Spring 5-2017

Conjoint Audiogram Estimation via Gaussian Process Classification

James DiLorenzo
Washington University in St Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

School of Engineering and Applied Science

Department of Computer Science and Engineering

Thesis Examination Committee:
Roman Garnett, Chair
Dennis Barbour
Sanmay Das

Conjoint Audiogram Estimation via Gaussian Process Classification

by
James C. DiLorenzo

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

May 2017

Saint Louis, Missouri

Contents

List of Figures

Acknowledgements

First, I would like to thank my lab director Dr. Barbour. I came to his lab with no experience whatsoever in biomedical engineering and only limited exposure to machine learning. His guidance helped me identify interesting and challenging projects while providing the long-term vision to keep me motivated.

I was fortunate to get the opportunity to work with Dr. Garnett. The depth and breadth of his knowledge is incredible. His Bayesian machine learning class fundamentally changed the way I interact with data. He was encouraging and eager to help throughout the development of this entire thesis.

To the members of the Barbour Lab, thank you. Jeff and David, I looked up to you both throughout the entire process. You took me on as a mentee and taught me more about what it means to be a scientist than I can put into words. I look forward to seeing what you do next. Kiron and Braham, I am grateful to have you as my peers. To Kiron in particular, thank you for bringing me into the Barbour lab. This thesis would not even be an idea without you.

Finally to my wife Hillary, your strength and kindness inspire me every day. I am so grateful to walk through life with you.

James C. DiLorenzo

Washington University in St. Louis

May 2017

For Hillary

ABSTRACT OF THE THESIS

Conjoint Audiogram Estimation via Gaussian Process Classification

by

James C. DiLorenzo

Master of Science in Computer Science

Washington University in St. Louis, 2017

Research Advisor: Professor Roman Garnett

In traditional audiometry, a clinician seeks to estimate her patient's auditory response through the sequential delivery of various individual tests. These tests are treated as independent and correlation is assessed after each individual test has been completed, resulting in a diagnosis. Treating tests as independent impedes both accuracy and efficiency by ignoring correlations in conditions known to influence physiological response, for instance age, genetics, and exposure to noise. This thesis advances the existing framework for audiometry via Gaussian Processes by allowing for the estimation of audiogram thresholds for both ears simultaneously. The resulting model estimates both correlated and uncorrelated right- and left-ear audiograms with higher efficiency than was previously achievable. This work lays a foundation for building further estimation between discrete psychometric spaces.

1. Introduction

1.1. Psychometric Functions

Psychometric functions model an individual's task performance in response to some sensory stimulus. Much attention has been given to the estimation of unidimensional psychometric functions, or psychometric curves (PCs). One of the first methods for estimating PCs was the method of constant stimuli, which continues to see widespread use today and was first described in Gustav Fechner's famous *Elemente der Psychophysik* in 1860 (Fechner 1860). The method of constant stimuli randomly presents a fixed number of equally spaced stimuli with some repetition. While this method accurately predicts the target PC, it is time consuming in practice. This inefficiency led to the development of adaptive procedures for psychometric estimation. Adaptive procedures use subject response to influence the intensity of subsequent stimuli delivery with the goal of achieving similar accuracy in fewer observations including transformed up-down methods (Levitt 1971) and parameter estimation by sequential testing (PEST) (Taylor and Creelman 1967).

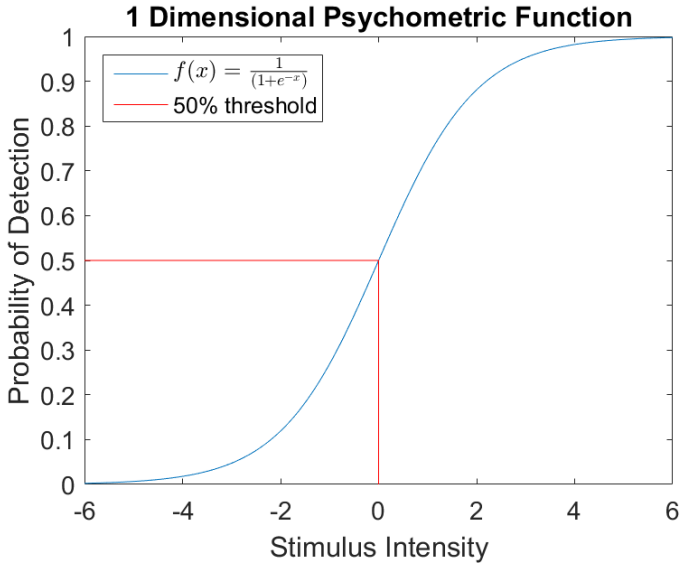

Figure 1: Depiction of 1-Dimensional Psychometric Function as a function of stimulus intensity

Inference of the PC falls broadly into two categories: parametric and nonparametric. Parametric models assume that the subject's true psychometric function follows some explicit formula which can be identified by its parameters. A typical set of parameters for PC estimation may include the threshold $\boldsymbol{\alpha}$, the intensity at which more than some fixed fraction of stimuli are observed, and slope $\boldsymbol{\beta}$, the rate at which subject response changes as a function of stimulus intensity (Hall 1968). By contrast, nonparametric models make no assumptions about the structure of the PC, but rather estimate values of the PC from observing data. Examples of nonparametric methods for PC estimation include spline estimation (Schoenberg 1946) in which the interpolant is a piecewise polynomial that passes through the observed data, and Gaussian Processes, a machine learning model in which the posterior distribution of the predicted values is jointly Gaussian, and this Gaussian has its mean and covariance matrix defined by the observed data (Williams and Rasmussen 1996). In both cases, the estimation of the model is derived from the observed data and not a set of fixed parameters.

Both adaptive techniques and inference methods leverage observed data to achieve their respective goals. In practice, it is often possible to improve performance by leveraging an expert's domain knowledge. For instance, an experienced audiologist may be able to construct an audiogram with fewer stimuli deliveries by making use of their knowledge of published reports or subject histories. Adaptive techniques and inference methods can both be improved by leveraging domain-specific knowledge. This domain-specific knowledge is known as a prior belief in statistical literature. Statistical inference on PCs can utilize both subject responses and prior beliefs via Bayes rule in a class of inference collectively referred to as Bayesian inference. Bayesian methods can be applied to both querying strategies (i.e. towards the improvement of adaptive techniques) or to PC inference itself. The earliest use of Bayesian inference and adaptive techniques is the QUEST method (Watson

and Pelli 1983). Bayesian methods are particularly promising for estimating PCs and continue to receive attention, particularly in the machine learning literature.

1.2. Audiometry

Audiometry presents interesting challenges for many psychometric function estimation techniques because the input space is inherently two-dimensional. A subject's response to stimuli depends not only on the intensity of sound delivered, but also on its frequency. This increase in dimensionality makes approaches such as the method of constant stimuli impractical. In general, the number of samples required to maintain a certain sample density increases exponentially with the number of dimensions of the input space. This is one facet of what is commonly referred to as the "curse of dimensionality" in machine learning literature.

The most commonly used method for clinical audiogram estimation is pure -tone audiometry (PTA) using a modified Hughson-Westlake (HW) procedure (Hughson and Westlake 1944, Carhart and Jerger 1959). Originally developed in response to the drastic increase in noise induced hearing loss among veterans after World War I and World War II, the HW procedure proceeds octave by octave (or semi-octave), delivering tones in decreasing 5dB increments until locating the threshold as measured by some number of "reversals."

