Published in final edited form as:
J Math Psychol. 2012 June 1; 56(3): 179–195. doi:10.1016/j.jmp.2012.04.002.

A Predictive Approach to Nonparametric Inference for Adaptive Sequential Sampling of Psychophysical Experiments

Stephan Poppe[a], **Philipp Benner**[a], and **Tobias Elze**[b]

Stephan Poppe: stephan.poppe@mis.mpg.de; Philipp Benner: philipp.benner@mis.mpg.de; Tobias Elze: tobias.elze@schepens.harvard.edu

[a]Max Planck Institute for Mathematics in the Sciences, Inselstr. 22, 04103 Leipzig, Germany

[b]Schepens Eye Research Institute, Harvard Medical School, 20 Staniford Street, Boston, MA 02114

Abstract

We present a predictive account on adaptive sequential sampling of stimulus-response relations in psychophysical experiments. Our discussion applies to experimental situations with ordinal stimuli when there is only weak structural knowledge available such that parametric modeling is no option. By introducing a certain form of partial exchangeability, we successively develop a hierarchical Bayesian model based on a mixture of Pólya urn processes. Suitable utility measures permit us to optimize the overall experimental sampling process. We provide several measures that are either based on simple count statistics or more elaborate information theoretic quantities. The actual computation of information theoretic utilities often turns out to be infeasible. This is not the case with our sampling method, which relies on an efficient algorithm to compute exact solutions of our posterior predictions and utility measures. Finally, we demonstrate the advantages of our framework on a hypothetical sampling problem.

Keywords

Adaptive Sequential Sampling; Optimal Design; Active Learning; Predictive Inference; Psychophysics; Efficient Statistical Computations

1. Introduction and Motivation

The application of adaptive measurement methods have a long tradition in psychophysics. The need for such methods is mainly due to the limited number of measurements that can be taken during experiments. Most of the classical methods are motivated by their simplicity, both conceptually and computationally. With the advent of modern computers and the continuing progress in statistical theory, the development of more sophisticated adaptive sampling procedures has recently seen much progress.

Especially the consideration of Bayesian experimental designs based on the information theoretic description of experimental objectives and their numerical approximation (cf.

Correspondence to: Stephan Poppe, stephan.poppe@mis.mpg.de.

MacKay (1992); Chaloner and Verdinelli (1995)) has moved into the focus of contemporary research. Recent developments are, for instance, the Ψ-method introduced in Kontsevich and Tyler (1999), the consideration of multidimensional stimulus spaces in Kujala and Lukka (2006), and the framework of adaptive design optimization (ADO) for model discrimination proposed in Cavagnaro et al. (2010). Another interesting example is given by Kujala (2010) who considers random cost as a further constraint for experimental observations.

Our contribution to the field is twofold. First, we address the case where no particular statistical model in form of parametric curves can be assumed. We present a complete formal description of a suitable framework for such a nonparametric setting. Second, we do not rely on any numerical approximation and the quantities of interest can be computed efficiently and exactly.

Our method applies to the following experimental setting: In a psychophysical experiment, the causal relation $X \rightarrow Y$ between a physical stimulus X and the psychological response Y of an observer is investigated. Before the experiment, a discrete set of L stimuli $\mathcal{X} = \{x_1, \ldots, x_L\}$ and K possible responses $\mathcal{Y} = \{y_1, \ldots, y_K\}$ is determined. The stimuli is considered to be ordinal, i.e. the set of stimuli is assumed to be linearly ordered, for instance by strength or any other property associated with the physical parameters. The actual experiment is then performed in a sequential manner. That is, at the n-th stage of the experiment, a particular stimulus $X_n \in \mathcal{X}$ is set and the participant's response Y_n to that stimulus is recorded.

Figure 1 outlines such a experiment for illustrative purposes, taken from Elze et al. (2011), Experiment 2. Figure 1A shows the experimental setup that involves a two-alternative-forced-choice (2AFC) discrimination task: An observer has to report which of two possible target stimuli has been presented on a computer monitor. The discrimination performance was impaired by the presentation of a second stimulus, the so-called mask stimulus, at a position close to the target location. Within this experimental setting, the actual stimulus of interest X is the time interval between the offset of the target and the onset of the mask, the so-called interstimulus interval (ISI) that takes values in $\mathcal{X}$ = {20 ms, 40 ms, …, 200 ms}, whereas the response Y takes values in $\mathcal{Y}$ = {correct, incorrect}, depending on whether or not the observer reported the correct target stimulus.

One of the most often considered approaches to model such an experimental situation is to assume a multinomial sampling law. More specifically, given any stimulus $\{X = x\}$ the generation of the response Y is thought to be described by multinomial parameters $\boldsymbol{p}_{x,\mathcal{Y}} = (p_{x,y})_{y \in \mathcal{Y}} \in \Delta_{\mathcal{Y}}$, where

$$\Delta_{\mathcal{Y}} = \left\{ \boldsymbol{p}_{\mathcal{Y}} = (p_y)_{y \in \mathcal{Y}} \in \mathbb{R}^{\mathcal{Y}} \mid p_y \geq 0 \text{ for all } y \in \mathcal{Y} \text{ and } \sum_{y \in \mathcal{Y}} p_y = 1 \right\} \tag{1}$$

is the *probability simplex*, such that the conditional probability of the event $\{Y = y\}$ given $\{X = x\}$ is

$$\mathbf{P}[Y = y | X = x, \boldsymbol{p}_{\mathcal{X},\mathcal{Y}}] = p_{x,y}.$$

Within this setting, the statistical task is then entirely focused on the estimation of the *psychometric rates* $\boldsymbol{p}_{\mathcal{X},\mathcal{Y}} = (\boldsymbol{p}_{x,\mathcal{Y}})_{x \in \mathcal{X}} \in \Delta_{\mathcal{Y}}^{\mathcal{X}}$. A low-dimensional parametric family of

functions $\mathcal{F} = \{f_\theta | \theta \in \Theta\}$ can be used to introduce dependencies among the psychometric rates, such that

$$\mathbf{P}\left[Y=y|X=x,\theta\right]=f_\theta(x,y).$$

This facilitates the inference task by exploiting structural knowledge about the interlink between the stimuli. The function f_θ is commonly termed *psychometric function* in the particular case of a binary response, e.g. a 2AFC experiment as outlined above. Especially sigmoid curves are a common choice if the stimulus X can be considered to be real-valued. Their geometric parameters such as slope and threshold can serve to actually define relevant psychophysical quantities of interest. There exists a vast literature on the statistical inference of psychometric rates and functions, see for instance Wichmann and Hill (2001); Kuss et al. (2005), which provide a good entry point into the literature.

From a mere statistical point of view, this *parametric approach* seems to be a reasonable strategy as it allows *sharing statistical strength* across stimuli. By learning the psychometric rate for a particular stimulus we also learn about all other rates because dependencies are introduced by the parametric family. Hence, the parametric approach allows seemingly good estimates even with few experimental data. Nevertheless, this modeling approach is unsuitable and can even bear the risk of a severe bias if there is no or only vague knowledge about the potential shapes of the functions f_θ and no member of the proposed parametric family $\mathcal{F}$ does match with the actual psychometric rates. For instance, in the above example it seems hard to motivate any plausible regular functional form (see Figure 1B).

