic linear kernel, respectively. Instead of using a covariance matrix, the covariance of the test dimension is always set to 1, i.e., $K_t(x_t, x_t') = 1$. Thus the full kernel, the structure of which stays the same, is:

$$K(x, x') = K_t(x_t, x_t')[K_I(x_I, x_I') + K_f(x_f, x_f')]$$

Since $K_t(x_t, x_t') = 1$, this equation can be simplified to:

$$K(x, x') = K_I(x_I, x_I') + K_f(x_f, x_f')$$

Changed Model

The changed model, like the same model, is basically the same as as the model introduced in BADE, but with important differences. First, the mean function is discrete, using a separate constant mean function for $x_t = 1$ and $x_t = 2$. As for BADE, the covariance kernel for the test feature is:

$$K_t(x_t, x_t') = \begin{cases} s_{11} & \text{if } x_t \equiv x_t' = 1 \\ s_{22} & \text{if } x_t \equiv x_t' = 2 \\ s_{12} & \text{if } x_t \neq x_t' \end{cases}$$

Unlike for BADE, the parameters of K_t are fixed, such that $s_{11} = s_{22} = 1$. We fix the values of s_{11} and s_{22} because this allows the value of s_{12} to be interpreted as correlation between the functions, since

$$Corr(x, y) = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

where $cov(x, y)$ is s_{12}, and σ_x and σ_y are s_{11} and s_{22} respectively. This is important in the BAMS portion of the procedure.

Inference and Likelihood Functions

Like for BADE, the likelihood function used for BADS is the cumulative Gaussian likelihood for binary classification. Also as for BADE, the inference function used was expectation propagation.

Hyperparameter Learning

Early versions of BADS began with hyperparameter learning turned off with the intent of establishing the second model to be similarly constrained as the original model. This resulted in numerical stability issues, as the resulting differential entropy function returned smaller and smaller values, until underflow made BALD unviable. Learning new hyperparameters at each iteration for each model prevented this from occurring while still allowing for rapid discrimination between whether the underlying target function was the same or different from the prior model.

Hyperparameter learning was performed in the same way as for BADE, using a double gradient descent approach with the last iteration hyperparameters for one iteration of gradient descent and a Gaussian prior centered at the prior model's final hyperparameters for the other iteration. Hyperparameters were only updated for the mean function AND the frequency and intensity portions of the kernel, but not for the test kernel.

Bayesian Active Model Selection

All hyperparameters are initialized to the values used by the prior model except for $K_t(x_t, x_t')$, which is fixed to have $s_{11} = s_{22} = 1$ as described for the changed model. At each iteration, we calculate the mutual information between y* and the unknown model m for every candidate point in x*, using the following equation:

$$\underset{x^*}{\operatorname{argmax}} H(m|D) - E_{y^*}[H(m|D, x^*, y^*)]$$

Like the entropy equation in Aim 1, this is often intractable and is rewritten, similarly, as

$$\underset{x^*}{\operatorname{argmax}} H(y^*|x^*, D) - E_M[H(y^*|D, x^*, m)]$$

Next, it is necessary to derive formulas to compute each term in the above equation.

To calculate $H[P(y^*|x^*, D)]$, we first need to calculate $P(y^*|x^*, D)$. This can be expanded out over our models as

$$P(y^*|x^*, D) = \sum_{m \in M} P(y^*|x^*, D, m)P(m|D)$$

Similarly, we need to calculate $E_M[H(y^*|D, x^*, m)]$, which can be written as

$$E_M[H(y^*|D, x^*, m)] = \sum_{m \in M} H[P(y^*|x^*, D, m)]P(m|D)$$

Calculating P(m|D)

$P(m|D)$ is the probability of the model given the data, and shows up in both $H(y^*|x^*, D)$ and $E_M[H(y^*|D, x^*, m)]$. It can be expanded out, using Bayes' rule to give

$$P(m|D) = \frac{p(y|X, m)p(m)}{\sum_i p(y|X, m)p(m)},$$

where $p(m)$ is the prior for the model probability. Assuming a uniform prior over models, this reduces to, for our case,

$$P(m|D) = \frac{P(y|X, m)}{P(y|X, m =' same') + P(y|X, m =' different')}$$

Next, we must calculate $p(y|X, m)$ for each model. In the case of $p(y|X, m = `same')$ this is relatively straightforward, since the hyperparameters $\theta$ are fixed at each iteration, and we find $p(y|X, m = `same) = p(y|X, m = `same', \theta)$. Therefore, $p(y|X, m = `same', \theta)$ is easy to calculate, as expectation propagation returns the negative log marginal likelihood $-lnp(y|X, m =' same', \theta)$.

Next we have to calculate $P(y|X, m =' different')$. This is a bit trickier since we now have to worry about s_{12}. Recall that s_{12} can be interpreted as the correlation between the test domains, with correlation being measured between -1 and 1. This leads us to the integral

$$p(y|X, m =' different') = \int_{-1}^{1} p(y|X, m, s_{12})p(s_{12}|m)ds_{12}$$

Since computing the integral analytically is intractable, we create an approximation using quadrature, specifically the trapezoid rule. For computational purposes, we create a vector of s_{12} values between -1 and 1, $\vec{s}_{12}$. Next, we calculate

$$p(y|X, m =' different', s_{12})p(s_{12}|m =' different')$$

for each value of $s_{12,i}$ in $\vec{s}_{12}$. Under a uniform prior belief for the two possibilities,

$$p(s_{12}|m) = \frac{1}{2}.$$

This simplifies the integral, which becomes

$$p(y|X, m =' different') = \frac{1}{2}\int_{-1}^{1} p(y|X, m, s_{12})ds_{12}.$$

Like for the same model, we can calculate $p(y|X, M, s_{12})$ using the negative log marginal likelihood returned by a call to expectation propagation. We create a vector of $lnp(y|X, m, s_{12,i})$ values by repeatedly using the expectation propagation algorithm on each value $s_{12,i}$ in $\vec{s}_{12}$. For the sake of numerical stability while calculating the above integral, we subtract out the maximum value of the vector from each element to prevent underflow. We follow this by exponentiating and taking the integral using the trapezoidal approximation, and finish by correcting our integral by multiplying by the exponent we subtracted out.

Calculating P(y*|X*,D,m)

The second term needed for computing the entropy in both the marginal and individual entropies is the predictive distribution $P(y^*|x^*, D, m)$. For the same model, calculating $P(y^*|x^*, D, m = same)$ is relatively straightforward as it is a Bernoulli distribution with:

$$P(y^*|x^*, D, m = same) = \int \phi(f) N(f; \mu, \sigma^2) df = \phi\Big(\frac{\mu}{\sqrt{1+\sigma^2}}\Big).$$

Calculating $P(y^*|x^*, D, m = different)$ is a bit more complicated. Because we have one free hyperparameter in our different models, $K_t(x_t, x_t')$ matrices s_{12}, we take a weighted sum over the values

$$P(y^*|x^*, D, m = different) = \sum_{i \in s_{12}} \phi\Big(\frac{\mu_i}{\sqrt{1+\sigma_i^2}}\Big) p(s_{12,i}|D).$$

As when calculating P(m|D), we choose discrete values of s_{12} between -1 and 1.