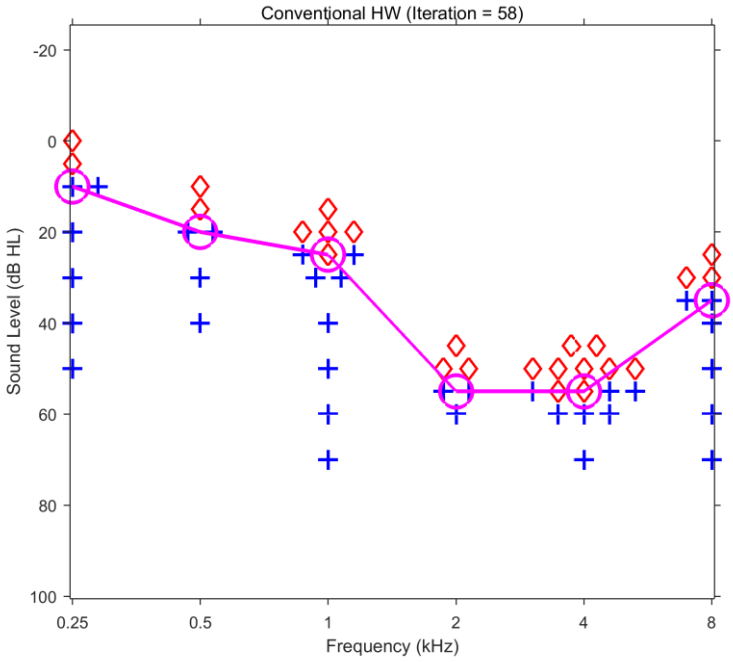

Figure 2: Hughson-Westlake Procedure Example adapted from (Barbour, Song 2015)

The HW procedure was an important first standard in audiometry. However, the HW procedure comes with some disadvantages. First, the HW procedure only queries a fixed number of frequencies. Practitioners can linearly interpolate between frequencies to obtain inter-octave thresholds. However, this approach is prone to missing narrow-banded notches common in noise induced hearing loss. A second major drawback of the HW procedure is that many of the tones delivered are uninformative. In particular, when moving from one frequency to the next, stimuli are delivered well above threshold. Third, predictable stimulus presentation sequences allow for noncooperative subjects to subvert the test. Finally, left- ear and right- ear audiograms are treated as independent. This assumption ignores important nonphysioloigical factors that contribute to both a subject's left- and right-ear audiograms. These factors include genetics, age, and environmental exposure and are major drivers of both noise-induced and age-related hearing loss.

A number of approaches have been developed in response to the shortcomings of the HW procedure. A 2013 review of techniques for pure-tone audiometry found that automated audiograms produced similar results to manual audiograms, with an average absolute difference of 4.2 dB HL (Mahomed, Eikelboom et al.). PEST (Taylor and Creelman), maximum likelihood estimation (Watson and Pelli), and numerous Bayesian methods attempt to address sampling inefficiencies (King-Smith, Grigsby et al. 1994, Guan 2011). Békésy audiometry and Audioscan® deliver continuous audiogram threshold estimates at the cost of slower testing (Meyer-Bisch 1996, Ishak, Zhao et al. 2011). Of note is the use of Gaussian Processes for audiometry. Gaussian Processes deliver accurate, continuous threshold estimates with a sampling schema that is difficult to subvert (Song, Garnett et al. 2017). However, none of the described methods are able to leverage test information between ears. In this paper, we extend the existing Gaussian Process model for continuous audiogram estimation to allow for querying and estimation in both the ipsilateral and contralateral ear. The resulting audiogram estimation technique is more efficient than the single-ear GP model without sacrificing accuracy. Additionally, the new model takes an important first step towards the goal of sharing information between disjoint tests in any testing battery.

2.Machine Learning Background

2.1. Introduction to Machine Learning

The goal in supervised machine learning is to train a model to estimate some underlying function $\boldsymbol{y(x)}$ from a set of labeled data $\boldsymbol{D = \{(x_i, y_i)\}_{i=1}^{n}}$, where $\boldsymbol{x_i}$ is the feature vector for observation $\boldsymbol{i}$ and $\boldsymbol{y_i}$ is the value of observation $\boldsymbol{i}$. $\boldsymbol{D}$ may be noisy (Mohri, Rostamizadeh et al. 2012). Namely, if there is some true underlying function $\boldsymbol{f(x)}$, then observations $\boldsymbol{y_i = f(x_i) + \epsilon}$. If one wants to improve the estimation of their model, a common approach is to simply train the model on additional data. This approach works particularly well for tasks such as image and speech recognition, where access to additional data is relatively inexpensive. Querying data is much more difficult in perceptual studies. Collecting additional data involves querying a subject and recording their response. Subject fatigue can lead to non-stationarity in subject responses, making it imperative that any predictive model is efficient in its sampling.

2.2. Motivation for the Use of Gaussian Processes

The need for sampling efficiency led to the choice of the Gaussian Process (GP) model. Also known as kriging, GPs were designed to estimate an unknown underlying function where access to data is expensive (Rasmussen and Williams 2006). Unlike other machine learning models that only give a point estimate of the underlying function value for a given input, GPs provide a posterior distribution of the model's belief of the underlying function value for a given test point. This distribution can be thought of as the model's uncertainty about its prediction and gives rise to techniques collectively known as active sampling. In this work, the GP model uses active sampling

to select a sequence of points $\boldsymbol{X} = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\}$ to query to maximize the rate at which it gains information about the subject's audiogram.

2.3. Gaussian Process Models

2.3.1. Introduction to Gaussian Processes

Gaussian Processes are a model for probabilistic inference about some function of interest $\boldsymbol{f}$. That is, instead of simply producing pointwise estimates $\hat{\boldsymbol{f}}(\boldsymbol{x})$, a GP returns a probability distribution $\boldsymbol{p}\left(\boldsymbol{f}(\boldsymbol{x})\right)$. A practitioner may encode domain-specific knowledge of $\boldsymbol{f}$ through a prior distribution. The GP is typically then conditioned on observed data $\boldsymbol{D}$ to form a posterior distribution $\boldsymbol{p}(\boldsymbol{f}|\boldsymbol{D})$. Formally, a GP is a collection of random variables such that the joint distribution of any finite subset of these random variables is a multivariate Gaussian distribution. (Rasmussen and Williams 2006) It is easier however to think of GPs as distributions over functions. Just as a variable drawn from a Gaussian distribution is specified by the distribution's mean and covariance, i.e. $\boldsymbol{p}(\boldsymbol{x}) \sim \boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$, a function drawn from the prior distribution of a GP is specified by its mean and Kernel functions, i.e. $\boldsymbol{p}(\boldsymbol{f}) \sim \boldsymbol{GP}(\boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}'))$. The mean function encodes the central tendency of functions drawn from the GP. The Kernel function encodes information about the shape these functions may take. Kernel functions can vary widely in construction and have a large impact on the posterior distribution of the GP. In general, Kernel functions are designed to express the belief that "similar inputs should produce similar outputs" (Duvenaud 2014). The GP model can be used in both classification and regression settings and allows us to condition our prior beliefs after observing data to produce a new posterior belief about function values via Bayes' rule:

$$\textbf{posterior} = \frac{\textbf{prior} \times \textbf{likelihood}}{\textbf{marginal likelihood}} \tag{2.1}$$

2.3.2. Gaussian Process Regression

The GP model for audiogram estimation gives probabilistic estimates for the likelihood of tone detection. However, to properly build up a framework for GP classification it is important to first examine GP regression.

In a typical regression problem, inputs $\boldsymbol{X}$ and outputs Y take on real values and are related through some function $\boldsymbol{f}$ of which we have access to only noisy observations. For convenience, this example assumes that noise is drawn independently and identically from a Gaussian distribution with mean 0 and standard deviation $\boldsymbol{s}$:

$$\boldsymbol{x_i} \in \mathbb{R}^{\boldsymbol{d}}, \qquad \boldsymbol{y_i} \in \mathbb{R} \tag{2.2}$$

$$\boldsymbol{y}(\boldsymbol{x_i}) = \boldsymbol{f}(\boldsymbol{x_i}) + \boldsymbol{\epsilon}, \boldsymbol{e} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{s^2}) \tag{2.3}$$

Before observing any data, the GP by definition implies a joint distribution on the function values of any set of input points:

$$\boldsymbol{p}(\boldsymbol{f}(\boldsymbol{X}) \,|\boldsymbol{X}) = \boldsymbol{N}\,(\boldsymbol{\mu}(\boldsymbol{X}), \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})) \tag{2.4}$$

More importantly, GPs allow us to condition the predictive distribution over unseen points $\boldsymbol{X}_*$ on (possibly noisy) observations of $\boldsymbol{f}$. Let $\boldsymbol{Y} = \boldsymbol{f}(\boldsymbol{X})$ be noisy observations of $\boldsymbol{f}$ at training inputs $\boldsymbol{X}$, and let $\boldsymbol{f}_* = \boldsymbol{f}(\boldsymbol{X}_*)$ be the test outputs of interest. Then the joint distribution implied by the GP is:

$$p\begin{pmatrix} Y \\ f_* \end{pmatrix} = N\left(\begin{bmatrix} \mu(X) \\ \mu(X_*) \end{bmatrix}, \begin{bmatrix} K(X,X) + s^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right)$$

An application of Bayes' rule yields:

$$p(f_*|X_*, D) = N\left(\mu_{f|D}(X_*), K_{f|D}(X_*, X_*)\right)$$

where

$$\mu_{f|D}(x) = \mu(x) + K(x,X)\left(K(X,X) + s^2 I\right)^{-1}(Y - \mu(X))$$

$$K_{f|D}(x,x') = K(x,x') - K(x,X)\left(K(X,X) + s^2 I\right)^{-1}K(X,x')$$

(Rasmussen and Williams 2006).