This form of bias is of course avoided by allowing the psychometric rates to freely range over $\Delta_{\mathcal{Y}}^{\mathcal{X}}$, which we refer to as the *nonparametric* approach[1]. Clearly, much more data is then needed to draw informative inferences.

In this paper, we use an intermediate approach fairly balancing the advantages and disadvantages of both approaches by exploiting the fact that $\mathcal{X}$ is of ordinal structure. Loosely speaking, we allow neighboring stimuli to share statistical strength by joining their respective psychometric rates. Each possible way of joining neighboring stimuli imposes a particular partition on the stimulus space $\mathcal{X}$, which is then assessed by a suitable Bayesian inference scheme. The resulting model is a variant of the product partition model proposed by Hartigan (1990) and the inhomogeneous Bernoulli process with piecewise constant probabilities described in Endres et al. (2008).

In our description of the model and the respective adaptive sampling procedures we follow the *predictive paradigm* as pioneered, for instance, in Roberts (1965); de Finetti (1974); Geisser (1993), by putting special emphasis on the prediction for the observables, which are the stimulus-response outcomes of the sequential experiment. By imposing a particular epistemic condition of partial exchangeability, the psychometric rates naturally emerge as a particular limiting statistic of the data rather than an external quantity. The corresponding Bayesian model matches extensionally with a multinomial sampling model with unknown parameters.

[1]There is much ambiguity in the usage of the term *nonparametric* as different fields of statistics assign different meanings to what is actually meant by nonparametric. Here, we simply mean that no constraint in form of a parametric family is imposed that restricts the topological support for the psychometric rates $\boldsymbol{p}_{x,y}$ in $\Delta_{\mathcal{X}}^{\mathcal{Y}}$.

2. A Predictive Perspective on Sequential Experiments

2.1. Sequential Construction of Adaptive Sampling Processes

A reasonable adaptive experimental design requires that we have at least partial control of the sampling process concerning the presentation of stimuli. This experimental controllability can be subject to uncertainty, e.g. the experimental setup might be prone to errors in generating the required stimulus. For simplicity, we shall nonetheless assume that we can adjust the stimulus in the way we want. Our personal action policy, which is the subjective assessment of which stimulus might be best to choose, is then expressed by a probability measure $\mathbf{P}[X]$. By setting a stimulus X w.r.t. $\mathbf{P}[X]$, we subsequently observe a respective response Y, where our prediction is described by a conditional measure $\mathbf{P}[Y|X]$.

In the following, we shall extend this scenario to sequential sampling schemes. We consider that at any sampling step $n \in \mathbb{N}$ we can freely choose a stimulus X_n that results in the observation of an instance (X_n, Y_n). Let $E_n = (X_n, Y_n)$ denote the n-th experiment, such that we refer to the first n experiments of the overall *experimental process* $\boldsymbol{E} = (E_n)_{n \in \mathbb{N}}$ by $\boldsymbol{E}_n = (E_1, \ldots, E_n) = (\boldsymbol{X}_n, \boldsymbol{Y}_n)$, where $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ and $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n)$. Two important statistics that summarize the data from the experiments $\boldsymbol{E}_n$ are given by the *total count statistic* $\boldsymbol{n}_{\mathcal{X},\mathcal{Y}} = (\boldsymbol{n}_{x,y})_{x \in \mathcal{X}} = (n_{x,y})_{x \in \mathcal{X}, y \in \mathcal{Y}}$, where $n_{x,y}$ is the (absolute) frequency of the event $\{X = x, Y = y\}$ in $\boldsymbol{E}_n$, and $\boldsymbol{n}_{\mathcal{X}} = (n_x)_{x \in \mathcal{X}}$ is the related *stimulus count statistic*, i.e. $n_x = \sum_{y \in \mathcal{Y}} n_{x,y}$. We illustrate this notation by the following short example.

Example 1—Suppose we run an experiment with a set of stimuli $\mathcal{X} = \{a, b, c, d\}$ and dichotomous responses $\mathcal{Y} = \{0, 1\}$. If in eight trials we observe that

$$\boldsymbol{E}_8 = \begin{pmatrix} \boldsymbol{X}_8 \\ \boldsymbol{Y}_8 \end{pmatrix} = \begin{pmatrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 \\ Y_1 & Y_2 & Y_3 & Y_4 & Y_5 & Y_6 & Y_7 & Y_8 \end{pmatrix} = \begin{pmatrix} a & b & a & c & d & b & b & c \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}, \tag{2}$$

then the two statistics just defined are

$$\boldsymbol{n}_{\mathcal{X},\mathcal{Y}} = \begin{pmatrix} \boldsymbol{n}_{a,\mathcal{Y}} & \boldsymbol{n}_{b,\mathcal{Y}} & \boldsymbol{n}_{c,\mathcal{Y}} & \boldsymbol{n}_{d,\mathcal{Y}} \end{pmatrix} = \begin{pmatrix} n_{a,0} & n_{b,0} & n_{c,0} & n_{d,0} \\ n_{a,1} & n_{b,1} & n_{c,1} & n_{d,1} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 2 & 2 & 0 \end{pmatrix},$$
$$\boldsymbol{n}_{\mathcal{X}} = (n_a, n_b, n_c, n_d) = (2, 3, 2, 1).$$

We now take a sequential and prediction oriented perspective. We fix our expectations about the experimental course $\boldsymbol{E}$ by specifying a sequence of conditional measures in form of kernels $\boldsymbol{\pi}_n(E_{n+1} \mid \boldsymbol{E}_n)$, $n \in \mathbb{N}_0$, where each $\boldsymbol{\pi}_n$ describes our uncertainty about the outcome of the $n + 1$-th experiment given the experimental data from the previous n sampling steps. Each such sequence of kernels $\boldsymbol{\pi}_n(\cdot \mid \cdot)$, $n \in \mathbb{N}_0$, defines a unique measure $\boldsymbol{\pi}$ on the space of experimental courses, such that marginally

$$\pi(\boldsymbol{E}_n) = \prod_{i=0}^{n-1} \pi_i(E_{i+1} | \boldsymbol{E}_i), n \in \mathbb{N}.$$

Thus, we can formally describe the experimental course $\boldsymbol{E}$ as a random process, which we call the *adaptive sampling process* $\boldsymbol{E}$ distributed with respect to $\boldsymbol{\pi}$ (i.e $\boldsymbol{E} \sim \boldsymbol{\pi}$). This process describes our belief dynamic since it is constructed from our *experimental predictions* $(\boldsymbol{\pi}_n)_{n\in\mathbb{N}_0}$. Each experimental prediction $\boldsymbol{\pi}_n$ can in turn be constructed from two separate kernels

$$\pi_n(E_{n+1}|\boldsymbol{E}_n)=\pi_n^X(X_{n+1}|\boldsymbol{E}_n)\pi_n^Y(Y_{n+1}|, X_{n+1}, \boldsymbol{E}_n).$$

There exists a very natural interpretation for each of these kernels in terms of experimental design and prediction:

The Stimulus Placement Rule: $\pi_n^X(X_{n+1}|\boldsymbol{E}_n)$ specifies our *action policy* in the n-th sampling step. It determines which stimulus X_{n+1} we select given the n previous experiments $\boldsymbol{E}_n$.