The last thing we need to do is to calculate $p(s12, i|D)$ for each value of $s_{12,i}$ in $\vec{s}_{12}$. This can be written as

$$P(s_{12,i}|D) = \frac{P(y|x, s_{12,i}) P(s_{12,i})}{\int_{-1}^{1} p(y|x, s_{12}) p(s_{12}) ds_{12}}$$

Assuming a uniform prior on $P(s_{12,i})$, the prior terms cancel, and this becomes a simple calculation once we have the likelihood vector composed of $p(y|x, s_{12,i})$ for all values of $\vec{s}_{12}$. We now have the predictive distribution $P(y^*|x^*, D, m)$ for both models and can compute the marginal expected entropy and the expected individual entropy over the models.

Since all of the distributions involved are Bernoulli distributions, we again use the Bernoulli entropy function for calculating H in BADS.

$$h(p) = -plog_2 p - (1-p)log_2(1-p)$$

$$\underset{x^*}{\text{argmax}} H(y^*|x^*, D) - E_M[H(y^*|D, x^*, m)]$$

With the BALD criterion calculated, we apply a few heuristics to stabilize point selection. BALD values are normalized to the range [0,1]. Next, Gaussian noise is added, and an inverse distance heuristic is applied such that BADS avoids resampling points in extremely close

proximity. Finally, the x^* that maximizes the BALD criterion after the heuristic has been applied is selected as the next point to sample.

The Bayes factor, defined as the ratio of the posterior probabilities of the models $\frac{P(M1|D)}{P(M2|D)}$ is a useful metric for measuring model probabilities, as it can be computed on the fly. If one model exceeds a Bayes factor of 100 (interpreted as one model being 100× more likely) relative to the other model, this can be interpreted as BADS converging to a selected model.

Experimental Setup

To test our methods, we used the ground truth audiograms described in section 2.1. These ground truths were then used to compute the "prior" estimated audiograms used in both BADE and BADS. These prior audiograms were computed using the existing GP AMLAG framework (Song, Garnett, & Barbour, 2017). For these prior audiograms, hyperparameters for learning were initialized to the values learned in the conjoint framework (DiLorenzo, 2017). Next, 15 points were selected using the pseudo-random Halton sampling method (Halton, 1964) to provide a stable basis for learning the GP model. Another 85 points were actively sampled using a BALD criterion (Houlsby et al., 2011) normalized to the range $[0, 1]$ with a small amount of Gaussian noise. The resulting GPs are shown in Figure 6.

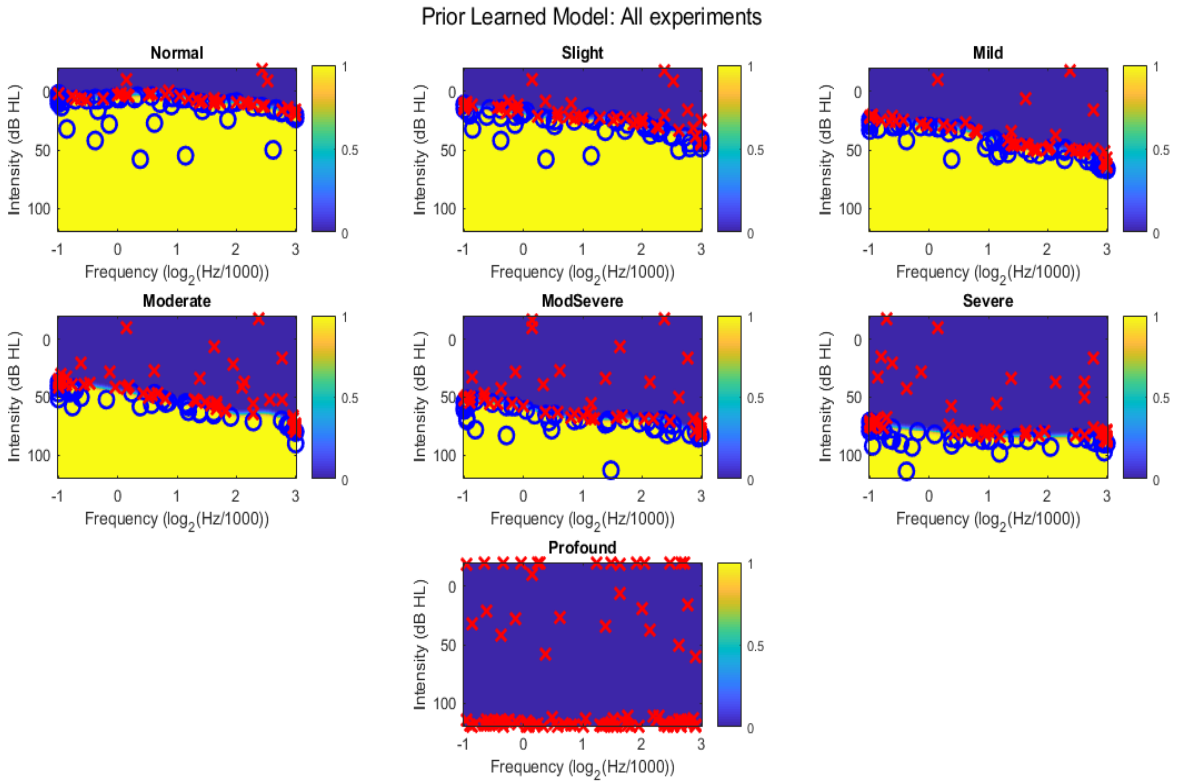


Figure 6. The Prior Learned Model for all Experiments

The prior learned model for each ground truth. X's represent unheard tones, while O's represent heard tones. These are approximations of the actual ground truths in Figure 5.

Both BADE and BADS were run with each combination of a prior audiogram from the set of prior audiograms, above, and a ground truth from the set of ground truths for a total of 49 combinations. For each combination, BADE was run for 60 iterations, while BADS was run for 20 iterations. The algorithm was only run once for each combination.