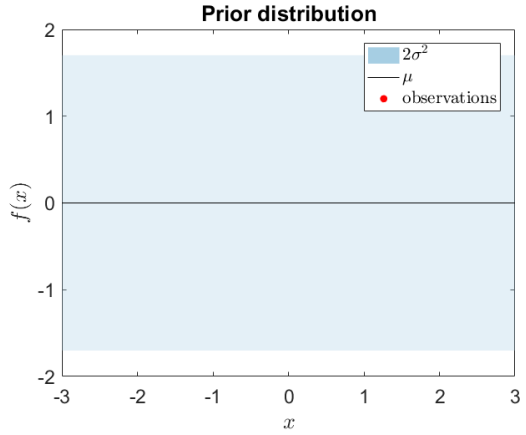

Figure 3: GP prior mean and variance

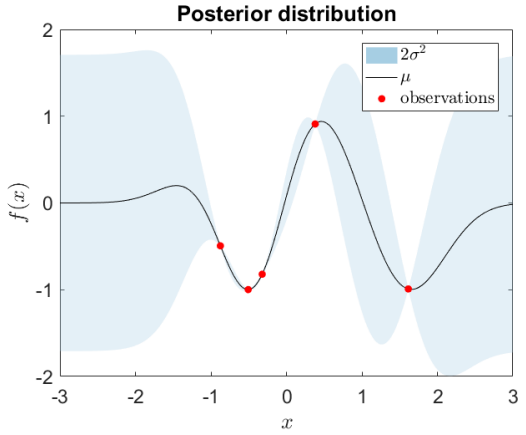

Figure 4: GP posterior mean and variance after 5 observation

2.3.3. Gaussian Process Classification

In classification problems, the target function shifts from producing real valued outputs, i.e. $y \in \mathbb{R}, y(x_i) = f(x_i) + \epsilon$, to a discrete space, where y_i can take on a fixed nmber of classes $C_1, \ldots, C_m$. Of particular interest in this thesis is the special case of binary classification, where outputs can take on one of two classes: $y_i \in \{0, 1\}$. Linear classification methods instead assume that the class-conditional probability of belonging to the "positive" class is a nonlinear

transformation of an underlying function known as the latent function. This applies the following transformation to equation the likelihood:

$$\boldsymbol{p(y(x_i) = 1) = \Phi\big(f(x_i)\big)} \tag{2.5}$$

$\boldsymbol{\phi}$ can be any "sigmoid" (s-shaped) function. Common choices of sigmoidal functions include the logistic function $\boldsymbol{\phi(x) = \frac{\exp(x)}{1+\exp(x)}}$ and the cumulative Gaussian $\boldsymbol{\sigma(x) = \int_{-\infty}^{x} \frac{e^{-z^2}}{\sqrt{2\pi}} dz}$

There is one further complication to the GP Classification problem. From Bayes' rule, the posterior distribution can be written as:

$$\boldsymbol{p(f\,|D) = \frac{1}{Z}\, p(\,f\,|X)\, p(y\,|f) = N(X,X))} \prod_{\boldsymbol{i}} \boldsymbol{p(y_i|f_i)} \tag{2.6}$$

Where $\boldsymbol{Z}$ is a normalization factor that is approximated in the schemes discussed below. In the regression setting, the posterior distribution is easy to work with directly because it is the product of a Gaussian prior and a Gaussian likelihood. However, likelihood is sigmoidal in the classification setting. Unfortunately, the product of a Gaussian distribution with a sigmoidal function does not produce a tractable posterior distribution. The model must instead approximate the posterior. Common approximation schemes include expectation propagation and Laplace approximation (Rasmussen and Williams 2006). Laplace approximation attempts to approximate the posterior distribution by fitting a Gaussian distribution to a 2nd order Taylor expansion of the posterior around its mean. Expectation propagation attempts to approximate the posterior distribution by matching the first and second moments, the mean and variance, of the posterior distribution.

2.3.4. Hyperparameter Selection

It was previously mentioned that Kernel functions encode information about the shape and smoothness of the functions drawn from a GP. While the GP itself is a nonparametric model, many Kernel functions themselves have parameters known as hyperparameters $\boldsymbol{\theta}$. The setting of hyperparameters exerts great influence over the predictive distribution of the GP. For instance, the popular squared exponential kernel is parameterized by its length scale $\boldsymbol{\ell}$ and output variance $\boldsymbol{\sigma}$ (Duvenaud 2014)

$$\boldsymbol{K(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\,\ell}\right)} \tag{2.7}$$

The model belief about the hyperparameters can be computed again via Bayes rule:

$$\boldsymbol{p(\theta|D, H) = p(Y \mid X, \theta)\, p(\theta|H)} \tag{2.8}$$

where $\boldsymbol{p(\theta|H)}$ is the hyperparameter prior, which can be used to encode domain knowledge about the settings of hyperparameters or may be left uninformative (Rasmussen and Williams 2006). This posterior distribution is often computationally intractable, and thus settings of the hyperparameters may be chosen through optimization algorithms such as gradient descent.

2.3.5. Active Learning

One notable advantage of the GP model is that its probabilistic predictions give rise to a set of techniques collectively known as "active learning." Active learning, sometimes called "optimal

experimental design," allows a machine learning model to choose the data it samples to perform better with less training (Settles 2011). To contrast with adaptive techniques, queries in active learning are chosen in such a way as to minimize some utility function. For example, an active learning query may select a point designed to minimize the expected error of the model against the latent function. In general, the application of active learning proceeds as follows: first, use the existing model to classify unseen data; next, find the "best" next point to query based on some objective function and query the data via an oracle (for instance, a human expert); finally, retrain the classifier and repeat these steps until satisfied.

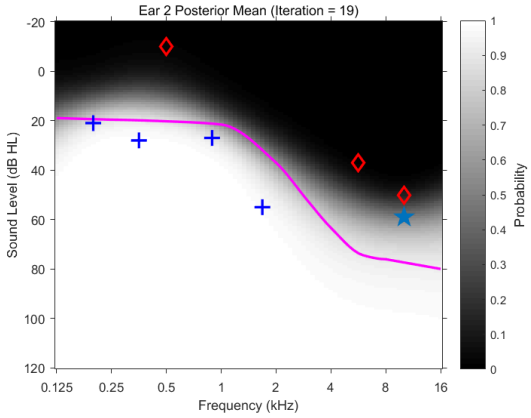

Figure 5: GP Audiogram posterior mean with next point

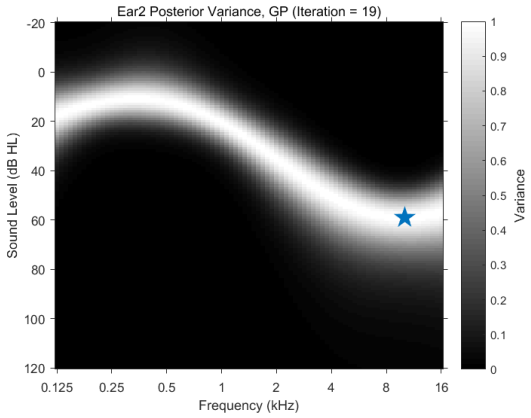

Figure 6: GP Audiogram posterior variance with next point

The most common form of active learning is uncertainty sampling (Lewis and Gale 1994, Settles 2011). Models employing uncertainty sampling will query areas about which the model is most uncertain. In the case of probabilistic classification, including GP classification, uncertainty sampling corresponds to querying the instance whose probability of being positive is closest to 0.5. This model can rapidly identify a class boundary for a target function of interest. The performance of

this method in estimating an underlying function degrades if the function is itself probabilistic instead of binary, for instance if the target function models some probability of stimulus detection. Because uncertainty sampling always attempts to query exactly where $\boldsymbol{p(y = 1\,|x, D) = 0.5}$, the model under-explores the input space. In the context of psychometric functions, the model cannot learn effectively about slope because every query under uncertainty will occur at the best current estimate of the threshold.