The Response Prediction Rule: $\pi_n^Y(Y_{n+1}|X_{n+1}, \boldsymbol{E}_n)$ is our prediction of the response knowing the outcome of the last n experiments $\boldsymbol{E}_n$, i.e. which response Y_{n+1} we expect given the actual stimulus X_{n+1}.

Given a particular assignment for the prediction rule π_n^Y, we want to learn the stimulus-response relation $X \rightarrow Y$ in an optimal manner with regard to the inference scheme and external constraints. Thus, we want to derive a placement rule π_n^X that allows us to adapt the experimental course to our objectives.

In order to understand the logic of adaptive sampling strategies it is worthwhile to first consider the case of a non-adaptive design. Such a conventional design consists of a fixed sequence of stimuli $\mathbf{x}^*=(x_n^*)_{n\in\mathbb{N}}$, such that

$$\pi_n^X(X_{n+1}=x|\boldsymbol{E}_n)=\begin{cases} 1 & \text{if } x=x_{n+1}^* \\ 0 & \text{otherwise} \end{cases}.$$

Clearly, such an action policy does not take any information about the already collected data into account. A more reasonable strategy allows the decision for a stimulus x_{n+1}^* to depend on the outcomes of the n foregoing experiments $\boldsymbol{E}_n$, i.e. $x_{n+1}^*=x_{n+1}^*(\boldsymbol{E}_n)$. More precisely, instead of describing one fixed sequence of stimuli we rather specify a *decision rule* that determines a stimulus x_{n+1}^* on the basis of the previous experiments $\boldsymbol{E}_n$. Many of the classical adaptive procedures for testing psychometric functions, such as PEST (Taylor and Creelman (1967)), QUEST (Watson and Pelli (1983)) and the up-down procedures (Levitt (1971)) can be described that way. A comprehensive review of these methods can be found in Leek (2001).

One principled way to obtain a decision rule is to specify a *utility measure* $U_{n+1}(x, \boldsymbol{E}_n)$, which quantifies the utility of a stimulus x based on the outcome of the previous experiments $\boldsymbol{E}_n$. Given such a measure, an *optimal stimulus* is determined by

$$x_{n+1}^*(\boldsymbol{E}_n)=\arg\max_{x\in\mathcal{X}} U_{n+1}(x, \boldsymbol{E}_n). \tag{3}$$

A proper placement rule in the case of multiple optimal stimuli $\mathcal{O}_{n+1}(\boldsymbol{E}_N) \subseteq \mathcal{X}$ is given by

$$\pi_n^X(X_{n+1}|\boldsymbol{E}_n)=\begin{cases}\frac{1}{|\mathscr{O}_{n+1}(\boldsymbol{E}_n)|} & \text{if } X_{n+1}\in\mathscr{O}_{n+1}(\boldsymbol{E}_n)\\ 0 & \text{otherwise}\end{cases}.$$

Hence, by taking on the utility-oriented perspective, the problem of determining a placement rule becomes a problem of choosing suitable utility measures U_{n+1}, $n\in\mathbb{N}_0$. Many possible utility measures exist and which one we choose depends solely on our objectives. For instance, if we want to place the stimuli in a random but balanced manner, then we could choose the utility measure

$$U_{n+1}(x,\boldsymbol{E}_n)=-n_x(\boldsymbol{E}_n),$$

where n_x is the stimulus count statistic of x. Clearly, the respective placement rule selects stimuli that have seen the least trials. This *random uniform sampling* scheme has also been called *method of constant stimuli* (cf. McKee et al. (1985)) within the context of psychophysical experiments. It is usually considered as a non-adaptive strategy (Watson and Fitzhugh (1990)).

Another common and more sensible strategy to obtain a placement rule is suggested by the theory of optimal sequential decisions under uncertainty (cf. DeGroot (2004); Berger (1993)). In principle we should consider that only a finite number of experiments is performed, say $N\in\mathbb{N}$, such that we should formulate our objectives in form of a *global utility* measure $u(\boldsymbol{E}_N)$ for the outcome of the overall experimental course $\boldsymbol{E}_N$. As a matter of rationality, one should choose a sequence of descision rules $\boldsymbol{x}_N^*(\boldsymbol{E}_{N-1})=(x_1^*, x_2^*(\boldsymbol{E}_1),\ldots, x_N^*(\boldsymbol{E}_{N-1}))$, such that the *expected global utility*

$$\bar{u}=\sum_{\boldsymbol{y}_N\in\mathscr{Y}^N} u(\boldsymbol{x}_N^*,\boldsymbol{y}_N)\prod_{n=0}^{N-1}\pi_{n+1}^Y(Y_{n+1}=y_{n+1}|X_{n+1}=x_{n+1}^*,\boldsymbol{X}_n=\boldsymbol{x}_n^*,\boldsymbol{Y}_n=\boldsymbol{y}_n)$$

is maximized. This optimization problem is highly non-trivial, but can be solved, at least in principle, with backward induction (Berger (1993); Bernardo and Smith (1995); DeGroot (2004)). For a concise description of the backward induction method see in particular Müller et al. (2007). This procedure leads to a sequence of *local utility* measures $u_{n+1}(x, y, \boldsymbol{E}_n)$ that are induced from both the global utility u and the predictions $\boldsymbol{\pi}^Y$. The optimal stimulus for the $n+1$-th experiment is determined by maximizing the *expected local utility*, i.e.

$$x_{n+1}^*(\boldsymbol{E}_n)=\arg\max_{x\in\mathscr{X}}\sum_{y\in\mathscr{Y}} u_{n+1}(x,y,\boldsymbol{E}_n)\pi_n^Y(Y_{n+1}=y|X_{n+1}=x,\boldsymbol{E}_n). \quad (4)$$

The local utility measure $u_{n+1}(x, y, \boldsymbol{E}_n)$ for a particular stimulus-response (x, y) is the expected global utility $\bar{u}$ as of the $n+1$-th sampling step given that all subsequent decisions are made in the same optimal manner. Although this scheme leads in principle to an optimal design, in most cases, except for trivial settings, the actual computation turns out to be infeasible.

It is primarily for this reason that various approximations to such an optimal design have been developed. One of them is to resort to a *myopic adaptive optimal design* by directly

specifying local utility measures $u_{n+1}(x, y, \boldsymbol{E}_n)$, $n = 0, \ldots, N-1$, in an attempt to optimize the global utility. Likewise, the optimal stimulus is determined by (4), such that we optimize the next step, but with the hope that this also maximizes the global utility. We shall discuss particular choices for such measures based on information theoretical considerations in section 3.7.