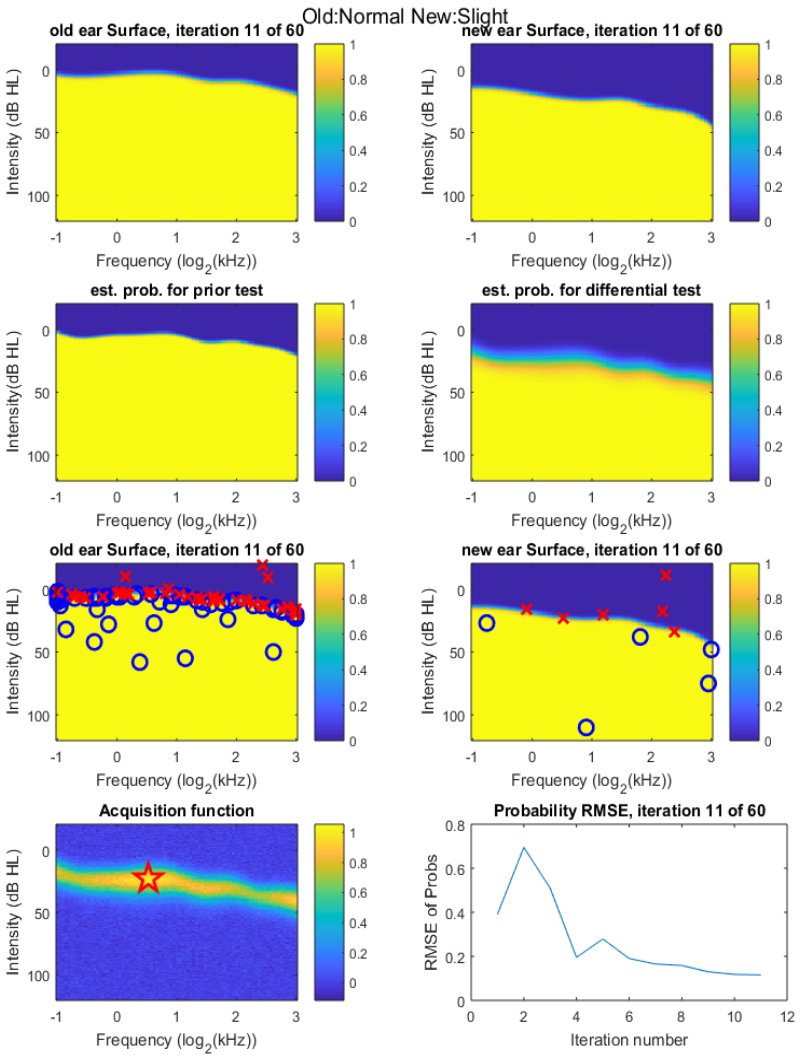


Figure 7. Differential Estimation Example: Different

Example experiment figure from BADE at iteration 11. Top 3 rows show ground truth, estimated model, and ground truth with tones superimposed, respectively. Left column shows the prior model, while the right column shows the target. Bottom left shows the acquisition function, which indicates the maximum value to sample as the next point. Bottom right shows the posterior RMSE of the predicted probabilities. In as few as 11 tones, the threshold has been almost exactly approximated, though the spread of the distribution will take a bit longer to identify. The posterior already has the correct shape, matching the ground truth (upper right plot).

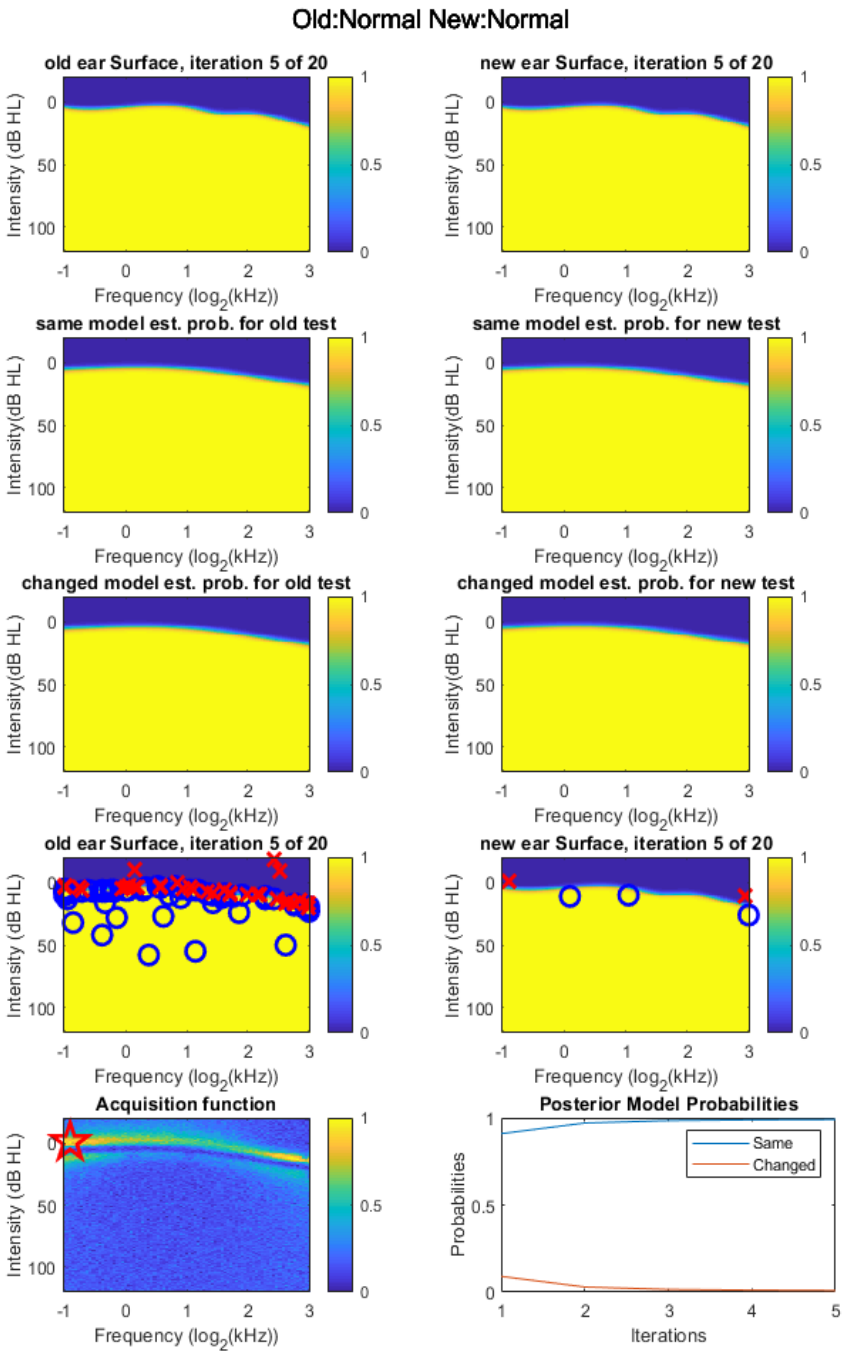


Figure 8. Bayesian Active Differential Selection Example: Same

In progress experiment figure for BADS at iteration 5 of 20. Both the prior and target hearing loss classifications are the same: Normal hearing. Acquisition has sampled both above and below threshold, stabilizing the posterior probability p(M|D).