Bayesian Active Learning by Disagreement (BALD) attempts to circumvent this problem via an information theoretic approach (Houlsby, Huszár et al. 2011). The BALD method assumes the existence of some latent hyperparameters $\boldsymbol{\theta}$ that control the relationship between inputs and outputs $\boldsymbol{p(y\,|x, \theta)}$. For example, when performing GP regression with a squared exponential kernel, $\boldsymbol{\theta}$ would be the length scale and noise parameters. Further, under the Bayesian framework, it is possible to infer a posterior distribution over the parameters $\boldsymbol{p(\theta|D)}$. Each possible setting of $\boldsymbol{\theta}$ represents a distinct hypothesis about the relationship between inputs and outputs. The goal of the BALD method is to reduce the number of viable hypotheses as quickly as possible by minimizing the entropy of the posterior distribution of $\boldsymbol{\theta}$. To that end, BALD queries the point $\boldsymbol{x}$ to maximize the decrease in expected entropy:

$$\underset{\boldsymbol{x}}{\mathbf{argmax}}\, \boldsymbol{H}[\boldsymbol{\theta}\,|\boldsymbol{D}] - \mathbb{E}_{\boldsymbol{y} \sim \boldsymbol{p}(\boldsymbol{y}\,|\boldsymbol{x},\boldsymbol{D})}[\;\; \boldsymbol{H}[\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{D}]] \tag{2.9}$$

Where $\boldsymbol{H}[\boldsymbol{\theta}|\boldsymbol{D}]$ is Shannon's entropy of $\boldsymbol{\theta}$ given $\boldsymbol{D}$. This expression can be difficult to compute directly because the latent parameters often exist in high dimensional space. However, equation (2.9) can be rewritten in terms of entropies in the1-dimensional output space as follows:

$$\underset{\boldsymbol{x}}{\mathbf{argmax}}\, \boldsymbol{H}[\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{D}] - \mathbb{E}_{\boldsymbol{\theta} \sim \boldsymbol{p}(\boldsymbol{\theta}|\boldsymbol{D})}[\boldsymbol{H}[\boldsymbol{y}\,|\boldsymbol{x}, \boldsymbol{\theta}]] \tag{2.10}$$

This expression can be computed in $\boldsymbol{O(1)}$ time (Houlsby, Huszár et al. 2011), making it easy to work with in practice. In (2.10), BALD selects the $\boldsymbol{x}$ for which the entire model is most uncertain about $\boldsymbol{y}$ (high $\boldsymbol{H[Y|X,d]}$), but for which the individual predictions given a setting of the hyperparameters are very confident. This can be interpreted as "seeking the $\boldsymbol{x}$ for which the parameters under the posterior disagree about the outcome the most," (Houlsby, Huszár et al. 2011)

3.Methods

3.1. Introduction

GP Classification was used to simultaneously estimate the right-ear and left-ear audiograms of simulated subjects. The model produces continuous audiogram estimates across the entire input space and tones were actively sampled to reduce the number of stimuli required to achieve acceptable error thresholds. The goals of this experiment were 1) to achieve error thresholds comparable to current state-of-the-art methods for audiogram estimation and 2) to achieve these results across both ears faster than would be otherwise possible with disjoint audiogram estimation.

3.2. Simulations

Simulated subjects have separate audiograms for each ear. These audiograms define the probability of stimuli detection over a two-dimensional input space consisting of frequency and intensity. Audiogram shapes were defined by one of four human audiogram phenotypes: older-normal, sensory, metabolic, and metabolic + sensory (Dubno, Eckert et al. 2013).

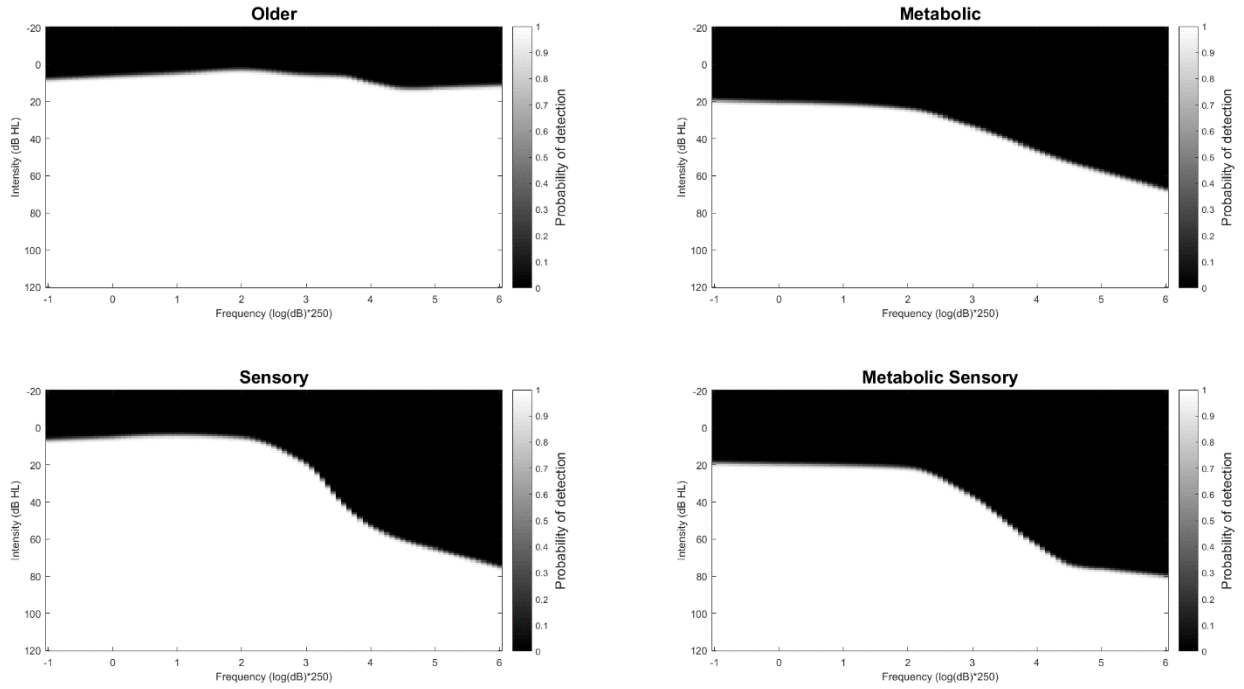

Figure 7: Simulated audiograms for each of the four human phenotypes identified by Dubno

In the context of this work, threshold is defined as a point $\boldsymbol{x}$ such that $\boldsymbol{p(y = 1|x) = 0.5}$. These standard phenotypes provide threshold estimates at octave frequencies, as would be typically observed by the HW procedure. Spline interpolation and linear extrapolation were used to generate a continuous threshold estimation across frequency space. At each frequency, a cumulative Gaussian was used to generate a sigmoidal psychometric curve to generate probability of tone detection outside of threshold (Song, Garnett et al. 2017). The cumulative Gaussian was parameterized by the intensity and threshold $\boldsymbol{(x, \mu)}$ as follows:

$$\boldsymbol{p(y = 1\,|x, \mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}} \tag{3.1}$$

Subject response is recorded by drawing a random number from the (0, 1) uniform distribution. A stimulus is recorded as "observed" if the random number is less than the probability of tone detection for that $\boldsymbol{(f, I)}$ under that ear's phenotypic model.

The four common phenotypes fall into either normal (older-normal) or pathologic categories (metabolic, sensory, metabolic+sensory). As such, simulations were run on three pairings of audiograms to reflect possible conjoint hearing conditions. These were: normal, in which simulated subjects have the older-normal phenotype in both ears; symmetric hearing loss, in which simulated subjects have metabolic hearing loss in one ear and metabolic+sensory hearing loss in the other; and asymmetric, in which simulated subjects have older-normal hearing in one ear and metabolic+sensory hearing loss in the other. Asymmetric hearing was defined by the two phenotypes with the greatest difference in threshold to demonstrate the flexibility of this model.

The results of four models were compared to determine relative sample efficiency and inference accuracy. The first model is the existing framework for GP audiogram estimation (Song, Garnett et al. 2017). This approach uses two GP models that do not share information, and queries alternate between the two input spaces. This functions as a control group to compare with other models. The second model uses conjoint audiogram estimation, but artificially constrains the model to alternate ears in its sampling. The motivation for this model is to see improvement in accuracy or efficiency just through an extension of the input space and covariance function. This model also provides an easy direct comparison to model 1 for explanatory purposes. Sampling in the third model is unconstrained. This allows the model to query one ear multiple times in a row if it deems fit. Finally, a fourth model was run using Halton sampling (Halton 1964) to demonstrate the importance of active sampling. All tests in all models were run with the same mean, likelihood, and inference functions. Models 2 and 3 use the same kernel. Model 1 uses the kernel described in 3.3.3 without the multiplicative inter-ear covariance.

Simulated subjects representing each of the three phenotypic pairings were tested 10 times under each of the three models. Each test consisted of 100 observations of data.

3.3. Gaussian Process Framework

3.3.1. Variable Space

Traditional pure-tone audiometry involves delivering tones in a two-dimensional continuous feature space, frequency and intensity. In this study, we augment the feature space to include a third discrete "ear" dimension, i.e. $\boldsymbol{x_i} = (\boldsymbol{f_i}, \boldsymbol{I_i}, \boldsymbol{e_i})$. In querying a simulated subject's audiogram, the model chooses in which ear to deliver the tone in addition to the frequency and intensity of tone delivered. Binary responses were recorded for each simulated tone delivery.