2.2. Partial Exchangeable Response Processes

Concerning the choice of a proper prediction rule π_n^Y, we already mentioned in the introduction that the response generation $X \to Y$ is usually thought to be governed by a multinomial sample law described by some unknown psychometric rates $\boldsymbol{p}_{\mathcal{X},\mathcal{Y}} \in \Delta_{\mathcal{X},\mathcal{Y}}$. From a predictive perspective this amounts to the assumption of a specific form of *partial exchangeability*, which has been introduced in de Finetti (1980) as a generalization of the concept of exchangeability (de Finetti (1937)). Roughly speaking, we have to assume that particular temporal orderings within any finite sequence of experiments $\boldsymbol{E}_n$ do not provide relevant information for our predictions.

In order to make this more precise, we need to introduce the following statistics and respective variables. We define the *response statistic* $\boldsymbol{y}_{n_{\mathcal{X}}}^{\mathcal{X}} = (\boldsymbol{y}_{n_x}^x)_{x \in \mathcal{X}}$, where $\boldsymbol{y}_{n_x}^x = (y_1^x, \ldots, y_{n_x}^x)$. Here $y_i^x = y$ indicates that in the i-th trial in which $\{X = x\}$ occurred the response outcome was the event $\{Y = y\}$. It is crucial to notice that the count statistic $\boldsymbol{n}_{\mathcal{X},\mathcal{Y}}$ can also be computed from the response statistic. The random variables related to the response statistic are denoted by $\boldsymbol{Y}^{\mathcal{X}} = (\boldsymbol{Y}^x)_{x \in \mathcal{X}}$ with $\boldsymbol{Y}^x = (\boldsymbol{Y}_i^x)_{i \in \mathbb{N}}$, such that $Y_i^x = y_i^x$. Since we often need to refer to a particular subset of $\boldsymbol{Y}^{\mathcal{X}}$, we also define $\boldsymbol{Y}_{\boldsymbol{n}_{\mathcal{X}}}^{\mathcal{X}} = (\boldsymbol{Y}_{n_x}^x)_{x \in \mathcal{X}}$, where $\boldsymbol{Y}_{n_x}^x = (Y_1^x, \ldots, Y_{n_x}^x)$, $x \in \mathcal{X}$, i.e. $\boldsymbol{Y}_{\boldsymbol{n}_{\mathcal{X}}}^{\mathcal{X}} = \boldsymbol{y}_{\boldsymbol{n}_{\mathcal{X}}}^{\mathcal{X}}$.

Example 2—In our previous example the response statistic $\boldsymbol{y}_{n_x}^x$ is given by

$$\boldsymbol{y}_2^a = (1, 0), \boldsymbol{y}_3^b = (1, 1, 0), \boldsymbol{y}_2^c = (1, 1), \boldsymbol{y}_1^d = (0),$$

such that we observed that

$$(\boldsymbol{Y}_2^a, \boldsymbol{Y}_3^b, \boldsymbol{Y}_2^c, \boldsymbol{Y}_1^d) = \begin{pmatrix} Y_1^a & Y_1^b & Y_1^c & Y_1^d \\ Y_2^a & Y_2^b & Y_2^c & \\ & Y_3^b & Y_3^c & \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & \\ & 0 & & \end{pmatrix},$$

We now proceed as follows. Instead of specifying the prediction rule π_n^Y directly, we rather fix a probability measure $\mathbf{P}$ for $\boldsymbol{Y}^{\mathcal{X}}$, i.e $\boldsymbol{Y}^{\mathcal{X}} \sim \mathbf{P}$, such that we treat $\boldsymbol{Y}^{\mathcal{X}}$ as a random process called the *response process*. This response process is meant to describe our personal beliefs about the observational part of the experiments irrespective of the actual underlying sequence of stimuli $\boldsymbol{X}_N$, which clearly depends on our action policy. A proper prediction rule is then given by

$$\pi_n^Y(Y_{n+1}=y|X_{n+1}=x, \boldsymbol{E}_n)=\frac{\mathbf{P}\left[\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}, Y^x_{n_x+1}=y\right]}{\mathbf{P}\left[\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}\right]}. \quad (5)$$

Furthermore, and more importantly, the response process $\boldsymbol{Y}^{\mathcal{X}}$ can be utilized to derive an intrinsic statistical model for $X \rightarrow Y$ by requiring the following form of partial exchangeability (cf. Link (1980); de Finetti (1980)):

The response process $\boldsymbol{Y}^{\mathcal{X}}$ is said to be *partial exchangeable* iff

$$\mathbf{P}\left[\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}=\boldsymbol{y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}\right]=\mathbf{P}\left[\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}=\boldsymbol{y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}\right]$$

for every two responses $\boldsymbol{y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}$ and $\tilde{\mathbf{y}}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}$, which share the same count statistic $\boldsymbol{n}_{\mathcal{X},\mathcal{Y}}$. This is equivalent to require that the response process $\boldsymbol{Y}_{\boldsymbol{n}_{\mathcal{X}}}$ is *summarized* by the count statistics $\boldsymbol{n}_{\mathcal{Y}}$ (cf. Lauritzen (1974)), i.e. every sequence $\boldsymbol{y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}$ with the same count statistic is predicted to be equally likely. We illustrate this notion of exchangeability by the following example.

Example 3—Consider we observed the data set in (2), see example 1. The alternative observations

$$(\boldsymbol{Y}^a_2, \boldsymbol{Y}^b_3, \boldsymbol{Y}^c_2, \boldsymbol{Y}^d_1)=\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & \\ & 0 & & \end{pmatrix} \text{ or } \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & \\ & 1 & & \end{pmatrix},$$

do all preserve the count statistics $\boldsymbol{n}_{\mathcal{X},\mathcal{Y}}$, such that we would judge them all to be equally likely under the condition of partial exchangeability. Note that all of these equivalent observations are related by particular permutations, i.e. by exchanging positions within each of the columns $\boldsymbol{Y}^x_{n_x}$, $x \in \mathcal{X}$, but not across them.

An import subclass of partially exchangeable response processes is described by the multinomial sampling laws[2], where

$$\boldsymbol{Y}^x \sim \text{Mult}[\boldsymbol{p}_{\mathscr{X},\mathscr{Y}}],$$

such that the marginal probability mass function is given by

$$f_{\text{Mult}[\boldsymbol{p}_{\mathscr{X},\mathscr{Y}}]}\left[\boldsymbol{y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}\right]=\prod_{x\in\mathscr{X},y\in\mathscr{Y}} p_{x,y}^{n_{x,y}}.$$

[2]The multinomial sampling laws described here apply to the categorical process $\boldsymbol{Y}^{\mathcal{X}}$ and are not to be confused with the related multinomial distribution for the respective count statistic $\boldsymbol{n}_{\mathcal{X},\mathcal{Y}}$.