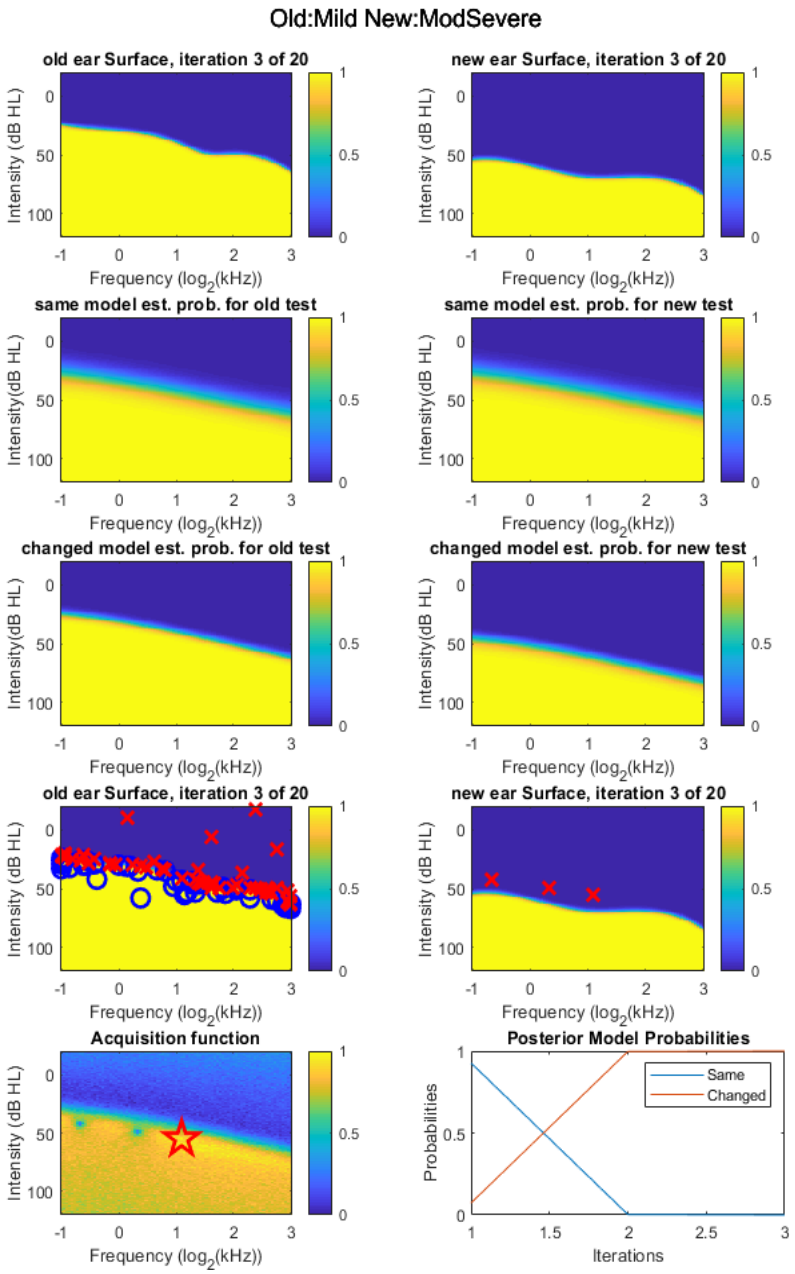


Figure 9. Bayesian Active Differential Selection Example: Different

In progress experiment figure for BADS at iteration 3 of 20. Prior model classification is “Mild hearing loss”, while target is “Moderate-Severe hearing loss”. Acquisition has sampled below the prior model threshold, where the prior model would have classified the tones as heard. These three tones were unheard, however. This results in a quick convergence of the posterior to “changed”.

Results: Bayesian Active Differential Estimation

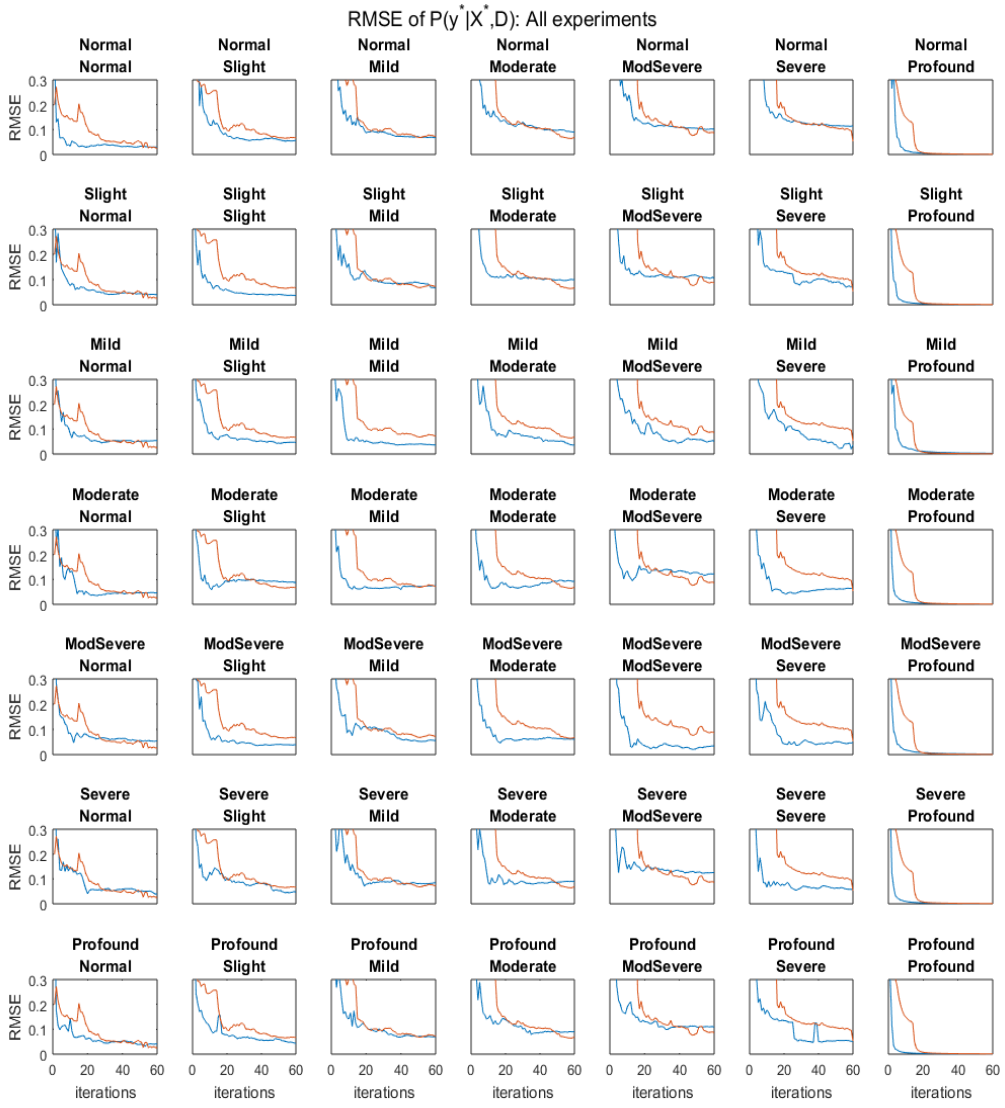


Figure 10. BASE RMSE of probabilities

Plot of root mean squared error (RMSE) in probability of the estimated model from the ground truth for each combination of hearing loss prior and target as a function of iterations. Note the model predicts a probability for each value in the frequency × intensity domain, and this is the error across the entire domain. The blue line in each plot is the result of BADE, while the orange line is the error of the prior audiogram for the first 60 iterations. In each title, the top loss type denotes the prior audiogram while the bottom title denotes the ground truth audiogram.