3.3.2. Mean Function

The model uses a constant mean function, $\boldsymbol{\mu(x) = c \; \forall x \in X}$. While this mean function is not representative of any of the phenotypic audiograms, deviation from the mean is captured in the posterior distribution of the GP Classification model.

3.3.3. Kernel Function

The GP Kernel function was derived from prior knowledge about the behavior of audiograms. Knowing that a subject's psychometric curve for any frequency is sigmoidal allows us to place a linear kernel in the intensity dimension:

$$\boldsymbol{K_I(x, x') = I \cdot I'} \tag{3.2}$$

Further, the model leverages the continuity of audiogram threshold by placing an isotropic squared exponential kernel with unit magnitude over the frequency domain

$$\boldsymbol{K_f(x, x') = e^{-\frac{(f-f')(f-f')^T}{2\ell}}} \tag{3.3}$$

where $\boldsymbol{\ell}$ is the length scale.

Finally, the model must incorporate some sort of covariance between ears. For this, the model uses a discrete covariance function which directly parameterizes relationships between every pair of points in the discrete space.

$$\boldsymbol{K_e(x, x') = \begin{cases} s_{11} \text{ if } x, x' \in e_1 \\ s_{12} \text{ if } e \neq e' \\ s_{22} \text{ if } x, x' \in e_2 \end{cases}} \tag{3.4}$$

This model can explicitly define the covariance between ears without having to relate them via some functional form. Computationally, this is done by modeling the discrete covariance as the Cholesky decomposition of a 2×2 matrix, $\boldsymbol{K = \Lambda\Lambda^T}$.

Finally, the model combines the covariance functions as follows:

$$\boldsymbol{K(x, x') = K_e(x, x') \times \left(K_f(x, x') + K_I(x, x')\right)} \tag{3.5}$$

3.3.4. Likelihood Function

The model uses the cumulative Gaussian likelihood function for binary classification, which is both standard for GP classification and accurately captures the sigmoidal behavior of psychometric functions.

3.3.5. Inference Function

The exact form of the posterior requires computing the product of the likelihood and prior distributions. In the case of GP Classification, the product of a Gaussian distribution with a sigmoidal function does not produce a tractable posterior distribution. The model must instead approximate the posterior. This model uses Expectation Propagation (EP) to approximate the posterior. Under EP inference, the model approximates each of the sigmoid likelihoods with moment-matching Gaussian distributions to derive a Gaussian posterior distribution (Gelman, Vehtari et al. 2014).

3.3.6. Active Sampling

Simulations were run until 100 data points had been collected. Because the model tends to be less accurate with very little data ($n << 10$), traditional active learning procedures would query regions that prior knowledge would indicate are very uninformative, for example extremely quiet tones (dB < -10). Thus, the first 15 points are delivered via a modification of Halton sampling. The typical Halton sampling method produces "well spaced" draws from the feature space (Halton 1964). In this modification, Halton samples were constrained to deliver tones below 60dB to protect the subject's hearing in the event of clinical application. The remaining 85 observations were delivered via BALD. The BALD procedure is described in greater detail in section 2.3.5.

3.3.7. Hyperparameter Learning

The GP classification model is fully parameterized by its mean and covariance functions. In the case of this model, the constant mean has one parameter, namely the constant, and the custom kernel function has four: one for the length of the squared exponential kernel, and three for the discrete kernel. Because the posterior distribution of the hyperparameters $\boldsymbol{p(\theta|D)}$ may be multimodal, standard gradient descent approaches run the risk of getting stuck in a local maximum. To circumvent this issue, the model performs gradient descent on two settings of hyperparameters after each observation. The first setting of hyperparameters comes from the most recent results of the model (or the hyperparameter prior in absence of any data). The second setting of hyperparameters is drawn from a multivariate Gaussian distribution whose mean is the hyperparameter prior derived in the following section. Gradient descent is performed on both settings of the hyperparameters, and the setting with higher likelihood $\boldsymbol{p(D\,|\theta)}$ is kept for the next iteration.

3.3.8. Hyperparameter Prior Selection

The first iterations of the model suffered greatly from inefficient early sampling. It was clear that the initial settings of the hyperparameters did not accurately model the types of audiograms that would be seen in a clinical setting. Fixing this issue involved learning reasonable priors on the hyperparameters to serve as a strong starting point for the model.

Each of the four common human phenotypes has at least one optimal setting to its hyperparameters to minimize model error. Because the kernel function is symmetric, there are ten unique pairs of

audiogram profiles that can be derived from the four phenotypes. Data was collected for each of the ten audiogram profile pairs far in excess of what would be collected in a clinical setting. First, 400 stimuli were delivered across both ears using Halton sampling. Then, an additional 100 stimuli were queried via BALD to gain additional sampling density around the threshold. Hyperparameters were learned using the modified gradient descent method discussed in 3.3.7. The same concept was repeated with varying numbers of Halton and BALD queries, but the final settings of the hyperparameters converged to within 2% of each after about 300 samples. The final setting of the hyperparameter priors was computed by taking an average of the hyperparameters in each of the ten pairs, weighted by the prevalence of those phenotypes in human populations (Dubno, Eckert et al. 2013). This method assumes that the phenotype for one ear is independent of the phenotype of the other ear in the same subject. This is not the case, and a possible improvement of the model would involve weighting each setting of the hyperparameters by the prevalence of that pair of phenotypes in humans.

3.3.9 Evaluation

I evaluated the performance of two variants of the conjoint audiogram estimation technique described above, and compared these results with those of the existing GP audiogram framework, which served as a baseline. All together, the performance of the following three methods were compared for each of my test cases:

- *Disjoint GP Audiogram Estimation (Disjoint)*: This method performs inference using two separate models of the existing GP audiogram framework. (Song, Garnett et al. 2017). Information in this approach is not shared between ears. Tone delivery alternated between left- and right- ears.

- *Alternating Conjoint GP Audiogram Estimation (Alternating Conjoint):* This method performs inference using the conjoint audiogram estimation extension of GPs described above. However, this method is artificially constrained to alternate samples between the left- and right- ears. This approach was included in hopes of demonstrating that conjoint audiogram estimation outperforms disjoint audiogram estimation even with the same sampling scheme.

- *Unconstrained Conjoint GP Audiogram Estimation (Unconstrained Conjoint):* This method also performs inference using the conjoint audiogram estimation extension of GPs described above. This method gives the model complete choice over which ear to query, as well as which frequency / intensity pair to deliver. This occasionally results in multiple stimuli being delivered to the same ear, particularly in cases where the model is more unsure of the audiogram in one ear than the other.

Each of the four phenotypes identified by Dubno could be further classified into either "normal" hearing or "pathological" hearing. Normal hearing was identified by the older-normal phenotype, whereas pathological hearing could be any of the metabolic, sensory, or metabolic + sensory phenotypes. From here, I identified three cases of interest to demonstrate the flexibility of the conjoint audiogram estimation framework:

- *Case 1 - Older Normal*: This case was defined as having the older-normal phenotype in both ears

- *Case 2 – Asymmetric Hearing Loss*: This case was defined as having the older-normal phenotype in one ear, and the metabolic + sensory phenotype in the other ear. This case represents more severe asymmetric hearing loss than is typical in human populations (Song, Wallace et al. 2015) but was included to demonstrate the flexibility of the conjoint models.

- *Case 3 – Symmetric Hearing Loss*: This case was defined as having metabolic + sensory hearing loss in one ear, and sensory hearing loss in the other. This case is more typical of hearing loss in human subjects. The sensory and metabolic + sensory phenotypes have slightly different thresholds. Distinct phenotypes were chosen for this case to more accurately reflect presentations of hearing loss in human subjects, where left- and right- ear audiograms are not typically identical.