This class is especially important because for each partial exchangeable random process there exists a mixing measure μ on $\Delta_{\mathcal{Y}}^{\mathcal{X}}$ (Link (1980); Bernardo and Smith (1995)), also called a *de Finetti measure*, such that

$$\mathbf{P}\left[\boldsymbol{Y}_{\boldsymbol{n}_{\mathcal{X}}}^{\mathcal{X}}=\boldsymbol{y}_{\boldsymbol{n}_{\mathcal{X}}}^{\mathcal{X}}\right]=\int_{\Delta_{\mathcal{Y}}^{\mathcal{X}}} f_{\mathrm{Mult}[\boldsymbol{p}_{\mathcal{X},\mathcal{Y}}]}\left[\boldsymbol{y}_{\boldsymbol{n}_{\mathcal{X}}}^{\mathcal{X}}\right]\mu(\mathrm{d}\boldsymbol{p}_{\mathcal{X},\mathcal{Y}}).$$

Therefore each partial exchangeable random process is a mixture of the multinomial sampling laws. Within this context, each $\boldsymbol{p}_{\mathcal{X},\mathcal{Y}} \in \Delta_{\mathcal{Y}}^{\mathcal{X}}$ can be interpreted as a possible limit of the relative frequencies, i.e.

$$\lim_{n_x\to\infty}\frac{\boldsymbol{n}_{x,\mathcal{Y}}}{n_x}=\boldsymbol{p}_{x,\mathcal{Y}},\ x\in\mathcal{X},$$

such that the measure μ expresses our prediction for these limits. Hence, under the condition of partial exchangeability we can formally identify these limits as the psychometric rates we are uncertain about, but with the important distinction that the rates are not external quantities but asymptotic statistics of the response process itself. The corresponding Bayesian model can thus be described by

$$\boldsymbol{Y}^{\mathcal{X}}\sim\mathrm{Mult}[\boldsymbol{p}_{\mathcal{X},\mathcal{Y}}],$$

whereas

$$\boldsymbol{p}_{\mathcal{X},\mathcal{Y}}\sim\mu.$$

By conditioning on finite data $\boldsymbol{Y}_{\boldsymbol{n}_{\mathcal{X}}}^{\mathcal{X}}$, the resulting response process is still partial exchangeable, such that the respective de Finetti measure $\mu(\cdot|\boldsymbol{Y}_{\boldsymbol{n}_{\mathcal{X}}}^{\mathcal{X}})$ describes our posterior belief about the psychometric rates and can be seen to be a Bayesian posterior of $\boldsymbol{p}_{\mathcal{X},\mathcal{Y}}$.

3. Response Processes with Proximally Related Stimuli

Here, we introduce a particularly useful instance of a partial exchangeable response process, namely the *multibin Pólya mixture urn* process, which allows a flexible modeling of similarities between stimuli. We first consider the case of only one stimulus and continue with the easiest non-trivial case of two stimuli before we develop the full process with multiple stimuli. Afterwards we shall consider the construction of respective adaptive sampling procedures.

3.1. One stimulus

Let us consider only one stimulus, i.e. $\mathcal{X} = \{x\}$, with the following probability assignment for the response process $\boldsymbol{Y}^{x}$: We say that $\boldsymbol{Y}^{x}$ is a *Pólya urn process* with parameters $\boldsymbol{\alpha}_{x,\mathcal{Y}} = (\alpha_{x,y})_{y\in\mathcal{Y}}$, where $\alpha_{x,y} > 0$, if

$$\mathbf{P}[\boldsymbol{Y}_{n_x}^{x}] = \frac{\mathrm{Beta}(\boldsymbol{n}_{x,\mathscr{Y}} + \alpha_{x,\mathscr{Y}})}{\mathrm{Beta}(\alpha_{\mathscr{Y}})},$$

where Beta is the multinomial beta function. The corresponding prediction rule is

$$\pi_n^Y(Y_{n+1} = y | X_{n+1} = x, \boldsymbol{X}_n, \boldsymbol{Y}_n) = \frac{n_{x,y} + \alpha_{x,y}}{n_x + \alpha_x},$$

where $\alpha_X = \Sigma_{y \in \mathscr{Y}} \alpha_{X,y}$, which is commonly known as the (generalized) Bayes-Laplace rule. The parameters $\boldsymbol{\alpha}_{X,\mathscr{Y}}$ have a natural interpretation as *pseudo counts* added to the actual counts $\boldsymbol{n}_{X,\mathscr{Y}}$. The term Pólya urn stems from the original introduction as an urn model in Eggenberger and Pólya (1923). In this picture, balls are successively drawn from an urn that is initially filled with K balls of different colors. Each color indicates a particular $y \in \mathscr{Y}$ and the respective ball is assigned an initial weight of $\alpha_{X,y}$. At each sampling step, a ball is thought to be drawn with chance given by its individual mass and put back to the urn with another ball of the same color and mass one. The so generated sequence of colors $\boldsymbol{Y}^X$ is then described by the above Pólya urn process, which is partial exchangeable.

The respective de Finetti measure is given by a Dirichlet distribution with parameters $\boldsymbol{\alpha}_{X,\mathscr{Y}}$, i.e.

$$\boldsymbol{p}_{x,\mathscr{Y}} \sim \mathrm{Dir}\left[\alpha_{x,\mathscr{Y}}\right],$$

where the respective density is

$$f(\boldsymbol{p}_{x,\mathscr{Y}}) = \frac{1}{\mathrm{Beta}(\alpha_{x,\mathscr{Y}})} \prod_{y \in \mathscr{Y}} p_y^{\alpha_{x,y}-1}.$$

Conditional on finite data $Y_{n_x}^x$, the resulting response process is again a Pólya urn process with parameters $\alpha'_{x,\mathscr{Y}}$ given by the update rule

$$\alpha'_{x,y} = \alpha_{x,y} + n_{x,y}, y \in \mathscr{Y},$$

i.e. the observed counts $\boldsymbol{n}_{X,\mathscr{Y}}$ are just added to the pseudo counts $\boldsymbol{\alpha}_{X,\mathscr{Y}}$. Hence, assuming a Pólya urn process for describing our personal predictions justifies the formal adoption of the standard Bayesian model of multinomial sampling and a subjective prior in form of the Dirichlet distribution.

3.2. Two stimuli

We now consider a stimulus space with two elements, say $\mathscr{X} = \{x_1, x_2\}$. Based on the previously described Pólya urn process we can think of two extremal urn models for the response process $\boldsymbol{Y}^{\mathscr{X}}$. First, we consider the case where the underlying mechanisms of ($X =$