The first method used to analyze the results of BADE was the root mean squared error of the entire surface probability estimation. This metric is useful, as it evaluates model performance across the entire domain. The result in Figure 10 shows that BADE converges to an estimated model more quickly than past approaches. When the prior audiogram and the ground truth are of the same hearing loss type, the model converges rapidly, as expected. Even when the hearing loss types are vastly different, however, the shared variance between hearing loss types still allows for a moderate speed up. When the hearing loss types were the same, RMSE generally converged to low error values (<5%), whereas hearing loss types that were different managed to achieve an error that was slightly higher, usually 5%-10%.

The second metric used to evaluate BADE performance was threshold error. This was chosen because the threshold is the relevant metric in a clinical setting for diagnosing hearing loss using current methods. For each frequency, the lowest intensity value with a probability larger than 0.5 was marked as the threshold intensity. This was done for both the ground truths as well as the estimated model. Next, we evaluated the performance of the model by comparing the root mean squared error between the ground truth and the target model. The results are shown in Figure 11. Also of interest is the number of iterations to 5 dB convergence, defined as a root mean squared error between ground truth and target model of less than 5 dB, as this is the approximate resolution of current methods such as Hughson Westlake. These results are included in Table 2 and Table 3.

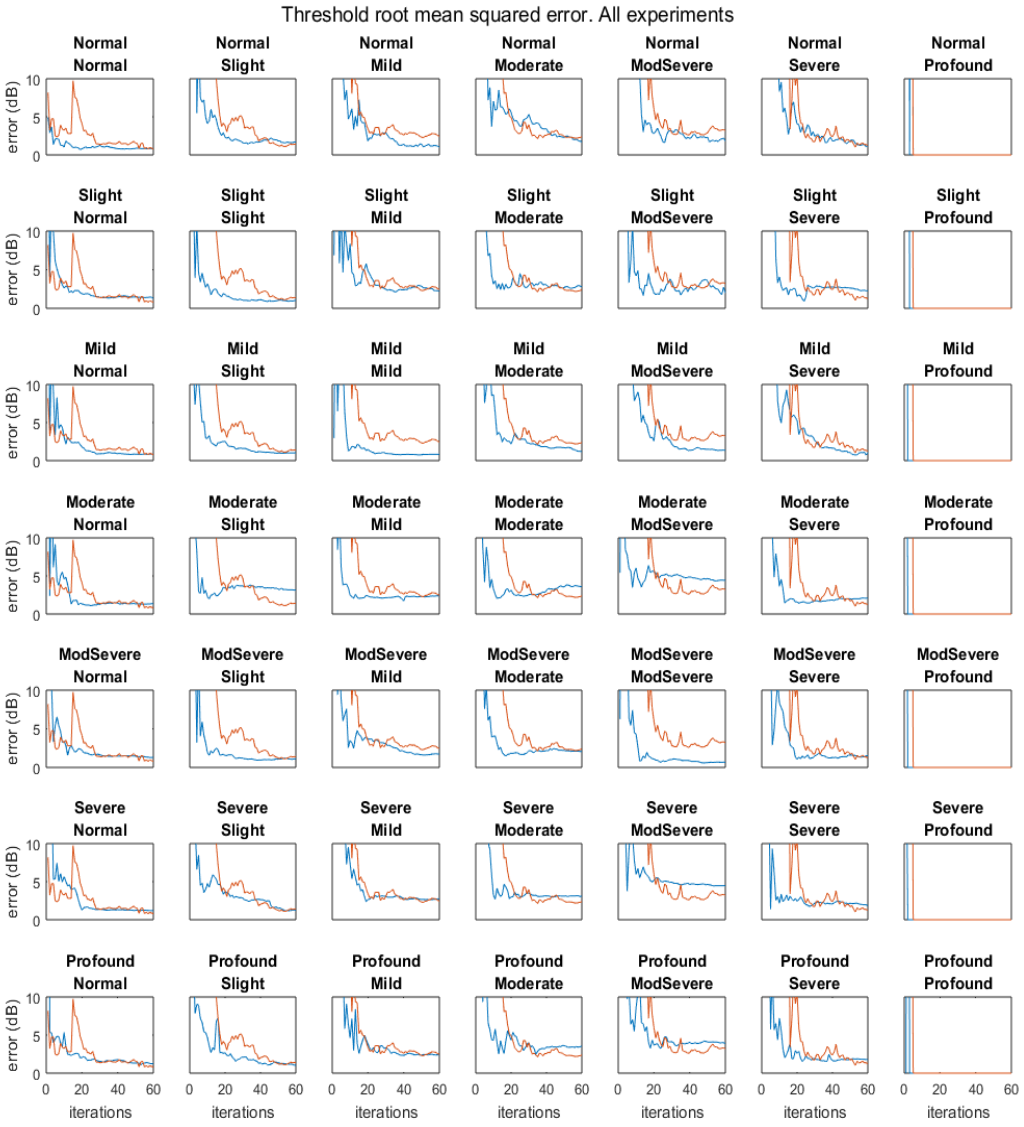


Figure 11. BADE Threshold RMSE

Plot of root mean squared error of threshold in dB HL of the estimated model from the ground truth for each combination of hearing loss prior and target as a function of iterations. The blue line in each plot is the result of BADE, while the orange line is the error of the prior audiogram for the first 60 iterations estimated with AMLAG. In each title, the top hearing loss type denotes the prior audiogram while the bottom title denotes the ground truth target audiogram.

Table 2

Iterations to 5 dB threshold error for prior models estimated with AMLAG

Ear Type	Normal	Slight	Mild	Moderate	ModSevere	Severe	Profound
Iterations	2	18	17	20	21	16	5

Table 3

Iterations to 5 dB threshold error for BADE.

Ear Type	Normal	Slight	Mild	Moderate	ModSevere	Severe	Profound
Normal	1	9	9	9	14	14	3
Slight	2	3	6	9	6	9	3
Mild	2	9	1	12	15	11	2
Moderate	2	5	6	5	8	7	2
ModSevere	4	4	9	9	9	6	2
Severe	9	6	14	10	5	5	2
Profound	4	10	10	9	17	9	1

Note: Rows: prior hearing loss type, columns: target model type

Figure 11 shows that BADE converges on the correct threshold much faster than current methods. For almost all cases where the prior and target models are the same hearing loss classification, BADE converges to within just a few dB of the actual threshold in roughly half the number of observations. The threshold for the profound hearing loss ground truth, which is essentially nonexistent, makes the last column of plots uninterpretable for threshold error.

Table 2 shows that standard methods converge within approximately 16-21 tones. Recall the first 15 samples of the standard GP method use Halton sampling, which is necessary to establish stability of the GP model. The model does not converge to the true threshold until after Halton sampling is complete. Note that while the table says that ‘normal’ converges in 2 iterations, this is misleading as the GP was still experiencing instability at this stage, as can be seen in the orange lines of the normal column in Figure 11.