I ran ten tests of each case-model pair, for a total of 90 tests. For each test, 100 tones were delivered to the simulated subjects. To avoid unstable hyperparameter learning and uninformative early querying, the first 15 tones were delivered via a modified Halton sampling algorithm. The modified Halton sampling algorithm constrained tone deliveries to be below 60dB. This prevents damaging subject hearing if this approach were to be tested in humans. Subsequent tones are sampled via BALD, with constraints for the disjoint and alternating conjoint cases as discussed above. Hyperparameters are learned via a modified gradient descent algorithm every iteration starting with iteration 16. Hyperparameter learning is off for the first 15 iterations of each test to prevent model instability. For each tone delivery I recorded the model posterior distribution across the entire input space. From here, I derived the $\boldsymbol{x}$ intercept of the latent function to calculate the 50% threshold, also known as $\boldsymbol{\alpha}$ for the model over a fine grid of frequencies from 0.125kHz to 16kHz. I evaluated accuracy for a single test using the mean absolute error between the estimated $\boldsymbol{\alpha}$ and the true $\boldsymbol{\alpha}$ at each frequency in the grid. Results were then averaged at each iteration across all 10 tests, to get the average $\boldsymbol{\alpha}$ error per iteration for each of the three models in each of the three cases. In addition to comparing average $\boldsymbol{\alpha}$ error per iteration, I also examined the average number of iterations required to have less than 5dB $\boldsymbol{\alpha}$ error in both ears. This value was chosen as a measure of "convergence"

because it is the minimum step size in the Hughson Westlake procedure. Thus, once the model is within 5dB $\boldsymbol{\alpha}$ error in both ears, it is within the margin of error for the Hughson Westlake procedure.

4. Results

Figure 8 depicts a representative run of the conjoint audiogram estimation algorithm. The ground truth for this figure was the asymmetric hearing loss case identified by older-normal hearing in one ear and metabolic + sensory hearing loss in the right ear. The first 15 samples were selected via Halton sampling to improve the stability of the GP model. Subsequent tones were sampled using BALD. The samples selected via BALD cluster around the predicted threshold, where they are more informative about the true audiogram threshold. Note that after 14 tone deliveries, the conjoint GP model has not yet identified the microstructure in the older-normal ear and is very unconfident about the threshold location in the metabolic + sensory ear. After 98 tone deliveries, the model has correctly identified the microstructure in the older-normal ear and confidently and accurately identifies the threshold in the metabolic + sensory ear.

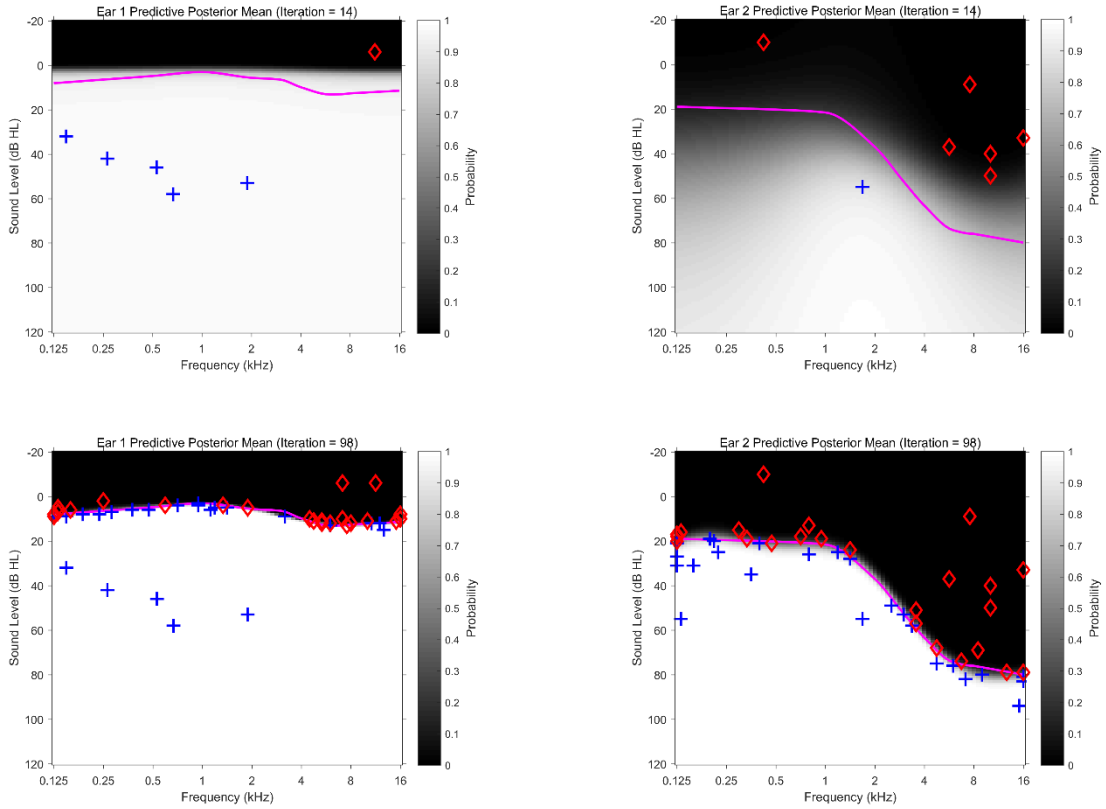

Figure 8: Example of posterior mean for simulated asymmetric hearing loss as estimated by the alternating conjoint GP audiogram estimation model. In all images, the predicted probability in tone detection is shown in grayscale. Blue plusses are heard stimuli and red diamonds are unheard stimuli. The true threshold from the simulation is shown in pink. (Top) Posterior mean after 14 samples. (Bottom) Posterior mean after 98 samples.

4.1 Case 1: Older Normal Hearing

Figure 9 shows the average $\boldsymbol{\alpha}$ error per iteration for case 1, which was defined as having the older-normal phenotype in both ears. Note that both conjoint approaches outperform the disjoint approach, particularly in the early iterations. While hyperparameter priors were learned in the same fashion, the disjoint approach has an additional multiplicative noise term for its squared exponential frequency kernel. It is possible that this additional hyperparameter leads to a degradation in early performance. However, as the models progress, they all approach around 1dB mean $\boldsymbol{\alpha}$ error.

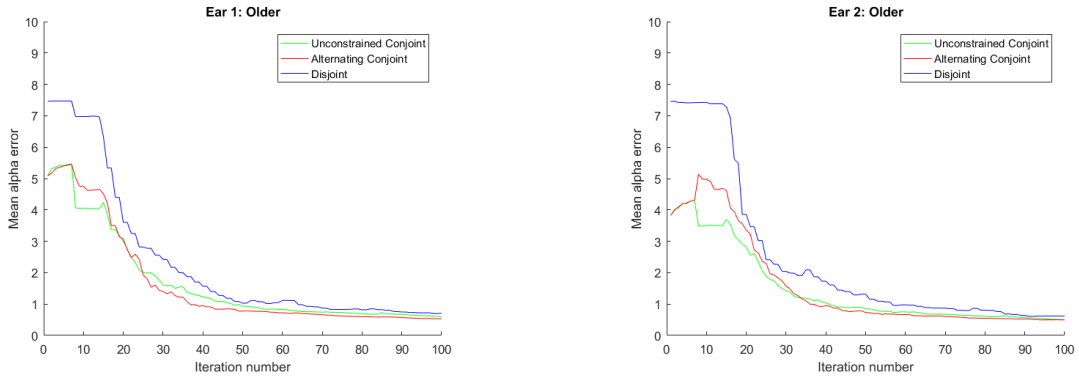

Figure 9: Mean α error per iteration case 1, no hearing loss

Figure 10 shows the average number of iterations required for each model to achieve 5dB average $\boldsymbol{\alpha}$ error in the normal hearing case. Both the unconstrained and alternating conjoint approaches require less than 2/3 the samples required in the disjoint approach. However, the conjoint approaches tend to have higher standard deviation than the disjoint approach. This is possibly because initial differences in the Halton sampling algorithm reinforce the constant threshold belief more in some iterations than others, and this constant threshold belief is stronger with more evidence, as would be the case if samples were shared among ears. It is worth noting that the 2nd ear of the older normal phenotype never observes higher than 5dB average $\boldsymbol{\alpha}$ error. This is because the older-normal phenotype is relatively constant (see **Figure 7**), and the GP has a constant mean prior. This makes the prior belief of the model much better for the older-normal phenotype than it is for any of the pathologic phenotypes.

	Ear 1: Older Normal	Ear 2: Older Normal
Unconstrained Conjoint	11.5 $\pm$4.5	0 $\pm$0*
Alternating Conjoint	15.3 $\pm$9.7	9.6 $\pm$10.3
Disjoint	18.1 $\pm$2.6	19.1 $\pm$2.5

Figure 10: Mean number of iterations required to achieve 5 dB α error for each ear, case 1

4.2 Case 2: Asymmetric Hearing Loss

Figure 11 shows the mean $\boldsymbol{\alpha}$ error per iteration for case 2, which was defined as having the older-normal phenotype in one ear and the metabolic + sensory phenotype in the other. Note that both conjoint approaches outperform the disjoint approach, particularly in the metabolic + sensory ear. It is also worth mentioning that the unconstrained conjoint method chooses to sacrifice some early performance in the older-normal ear in exchange for faster convergence in the metabolic + sensory ear. The unconstrained approach is able to make this choice because the model uncertainty is higher in early iterations on the metabolic + sensory phenotype than it is on the older-normal phenotype (**Figure 8**).