$x_1) \rightarrow Y$ and $(X = x_2) \rightarrow Y$ are identical. More precisely, we say that x_1 and x_2 are similar, denoted $x_1 \sim x_2$, if we expect that there is no difference in the generation of Y given either x_1 or x_2. This similarity allows us to define a *bin* $b = \{x_1, x_2\}$, such that we can exploit the similarity by joining the observations from both stimuli[3]. We call this binning scheme B_1 and conditional on it the response process $\boldsymbol{Y}^{\mathcal{X}}$ is characterized by

$$\mathbf{P}\left[\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}} | B_1\right] = \frac{\mathrm{Beta}(\boldsymbol{n}_{b,\mathcal{Y}} + \alpha_{b,\mathcal{Y}})}{\mathrm{Beta}(\alpha_{b,\mathcal{Y}})},$$

where $\boldsymbol{n}_{b,\mathcal{Y}} = (n_{b,y})_{y \in \mathcal{Y}}$ is the total count $\boldsymbol{n}_{\mathcal{X},\mathcal{Y}}$ of observations joined in bin b, i.e. $n_{b,y} = n_{x_1,y} + n_{x_2,y}$, $y \in \mathcal{Y}$, and $\boldsymbol{\alpha}_{b,\mathcal{Y}}$ describes the pseudo counts. The induced prediction rule is

$$\pi_n^Y(Y_{n+1} = y | X_{n+1} = x, \boldsymbol{E}_n, B_1) = \frac{n_{b,y} + \alpha_{b,y}}{n_b + \alpha_b},$$

where $x \in b$. The so described response process is partial exchangeable and the respective de Finetti measure $\mu(\cdot | B_1)$ can informally be described by the density

$$f(\boldsymbol{p}_{\mathcal{X},\mathcal{Y}}) = \frac{1}{\mathrm{Beta}(\alpha_{b,\mathcal{Y}})} \prod_{y \in \mathcal{Y}} p_{b,y}^{\alpha_{b,y}-1} \prod_{x \in b} \delta(\boldsymbol{p}_{b,\mathcal{Y}} - \boldsymbol{p}_{x,\mathcal{Y}}),$$

where δ is the Dirac delta function. That is, the measure $\mu(\cdot | B_1)$ concentrates on the *diagonal* of the product simplex $\Delta_{\mathcal{Y}}^{\mathcal{X}} = \Delta_{\mathcal{Y}}^{x_1} \times \Delta_{\mathcal{Y}}^{x_2}$, which is the set

$$\{\boldsymbol{p}_{\mathcal{X},\mathcal{Y}} \in \Delta_{\mathcal{Y}}^{\mathcal{X}} | p_{x_1,y} = p_{x_2,y}, y \in \mathcal{Y}\},$$

such that if we identify both psychometric rates $\boldsymbol{p}_{x_1,\mathcal{Y}}$ and $\boldsymbol{p}_{x_2,\mathcal{Y}}$ by just one rate $\boldsymbol{p}_{b,\mathcal{Y}} \in \Delta\mathcal{Y}$, then this rate is described by a Dirichlet distribution with parameters $\boldsymbol{\alpha}_{b,\mathcal{Y}}$. Similar to the case of a single stimulus, the update rule for the parameters is given by

$$\alpha'_{b,y} = \alpha_{b,y} + n_{b,y}, y \in \mathcal{Y}.$$

The second case we have to consider is that both mechanism $(X = x_1) \rightarrow Y$ and $(X = x_2) \rightarrow Y$ are independent. In order to describe a respective response process, consider two independent Pólya urns both filled with the same kind of colored balls, where we mark the two urns x_1 respectively x_2. In each sampling step we first decide from which urn to sample and then proceed as before. To be consistent with our notation we introduce a binning scheme B_2 with two separate bins $b_1 = \{x_1\}$ and $b_2 = \{x_2\}$. Instead of assigning weights to the urns directly, we assign them to our bins, i.e. $\boldsymbol{\alpha}_{b_1,\mathcal{Y}}$ for the first bin and $\boldsymbol{\alpha}_{b_2,\mathcal{Y}}$ for the second one. The respective response is given by the product measure

[3]The illustrative metaphor of a *bin* is taken from Endres and Földiák (2005); Endres et al. (2008), whereas cluster, block or component might serve equally well.

$$\mathbf{P}\left[\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}|B_2\right]=\frac{\mathrm{Beta}(\boldsymbol{n}_{b_1,\mathscr{Y}}+\alpha_{b_1,\mathscr{Y}})}{\mathrm{Beta}(\alpha_{b_1,\mathscr{Y}})}\frac{\mathrm{Beta}(\boldsymbol{n}_{b_2,\mathscr{Y}}+\alpha_{b_2,\mathscr{Y}})}{\mathrm{Beta}(\alpha_{b_2,\mathscr{Y}})}.$$

The so induced prediction rule is

$$\pi^Y_n(Y_{n+1}=y|X_{n+1}=x,\boldsymbol{E}_n,B_2)=\frac{n_{I_{B_2}(x),y}+\alpha_{I_{B_2}(x),y}}{n_{I_{B_2}(x)}+\alpha_{I_{B_2}(x)}},\tag{6}$$

where $x\in\mathscr{X}$ and $I_{B_2}(x)$ tells us the bin to which x is assigned. This response process is partially exchangeable and the respective de Finetti measure $\mu(\cdot|B_2)$ is a product measure on $\Delta^{\mathscr{X}}_{\mathscr{Y}}$, where $\boldsymbol{p}_{x_1,\mathscr{Y}}\sim\mathrm{Dir}[\boldsymbol{\alpha}_{b_1,\mathscr{Y}}]$ and $\boldsymbol{p}_{x_2,\mathscr{Y}}\sim\mathrm{Dir}[\boldsymbol{\alpha}_{b_2,\mathscr{Y}}]$ are independently distributed, i.e. the respective density is given by the product density

$$f(\boldsymbol{p}_{\mathscr{X},\mathscr{Y}})=\frac{1}{\mathrm{Beta}(\alpha_{b_1,\mathscr{Y}})}\prod_{y\in\mathscr{Y}}p_{x_1,y}^{\alpha_{b_1,y}-1}\times\frac{1}{\mathrm{Beta}(\alpha_{b_2,\mathscr{Y}})}\prod_{y\in\mathscr{Y}}p_{x_2,y}^{\alpha_{b_2,y}-1}.$$

Conditional on finite data $\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}},\mathscr{Y}}$, the respectively updated parameters $\alpha'_{B,\mathscr{Y}}$ are

$$\alpha'_{b_i,y}=\alpha_{b_i,y}+n_{b_i,y},\ y\in\mathscr{Y},\ i\in\{1,2\}.$$

We can utilize both schemes B_1 and B_2 to model our beliefs about the similarity and dissimilarity between the stimuli. If B_1 represents the hypothesis that $x_1\sim x_2$, then B_2 is the alternative hypothesis that $x_1\nsim x_2$. If we assess our a priori belief in B_1 with a probability of $\mathbf{P}[B_1]$, then our a priori belief for model B_2 is $\mathbf{P}[B_2]=1-\mathbf{P}[B_1]$. Our overall prediction for the response process is then given by the mixture measure

$$\mathbf{P}\left[\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}\right]=\mathbf{P}\left[\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}|B_1\right]\mathbf{P}[B_1]+\mathbf{P}\left[\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}|B_2\right]\mathbf{P}[B_2].\tag{7}$$

The induced prediction rule is

$$\pi^Y_n(Y_{n+1}=y|X_{n+1}=x,\boldsymbol{E}_n)=\frac{\mathbf{P}\left[\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}},Y^x_{n_x+1}=y\right]}{\mathbf{P}\left[\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}\right]},$$

which can be rewritten as

$$\pi^Y_n(Y_{n+1}=y|X_{n+1}=x,\boldsymbol{E}_n)=\sum_{i=1}^{2}\pi^Y_n(Y_{n+1}=y|X_{n+1}=x,\boldsymbol{E}_n,B_i)\mathbf{P}\left[B_i|\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}\right],\tag{8}$$

where

$$\mathbf{P}\left[B_i|\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}\right]=\frac{\mathbf{P}\left[\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}|B_i\right]}{\mathbf{P}\left[\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}\right]}\mathbf{P}[B_i], i\in\{1,2\}. \tag{9}$$

The latter probability measure $\mathbf{P}\left[B_i|\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}\right]$ can be interpreted as the Bayesian posterior assessment for each binning scheme B_i. The respective de Finetti measure for $\boldsymbol{p}_{\mathcal{X},\mathcal{Y}}$ is given by

$$\mu(\cdot)=\mathbf{P}[B_1]\mu(\cdot|B_1)+\mathbf{P}[B_2]\mu(\cdot|B_2),$$

which can be seen to be a model average over B_1 and B_2. One of our major goals is to generalize this mixture to multiple stimuli and responses, for which we want to develop adaptive sequential sampling strategies.