Table 3 shows that BADE converges to within 5 dB of the true threshold within roughly 10 iterations in almost all cases. Since BALD searches for points maximizing differential entropy, which are usually found near the threshold, the use of the prior belief allows the model to search for the threshold instantaneously without being encumbered by Halton sampling for the sake of stability.

Results: Bayesian Active Differential Selection

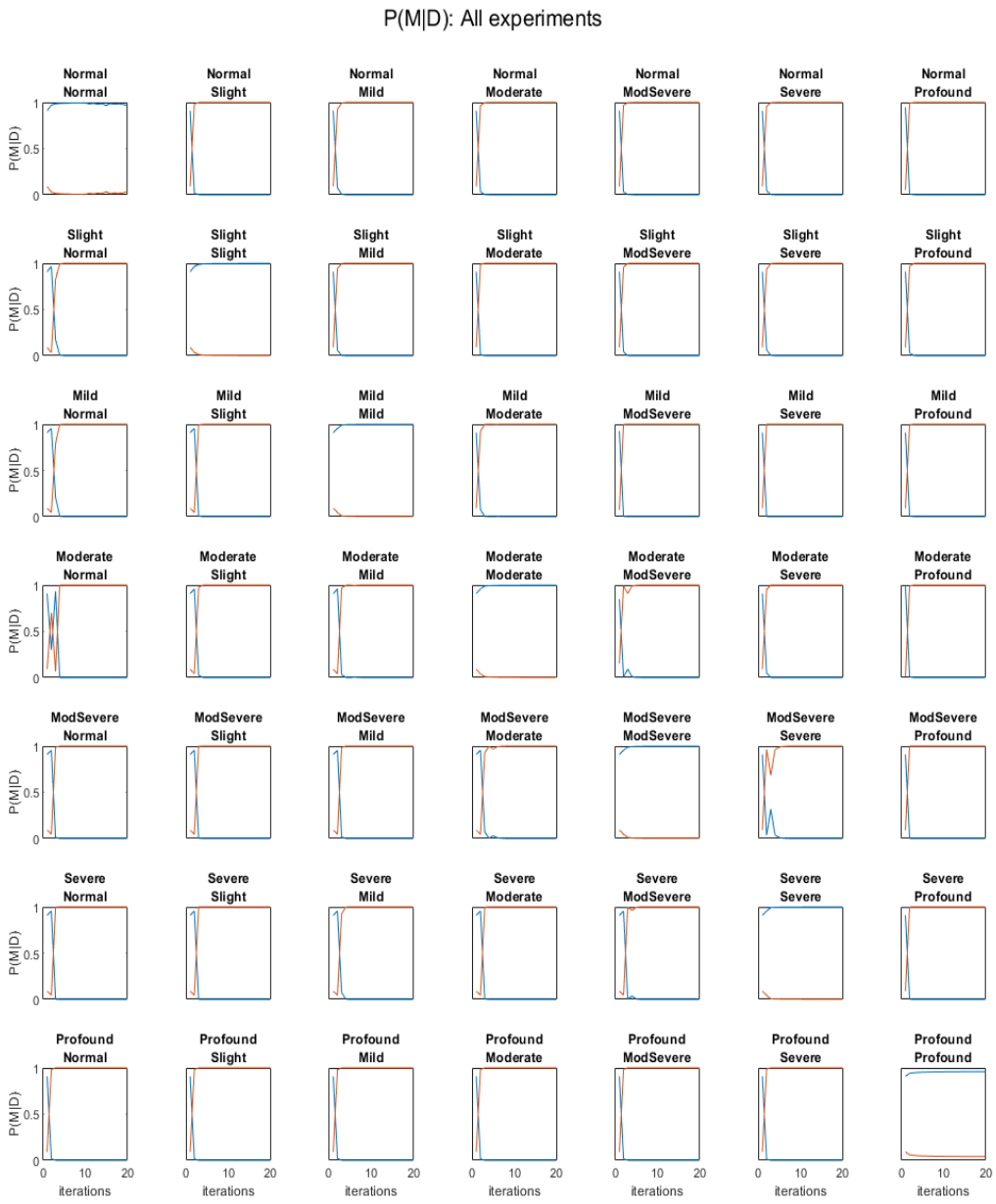


Figure 12. BADS Posterior Model Probabilities

Posterior model probabilities P(m|D). Blue lines represent P(m='same'|D), while orange lines represent P(m='different'|D). All models converge within 5 tones to near 100% probability of the correct model.

The first metric for measuring the success of BADS is the posterior model probabilities. BADS rapidly converged to the correct model classification within five tones in all cases, but usually approached the correct classification in as few as one tone, as can be seen in Figure 12. Some cases, such as the "Prior: Mild, Ground truth: Normal" took two tones. The "Prior: Moderate, Ground truth: Normal" and "Prior: ModSevere, Ground truth: Severe" combinations took slightly longer to converge, but only by a couple iterations. Due to the fact that convergence was so rapid, the logarithm of the Bayes Factor was used to further examine convergence to classification, as can be seen in Figure 13 and Figure 14. In Figure 14 the dashed red lines represent the criterion for a Bayes Factor of 100 in either model's favor, and the number of tones needed to cross this threshold can be seen in Table 4. When the prior model and the target model were different, BADS rapidly crossed the Bayes factor significance level and continued to increase to extremely high values. When the models were the same, BADS still rapidly converged past the significance criterion in all scenarios but the profound vs profound case. This can be explained by the fact that determining whether two functions are different can be proved by a single "counter example" where the observations between the two models do not match. On the other hand, showing that the models are the same requires a larger amount of points to verify that the threshold is similar across the frequency domain. All models were successfully classified in 6 or fewer tones, except for the profound vs profound case, which is a weird edge case due to total deafness.