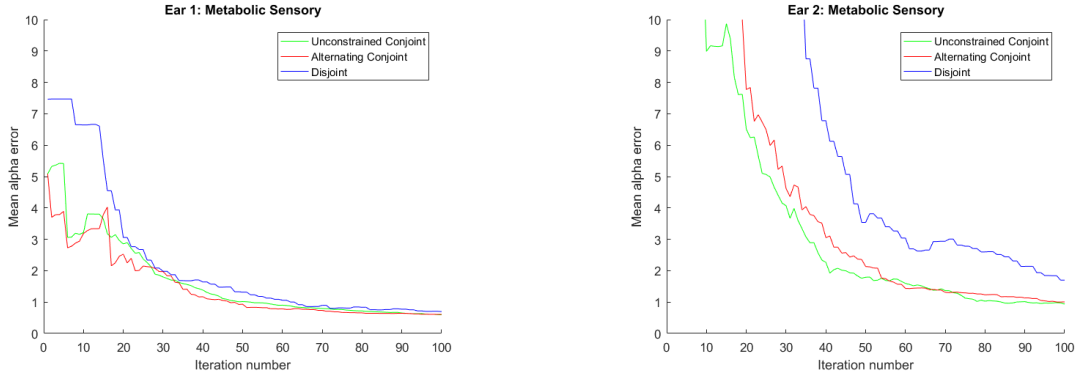

Figure 11: Mean α error per iteration, asymmetric hearing loss

Figure 12 shows the average number of iterations required for each model to achieve 5dB average $\boldsymbol{\alpha}$ error in the asymmetric hearing loss case. Both the unconstrained and alternating conjoint approaches require less samples than were required in the disjoint approach. As was observed in **Figure 11**, the unconstrained conjoint method sacrifices some early performance in estimating the older-normal phenotype in exchange for faster convergence in the metabolic + sensory phenotype. Thus, the number of tones required to achieve convergence in both ears is lower for the unconstrained conjoint approach than it is for the alternating conjoint approach. The limiting factor in all three methods was identifying the metabolic + sensory phenotype, which took an average of

27.9 tone deliveries split across both ears in the unconstrained case and 32.9 tone deliveries split across both ears in the alternating case. Put another way, unconstrained conjoint only requires 84.8% of the samples required for alternating conjoint to achieve convergence in both ears in this case.

	Ear 1: Older Normal	Ear 2: Metabolic Sensory
Unconstrained Conjoint	10 $\pm$5.2	27.9 $\pm$3.9
Alternating Conjoint	5.8 $\pm$6.1	32.9 $\pm$5.3
Disjoint	16.9 $\pm$2.3	46 $\pm$6.1

Figure 12: Mean number of iterations required to achieve 5 dB α error for each ear, case 2

4.3 Case 3: Symmetric Hearing Loss

Figure 13 shows the average $\boldsymbol{\alpha}$ error per iteration for case 3, which was defined as having the metabolic + sensory phenotype in one ear and the sensory phenotype in the other. Once again, both conjoint approaches outperform the disjoint approach, particularly in the metabolic + sensory ear. In this case, the unconstrained conjoint approach can leverage its ability to choose in which ear to deliver stimuli and achieves substantially faster convergence than even the alternating conjoint approach. Further, all three models continue to have more difficulty identifying pathological phenotypes. This suggests that there is room for improvement by using a more informative prior mean.

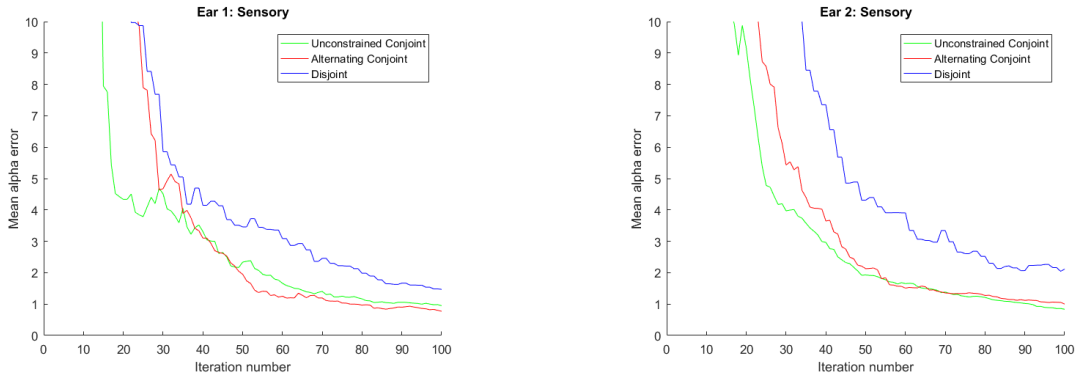

Figure 13: Mean α error per iteration, symmetric hearing loss

Figure 14 shows the average number of iterations required for each model to achieve 5dB average $\boldsymbol{\alpha}$ error in the symmetric hearing loss case. Both the unconstrained and alternating conjoint approaches require less samples than were required in the disjoint approach. Unlike in case 2, the unconstrained conjoint model can leverage its knowledge of inter-ear correlation to drastically improve the time to convergence in both ears by changing the distribution of ear samples. As a result, the unconstrained conjoint method is able to converge in both ears faster than the alternating conjoint approach can converge in either. Further, in this case, the unconstrained conjoint approach converges in both ears using 55.1% of the samples required in the disjoint approach.

	Ear 1: Metabolic Sensory	Ear 2: Sensory
Unconstrained Conjoint	29.2 $\pm$7.0	27.7 $\pm$4.1
Alternating Conjoint	31.1 $\pm$6.3	35.5 $\pm$5.0
Disjoint	41.8 $\pm$8.1	53 $\pm$11.4

Figure 14: Mean number of iterations required to achieve 5 dB αerror for each ear, case 3

4.4 Summary Results

Figure 15 shows the average number of tones required for each of the models to achieve below 5dB $\boldsymbol{\alpha}$ error in both ears for each case. Numbers here were selected as the last time the average $\boldsymbol{\alpha}$

error per iteration crossed below 5dB in both ears. This prevents the numbers from being overly optimistic. It is possible for the conjoint GP approach to start with a "lucky guess" of the true audiogram and have average $\boldsymbol{\alpha}$ error below 5dB for early iterations but have the $\boldsymbol{\alpha}$ error go up in early iteration, which tend to be more unstable. Presenting an average of the final cross below convergence solves this issue. To summarize the results of the three presented cases, both conjoint methods outperform the disjoint approach for every case. It is also apparent that the constant mean assumption performs substantially better on older-normal phenotypes than it does on any of the pathological phenotypes.

	Older Normal	Asymmetric	Symmetric
Unconstrained Conjoint	11.5 $\pm$4.5	27.9 $\pm$3.9	32.1 $\pm$4.7
Alternating Conjoint	17.3 $\pm$4.9	32.9 $\pm$5.2	36.1 $\pm$4.8
Disjoint	19.9 $\pm$2.4	46.0 $\pm$6.1	55.0 $\pm$8.2

Figure 15: Mean number of tones required for each of the three models to achieve better than 5dB average α error per iteration

Figure 16 reframes the results of **Figure 15** in relative terms. Regardless of the phenotype pairings, the unconstrained conjoint approach demonstrates approximately a 40% speedup in sampling efficiency over the disjoint approach. One interesting observation that is made clear from this example is that the performance of the alternating conjoint method is closest to that of the unconstrained conjoint method in the case of asymmetric hearing loss. This is because the conjoint model learns that there is less correlation between the two ears and, to learn a good audiogram estimation, must split its samples more evenly among both ears. This causes the unconstrained conjoint and alternating conjoint approaches to exhibit similar querying strategies in the asymmetric case, whereas in the older-normal and symmetric cases, their querying strategies are quite different.

	Older Normal	Asymmetric	Symmetric
Unconstrained Conjoint	60.65% $\pm$22.6%	58.36% $\pm$8.5%	57.79%$\pm$8.5%
Alternating Conjoint	71.5% $\pm$24.6%	65.63% $\pm$11.3%	86.93%$\pm$8.7%
Disjoint	100% $\pm$12%	100% $\pm$13.3%	100%$\pm$14.9%

Figure 16: Percentage of tones relative to disjoint required for each of the three models to achieve better than 5dB average αerror per iteration

5.Discussion

Bayesian methods continue to show promise in the estimation of psychometric functions. Posterior distributions allow for active sampling techniques which can produce fast, accurate psychometric estimations, even in the three-dimensional space explored here. This, coupled with the ability to encode domain specific knowledge in the form of a prior, gives Bayesian methods the robustness and flexibility to see substantial clinical application.