3.3. Multiple stimuli

Here, we briefly discuss the *multibin Pólya process* on discrete finite stimulus spaces $\mathcal{X}=\{x_1,\ldots,x_L\}$, where we introduce an equivalence relation $\sim_B$ that describes similarities in $\mathcal{X}$. This relation induces a partition B of $\mathcal{X}$ into $|B|$ bins, called a *multibin*. We then bin the data with respect to the resulting multibin $B=\{b_1,\ldots,b_{|B|}\}$, such that we get the binned count statistic $\boldsymbol{n}_{B,\mathcal{Y}}=(\boldsymbol{n}_{b,\mathcal{Y}})_{b\in B}$. In full analogy to the scenario of two stimuli, we set pseudo counts $\boldsymbol{\alpha}_{B,\mathcal{Y}}=(\boldsymbol{\alpha}_{b,\mathcal{Y}})_{b\in B}$ and fix the multibin Pólya urn process $\boldsymbol{Y}^{\mathcal{X}}$ with

$$\mathbf{P}\left[\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}|B\right]=\prod_{b\in B}\frac{\mathrm{Beta}(\boldsymbol{n}_{b,\mathcal{Y}}+\alpha_{b,\mathcal{Y}})}{\mathrm{Beta}(\alpha_{b,\mathcal{Y}})}.$$

The induced prediction rule is

$$\pi^{Y}_{n}(Y_{n+1}=y|X_{n+1}=x,\boldsymbol{E}_n,B)=\frac{n_{I_B(x),y}+\alpha_{I_B(x),y}}{n_{I_B(x)}+\alpha_{I_B(x)}},$$

where $I_B(x)$ denotes the bin of stimulus x given the multibin B, i.e. if $x\in b$ then $I_B(x)=b$. Likewise the above case of two separate stimuli, the respective de Finetti measure $\mu(\cdot|B)$ can be informally described by the density

$$f(\boldsymbol{p}_{\mathcal{X},\mathcal{Y}})=\prod_{b\in B}\frac{1}{\mathrm{Beta}(\alpha_{b,\mathcal{Y}})}\prod_{y\in\mathcal{Y}}p_{b,y}^{\alpha_{b,y}-1}\prod_{x\in b}\delta(\boldsymbol{p}_{b,\mathcal{Y}}-\boldsymbol{p}_{x,\mathcal{Y}}),$$

such that conditional on some finite data $\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}$ the relevant parameters $\boldsymbol{\alpha}_{B,\mathcal{Y}}$ are simply updated according to

$$\alpha'_{b,y}=\alpha_{b,y}+n_{b,y}, b \in B, y \in \mathcal{Y}.$$

3.4. A Hierachical Bayesian Model for Proximally Related Stimuli

We assumed that the stimulus space $\mathcal{X} = \{x_1, x_2, \ldots, x_L\}$ exhibits some well-ordering, as indicated by the indexing. We need to introduce some suitable terminology. Let $\mathcal{C}(\mathcal{X}) = \{b_{i,j} = \{x_i, x_{i+1} \ldots, x_j\} \subseteq \mathcal{X} \mid i \leq j\}$ denote the class of *consecutive bins*, where we call each partition B of $\mathcal{X}$ a *proximal multibin* if it consists only of consecutive bins. The class of all proximal multibins with m bins is denoted $\mathcal{P}_m(\mathcal{X})$, such that $\mathcal{P}(\mathcal{X})=\cup_{m=1}^{L}\mathcal{P}_m(\mathcal{X})$ constitutes the class of all proximal multibins. These definitions are illustrated by the following example.

Example 4—Consider a set of stimuli $\mathcal{X} = \{1, 2, 3\}$, such that

$$\begin{aligned}\mathcal{C}(\mathcal{X})&=\{b_{1,1}, b_{2,2}, b_{3,3}, b_{1,2}, b_{2,3}, b_{1,3}\}\\&=\{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}\}.\end{aligned}$$

Furthermore, $\mathcal{P}(\mathcal{X}) = \{B_1, B_2, B_3, B_4\}$, where

$$\begin{aligned}B_1&=\{b_{1,1}, b_{2,2}, b_{3,3}\}, \quad B_2=\{b_{1,2}, b_{3,3}\},\\B_3&=\{b_{1,1}, b_{2,3}\}, \quad B_4=\{b_{1,3}\}.\end{aligned}$$

For each consecutive bin $b \in \mathcal{C}$ we choose pseudo counts $\boldsymbol{\alpha}_{b,y}$ and construct the following response process: By fixing a priori beliefs $\mathbf{P}[B]$, $B \in \mathcal{P}(\mathcal{X})$, we can describe the full response process $\boldsymbol{Y}^{\mathcal{X}}$ as the mixture

$$\mathbf{P}\left[\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}\right]=\sum_{B\in\mathcal{P}(\mathcal{X})}\mathbf{P}[B]\mathbf{P}\left[\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}|B\right]=\sum_{B\in\mathcal{P}(\mathcal{X})}\mathbf{P}[B]\prod_{b\in B}\frac{\mathrm{Beta}(\boldsymbol{n}_{b,\mathcal{Y}}+\alpha_{b,\mathcal{Y}})}{\mathrm{Beta}(\alpha_{b,\mathcal{Y}})}. \quad (10)$$

We shall refer to this process as the *multibin Pólya mixture process*. The induced prediction rule is

$$\pi_n^Y(Y_{n+1}=y|X_{n+1}=x, \boldsymbol{E}_n)=\frac{\mathbf{P}\left[\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}, Y^x_{n_x+1}=y\right]}{\mathbf{P}\left[\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}\right]}, \quad (11)$$

which can be rewritten as

$$\pi_n^Y(Y_{n+1}=y|X_{n+1}=x, \boldsymbol{E}_n)=\sum_{B\in\mathcal{P}(\mathcal{X})}\pi_n^Y(Y_{n+1}=y|X_{n+1}=x, \boldsymbol{E}_n, B)\mathbf{P}\left[B|\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}\right], \quad (12)$$

where

$$\mathbf{P}\left[B|\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}\right]=\frac{\mathbf{P}\left[\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}|B\right]}{\mathbf{P}\left[\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}\right]}\mathbf{P}[B], B\in\mathscr{P}(\mathscr{X}), \tag{13}$$

is the posterior for the multibins $B \in \mathcal{P}(\mathcal{X})$. The Bayesian model that corresponds to the multibin Pólya mixture process is described by the hierarchical model

$$\begin{aligned} B &\sim \mathbf{P}[\cdot] \\ \boldsymbol{p}_{\mathscr{X},\mathscr{Y}} &\sim \mu(\cdot|B) \\ \boldsymbol{Y}^{\mathscr{X}} &\sim \text{Mult}[\boldsymbol{p}_{\mathscr{X},\mathscr{Y}}]. \end{aligned} \tag{14}$$