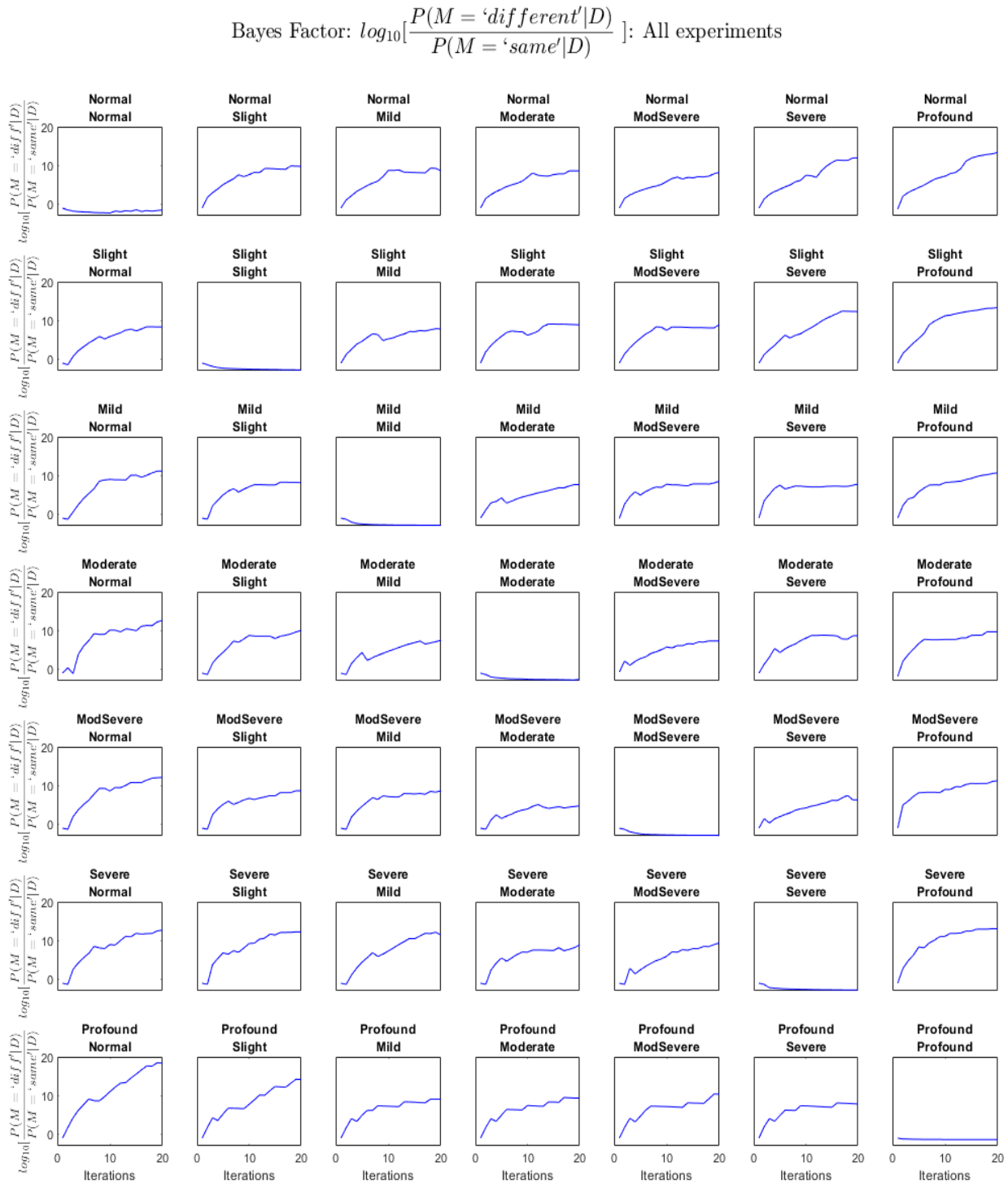


Figure 13. BADS Bayes Factor Plots

Plotted are log_{10} of the Bayes factor, defined as the ratio of probabilities P(m|D), as a function of iterations.

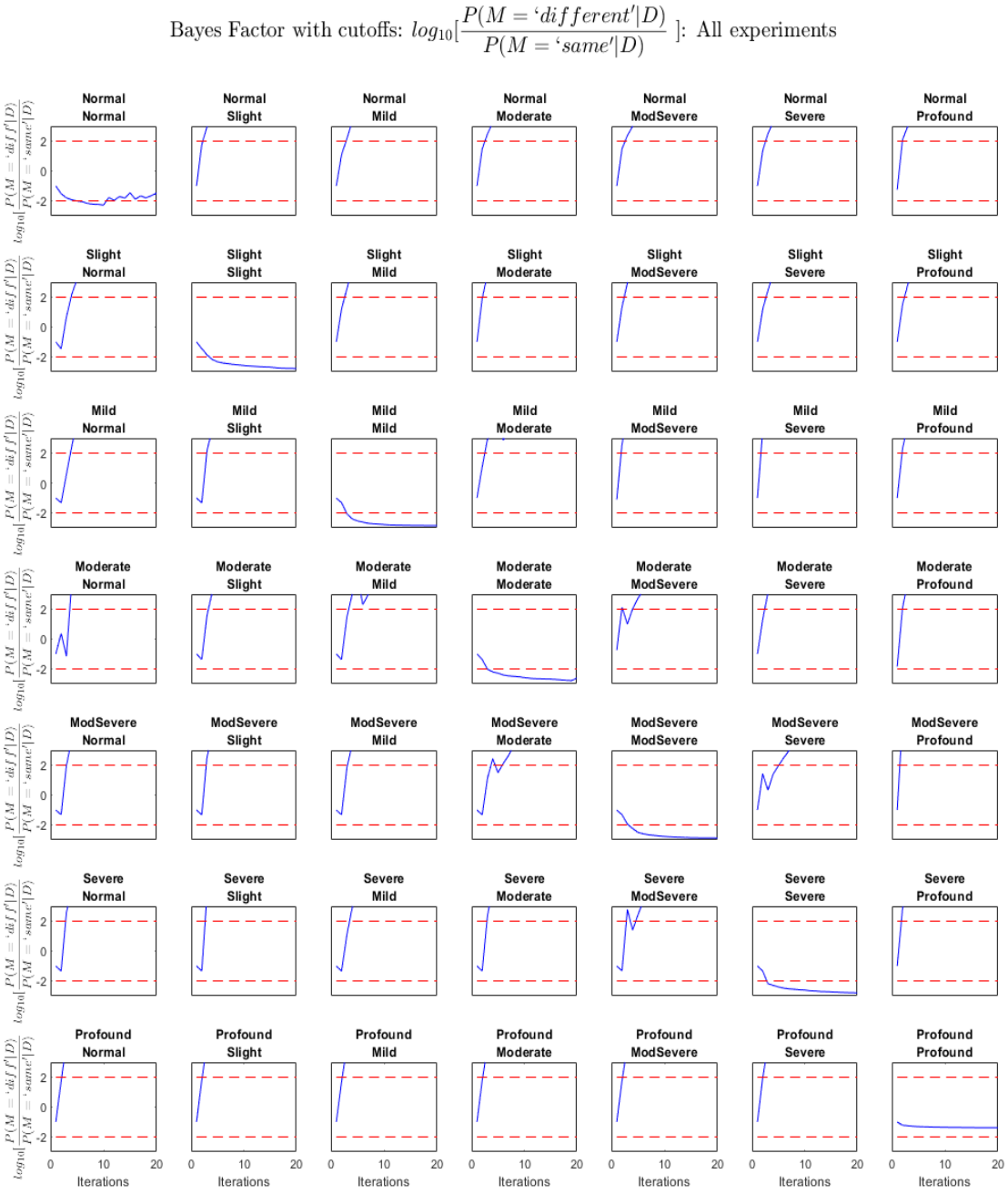


Figure 14. BADS Bayes Factor with cutoff

Plotted are log_{10} of the Bayes factor. Axis scaled to -3 to 3, with red dashed lines representing Bayes factors of 0.01 and 100 (reflecting when one model is 100 times more likely than the opposing model) as a cutoff for statistical significance.