Gaussian Processes also represent a significant conceptual shift for psychometrics. Psychometric function estimation has typically been a parametric task. The Gaussian Process model allows a diagnostician to infer directly about a patient's response to stimuli, hopefully allowing them to make clearer diagnoses. A nonparametric model could possibly bring new insights into pathologies that were previously unexplored because they were too complex to be modelled effectively by the clinical standard psychometric function for that stimulus.

To my knowledge, this is both the first application of GP classification to the estimation of multiple psychometric functions simultaneously and the first application of GP classification to a three-dimensional psychometric input space. The ability of GPs to efficiently sample higher dimensional input spaces allows them to extend to more complex problems.

There are several directions for future application in this space. First, these results need to be confirmed in a clinical setting. In the simulation space, one notable weakness of the approach was a relative dearth of ground truth audiograms from which to test. A reasonably straightforward extension of this work would be to assess the same model against a wider distribution of simulated audiograms. It is possible that there are pathologies that this model could not capture. I suspect that

this model may perform poorly if an individual has a very smooth threshold in one ear and a very notched threshold in the other, as may be the case for individuals with high levels of asymmetric noise exposure, for instance marksmen.

There are ways to extend this research beyond more rigorous testing. First, all three models had less trouble identifying the older-normal threshold than any of the pathological thresholds. This is because the constant mean assumption, while reasonable for older-normal, does not accurately reflect the pathological phenotypes. A more robust GP framework would encode some mean function that varies with respect to frequency. In each pathological phenotype, there is a region of constant threshold in low frequencies, followed by a decreasing threshold in high frequencies. The GP mean function should be able to model this behavior. Implementing this change would likely confer further increases in efficiency.

The conjoint GP model also allows for easy extension to higher dimensional discrete spaces. For example, the conjoint GP could also choose whether to deliver a tone via bone conduction or air conduction. This would help identify pathologies that are not identified by the current approach to conjoint audiogram estimation.

6.Conclusion

The goal of this thesis was to develop a framework for extending the input space of existing GP models and performing inference between discrete psychometric functions. A first step in moving forward in this line of research would be validating the simulated results with human studies. Another natural progression would be to apply the GP classification framework to other domains, for example vision or behavior. On the machine learning front in the audiology space alone, there are many directions in which this work could proceed. First, audiologists use both air-conduction

and bone-conduction pure tone audiometry to assess the source of a patient's hearing loss. It would be a nearly identical exercise to extend this framework to also allow for choice between delivering a tone via air conduction or bone conduction. A slightly more involved extension of the GP framework could add masking, a third continuous dimension that is delivered in the opposite ear to ensure that subject response to stimulus comes from the target ear. One notable challenge for masking would be developing ground truth data. Clinical state of the art involves using adaptive techniques to perform a limited grid search of the masking dimension along a fixed tone frequency and intensity. To this author's knowledge, the masking space has not been explored in as much depth as the frequency / intensity space. Finally, it would be particularly interesting to learn some warping function to infer between different tasks in the same domain.

Bayesian methods have taken time to gain traction amid much pushback from frequentist statisticians. However, the ability to encode prior information and hold probabilistic beliefs give Bayesian methods strength that is hard to deny. While GP classification models have some limitations, including poor scaling to large data relative to Neural Networks or SVMs (sparse GPs attempt to rectify this and are still an active area of research), they present a unique opportunity to gain inference about an individual subject in ways that were previously infeasible due to time or budgetary constraints. It is the hope of this author that one day Bayesian methods for diagnostics will be widely adopted by the medical community and will herald a shift in the way we perceive medicine.

The complexity of the GP model is a double-edged sword. While it can perform faster, more accurate audiogram estimates than current clinical methods, it is also substantially more difficult for clinicians to understand. I believe that for GPs to gain widespread clinical adoption, clinicians must

either a) receive substantial additional training in the new methods, or b) be willing to treat the querying and inference methods of the GP as a "black box" and simply rely on the results.

Irrefutable performance increases in the efficiency and accuracy of the GP audiogram could also lead to a significant increase in clinical adoption. The conjoint audiogram approach is an important step in this direction. Not only does the conjoint test confer a 40% speedup in sample efficiency over the existing GP audiogram approach (which itself is substantially faster than the Hughson Westlake procedure), but estimating both ears simultaneously allows the clinician to only perform one experimental setup. Additional tests can be incorporated by extending the model, with the goal of one day being able to perform an entire test battery simultaneously. At this point, the increases in efficiency and accuracy would be impossible to ignore and using approaches other than GP audiogram would be arcane or irresponsible.

References

Carhart, R. and J. Jerger (1959). "Preferred method for clinical determination of pure-tone thresholds." J Speech Hear Disord 24: 330-345.

Dubno, J. R., et al. (2013). "Classifying human audiometric phenotypes of age-related hearing loss from animal models." J Assoc Res Otolaryngol 14(5): 687-701.

Duvenaud, D. (2014). Automatic Model Construction with Gaussian Processes. Department of Engineering. Cambridge, England, University of Cambridge. Doctorate: 1-132.

Fechner, G. T. (1860). Elements of Psychophysics. New York, Holt, Rhinehart & Winston.

Gelman, A., et al. (2014). "Expectation propagation as a way of life." arXiv:1412.4869v2.

Guan, N. (2011). Bayesian Optimal Pure Tone Audiometry with Prior Knowledge. Electrical Engineering. Stockholm, Sweden, KTH Royal Institute of Technology.

**Hall, J. L. (1968). "Maximum-Likelihood Sequential Procedure for Estimation of Psychometric
Functions." The Journal of the Acoustical Society of America 44(1).**

Halton, J. H. (1964). "Algorithm 247: Radical-inverse quasi-random point sequence." Commun ACM 7(12): 701-702.

Houlsby, N., et al. (2011). "Bayesian active learning for classification and preference learning." arXiv preprint arXiv:1112.5745.

Hughson, W. and H. Westlake (1944). "Manual for program outline for rehabilitation of aural casualties both military and civilian." Trans Am Acad Ophthalmol Otolaryngol 48(Suppl): 1-15.

Ishak, W. S., et al. (2011). "Test-retest reliability and validity of Audioscan and Békésy compared with pure tone audiometry." Audiological Medicine 9(1): 40-46.

King-Smith, P. E., et al. (1994). "Efficient and unbiased modifications of the QUEST threshold method: theory, simulations, experimental evaluation and practical implementation." Vision Res 34(7): 885-912.

Levitt, H. (1971). "Transformed up-down methods in psychoacoustics." J Acoust Soc Am 49(2B): 467-477.

Lewis, D. D. and W. A. Gale (1994). A sequential algorithm for training text classifiers. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, Springer-Verlag New York, Inc.

Mahomed, F., et al. (2013). "Validity of automated threshold audiometry: A systematic review and meta-analysis." Ear and hearing 34(6): 745-752.

Meyer-Bisch, C. (1996). "Audioscan: a high-definition audiometry technique based on constant-level frequency sweeps - A new method with new hearing indicators." International journal of audiology 35(2): 63-72.

Mohri, M., et al. (2012). Foundations of Machine Learning, The MIT Press.

Rasmussen, C. E. and C. K. I. Williams (2006). Gaussian Processes for Machine Learning. Cambridge, MA, The MIT Press.

Schoenberg, I. J. (1946). "Contributions to the Problem of Approximation of Equidistant Data by Analytic Functions." Quarterly of Applied Mathematics 4(1): 55.

Settles, B. (2011). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin, Madison. 52.

Song, X. D., et al. (2017). "Psychometric function estimation by probabilistic classification." J Acoust Soc Am 141(4): 2513-2525.

Song, X. D., et al. (2015). "Fast, continuous audiogram estimation using machine learning." Ear and hearing 36 (6): e326-e335.

Taylor, M. M. and C. D. Creelman (1967). "PEST: Efficient estimates on probability functions." The Journal of the Acoustical Society of America 41(4A): 782-787.

Watson, A. B. and D. G. Pelli (1983). "QUEST: A Bayesian adaptive psychometric method." Perception & psychophysics 33(2): 113-120.

Williams, C. K. and C. E. Rasmussen (1996). Gaussian processes for regression. Adv Neural Inf Process Syst 8 (NIPS '95). D. Touretzky, M. Mozer and M. Hasselmo. Cambridge, MA, MIT Press.

Conjoint Psychometrics, DiLorenzo, M.S., 2017