Table 4

Iterations to Bayes Factor of 100 or 0.01.

Ear Type	Normal	Slight	Mild	Moderate	ModSevere	Severe	Profound
Normal	5	3	3	3	3	3	2
Slight	4	4	3	3	3	3	3
Mild	4	3	3	3	2	2	2
Moderate	4	4	4	3	2	3	2
ModSevere	4	3	4	4	4	6	2
Severe	3	3	4	3	3	3	2
Profound	3	3	3	3	3	3	NaN

Note: Rows: prior hearing loss type, columns: target model type

Another interesting result to examine is the probability distribution of $P(s_{12}|D)$. Hyperparameter s_{12} was represented computationally as a vector of values between -1 and 1. Part of the computation of the entropy function required calculation of this statistic. Initial development of this model used a maximum likelihood approach to estimating the entropy and log evidence of the predictive distribution $P(y^*|x^*, D, m)$. Due to the fact that the 'different' model had the ability to become functionally equivalent to the same model under the maximum likelihood approach, this led to early versions of BADS repeatedly sampling the same point, as the models did not disagree on the next point to select. Integrating over values of s_{12} helped solve this problem. The probability distribution of $P(s_{12}|D)$ is shown in Figure 15 for each combination of hearing loss types after 20 iterations. When the hearing loss types were the same, $P(s_{12} = 1|D)$ had extremely high probability, while other probabilities were much lower. For nearby off-diagonal combinations, $P(s_{12}|D)$ has a large probability mass near $s_{12} = 1$ is still

large, but not nearly as large as on the diagonal. For hearing loss types that were quite different (far off-diagonal), $P(s_{12}|D)$ had a much wider distribution, usually centered between 0 and 0.5, showing that the correlation between models was much lower.

Across almost all combinations except those involving the profound hearing loss type, the vast majority of probability density is massed close to $s_{12} = 1$, or at the very least is positive. This positive covariance between models help explain the speedup that BADE and conjoint methods generally provide.

Figure 15. Posterior Covariance Probabilities

Plot of probability distributions of $P(S_{12}|D)$.

Discussion and Conclusion

This work shows that utilizing information from previous models as a prior belief allows for a substantial speedup in audiogram estimation, and can even be leveraged with BAMS to rapidly determine whether functions are different. In the clinic, utilizing prior information from past audiograms, either from a specific patient, or the human population more generally, has the potential to drastically cut the amount of time it takes to run diagnostic tests for hearing loss as well as other applications. While this thesis specifically focused on hearing loss, the methods in this paper, specifically the exploitation of a discrete covariance matrix to link prior data to a fresh GP model, is widely applicable to a vast array of tasks that can be modelled using regression or classification with GPs.

Future work could involve integrating the methods introduced in this paper with the conjoint formulation of the bilateral audiogram to run BADE and BADS on both ears simultaneously to speed up acquisition of a conjoint model of patient hearing. Other work could include representing the prior belief using a sparse set of data that accurately model the prior audiogram to reduce runtime. Further in the future, it would be interesting to test BADE and BADS on other applications outside of audiometry.

References

Barbour, D. L., DiLorenzo, J. C., Sukesan, K. A., Song, X. D., Chen, J. Y., Degen, E. A., … Garnett, R. (2018). Conjoint psychometric field estimation for bilateral audiometry. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-1062-3

Carhart Raymond, & Jerger James F. (1959). Preferred Method For Clinical Determination Of Pure-Tone Thresholds. *Journal of Speech and Hearing Disorders*, *24*(4), 330–345. https://doi.org/10.1044/jshd.2404.330

Clark, J. G. (1981). Uses and abuses of hearing loss classification. *ASHA*, *23*(7), 493–500.

DiLorenzo, J. (2017). *Conjoint Audiogram Estimation via Gaussian Process Classification* (Washington University in St. Louis). https://doi.org/10.7936/K7ZK5F4V

Fechner, G. T. (1860). *Elemente der psychophysik*. Retrieved from http://archive.org/details/elementederpsych001fech

Gardner, J., Malkomes, G., Garnett, R., Weinberger, K. Q., Barbour, D., & Cunningham, J. P. (2015). Bayesian Active Model Selection with an Application to Automated Audiometry. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28* (pp. 2386–2394). Retrieved from http://papers.nips.cc/paper/5871-bayesian-active-model-selection-with-an-application-to-automated-audiometry.pdf

Halton, J. H. (1964). Algorithm 247: Radical-inverse Quasi-random Point Sequence. *Commun. ACM*, *7*(12), 701–702. https://doi.org/10.1145/355588.365104

Houlsby, N., Huszár, F., Ghahramani, Z., & Lengyel, M. (2011). Bayesian Active Learning for Classification and Preference Learning. *ArXiv:1112.5745 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1112.5745

Hughson, W., & Westlake, H. D. (1944). *Manual for program outline for rehabilitation of aural casualties both military and civilian: sponsored by the American Academy of Ophthalmology and Otolaryngology*. Omaha, Neb: Douglas Print. Co.

Levitt, H. (1971). Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2B), 467–477. https://doi.org/10.1121/1.1912375

Masterson, E. A., Tak, S., Themann, C. L., Wall, D. K., Groenewold, M. R., Deddens, J. A., & Calvert, G. M. (2013). Prevalence of hearing loss in the United States by industry. *American Journal of Industrial Medicine*, *56*(6), 670–681. https://doi.org/10.1002/ajim.22082

Minka, T. P. (2001). *A Family of Algorithms for Approximate Bayesian Inference* (PhD Thesis). Massachusetts Institute of Technology, Cambridge, MA, USA.

Rasmussen - 2004 - Gaussian Processes in Machine Learning.pdf. (n.d.). Retrieved from https://www.cs.ubc.ca/~hutter/EARG.shtml/earg/papers05/rasmussen_gps_in_ml.pdf

Rasmussen, C. E. (2004). Gaussian Processes in Machine Learning. In O. Bousquet, U. von Luxburg, & G. Rätsch (Eds.), *Advanced Lectures on Machine Learning* (Vol. 3176, pp. 63–71). https://doi.org/10.1007/978-3-540-28650-9_4

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge, Mass: MIT Press.

Song, X. D., Garnett, R., & Barbour, D. L. (2017). Psychometric function estimation by probabilistic classification. *The Journal of the Acoustical Society of America*, *141*(4), 2513. https://doi.org/10.1121/1.4979594

Song, X. D., Wallace, B. M., Gardner, J. R., Ledbetter, N. M., Weinberger, K. Q., & Barbour, D. L. (2015). Fast, Continuous Audiogram Estimation Using Machine Learning. *Ear and Hearing*, *36*(6), e326-335. https://doi.org/10.1097/AUD.0000000000000186

Taylor, M. M., & Creelman, C. D. (1967). PEST: Efficient Estimates on Probability Functions. *The Journal of the Acoustical Society of America*, *41*(4A), 782–787. https://doi.org/10.1121/1.1910407