There are many ways how to look at the so described hierarchical model. For instance, the model can be seen as a Bayesian regression model for the psychometric rates, where a non-trivial prior in form of

$$\mu(\cdot)=\sum_{B\in\mathscr{P}(\mathscr{X})}\mu(\cdot|B)\mathbf{P}[B]$$

is chosen. This prior assigns probability mass on particular diagonals of the product simplex, such that the psychometric rates become piecewise constant w.r.t. to a particular multibin, which are in turn assumed to be random quantities. From that point of view, the described model can be seen to be a generalization of the inhomogenous Bernoulli process described in Endres et al. (2008). Likewise, the model can be interpreted as a particular clustering model. Observations are clustered with respect to an unknown partition of the data space. From this point of view, the model is related to the so called *product partition model* as introduced by Hartigan (1990), which requires a particular prior structure for $\mathbf{P}[B]$ that will be discussed next.

3.5. Efficient Model Evaluation

The actual computation of the sum-product $\Sigma_{B\in\mathcal{P}(\mathcal{X})}\Pi_{b\in B}$ in equation (10) can become computationally very demanding and infeasible as it may take up to $\mathcal{O}(2^{L-1})$ steps. This is especially problematic for adaptive sampling where the relevant quantities have to be computed as quickly as possible. However, based on the computational approaches taken by Yao (1984); Barry and Hartigan (1992); Endres and Földiák (2005); Fernhead (2006); Hutter (2007) it can be shown that particular prior structures of $\mathbf{P}[B]$ lead to a drastic reduction of the computational effort. In fact, it can be reduced to $\mathcal{O}(L^3)$ if

$$\mathbf{P}[B]=\frac{1}{c(\beta,\gamma)}\beta_{|B|}\prod_{b\in B}\gamma_b, \tag{15}$$

where $|B|$ is the number of bins in B. The respective computational algorithm is given in lemma 1 (see appendix) to which we refer as ***Pro**ximal **M**ulti-**B**in **S**ummation* (ProMBS). It is an abstracted and generalized version of the algorithm presented in Endres and Földiák (2005). The parameters $\boldsymbol{\gamma}=(\gamma_b)_{b\in\mathcal{C}(\mathcal{X})}$, with $\gamma_b \geq 0$, assess the a priori importance of each consecutive bin $b \in \mathcal{C}(\mathcal{X})$ and have been also called *cohesions* in Barry and Hartigan (1992), whereas $\boldsymbol{\beta}=(\beta_1, \ldots, \beta_L)$, with $\beta_l \geq 0$, determine a relative weight for each class $\mathcal{P}_m(\mathcal{X})$, $m = 1, \ldots, L$, in $\mathcal{P}(\mathcal{X})$.

Given a set of parameters $(\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma})$ a particular multibin Pólya mixture process is fixed that describes our a priori belief about the response process $\boldsymbol{Y}^{\mathcal{X}}$. All relevant posterior quantities are determined by the updated parameters

$$\begin{aligned}\alpha'_{b,y} &= \alpha_{b,y} + n_{b,y} \\ \beta'_m &= \frac{\beta_m}{d(\alpha, \beta, \gamma, \boldsymbol{n}_{\mathcal{X},\mathcal{Y}})} \\ \gamma'_b &= \gamma_b \frac{\mathrm{Beta}(\boldsymbol{n}_{b,\mathcal{Y}} + \alpha_{b,\mathcal{Y}})}{\mathrm{Beta}(\alpha_{b,\mathcal{Y}})}\end{aligned}$$

where

$$d(\alpha, \beta, \gamma, \boldsymbol{n}_{\mathcal{X},\mathcal{Y}}) := \sum_{B \in \mathcal{P}(\mathcal{X})} \beta_{|B|} \prod_{b \in B} \gamma_b \frac{\mathrm{Beta}(\boldsymbol{n}_{b,\mathcal{Y}} + \alpha_{b,\mathcal{Y}})}{\mathrm{Beta}(\alpha_{b,\mathcal{Y}})}.$$

The ProMBS algorithm can be used to efficiently compute $d(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{n}_{x,y})$, which will reappear in many other relevant expressions. For instance, we can rewrite equation (10) as

$$\mathbf{P}\left[\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}\right] = \frac{d(\alpha, \beta, \gamma, \boldsymbol{n}_{\mathcal{X},\mathcal{Y}})}{c(\beta, \gamma)},$$

whereas the normalization constant in (15) is given by

$$c(\beta, \gamma) = d(\alpha, \beta, \gamma, 0),$$

such that the prediction rule in equation (11) becomes

$$\pi^{Y}_{n}(Y_{n+1} = y | X_{n+1} = x, \boldsymbol{E}_n) = \frac{d(\alpha, \beta, \gamma, \boldsymbol{n}^{+(x,y)}_{\mathcal{X},\mathcal{Y}})}{d(\alpha, \beta, \gamma, \boldsymbol{n}_{\mathcal{X},\mathcal{Y}})},$$

where $\boldsymbol{n}^{+(x,y)}_{\mathcal{X},\mathcal{Y}}$ is the count statistic $\boldsymbol{n}_{x,y}$ incremented by one count for the event $\{X = x,\ Y = y\}$. Likewise, given a stimulus $x \in \mathcal{X}$, the marginal posterior density of the respective psychometric rate $\boldsymbol{p}_{x,\mathcal{Y}}$ is given by

$$f(\boldsymbol{p}_{x,\mathcal{Y}} | \boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}) = \frac{d(\alpha, \beta, \tilde{\gamma}(\boldsymbol{p}_{x,\mathcal{Y}}), \boldsymbol{n}_{\mathcal{X},\mathcal{Y}})}{d(\alpha, \beta, \gamma, \boldsymbol{n}_{\mathcal{X},\mathcal{Y}})},$$

where

$$\tilde{\gamma}_b(\boldsymbol{p}_{x,\mathscr{Y}}) = \begin{cases} \gamma_b \dfrac{\prod_{y\in\mathscr{Y}} p_{x,y}^{n_{b,y}+\alpha_{b,y}-1}}{\mathrm{Beta}(\boldsymbol{n}_{b,\mathscr{Y}}+\alpha_{b,\mathscr{Y}})} & \text{if } x \in b \\ \gamma_b & \text{otherwise} \end{cases}.$$

The pointwise evaluation of this density can get computationally very expensive, but is nevertheless possible without Monte Carlo methods. Alternatively, the density $f(\boldsymbol{p}_{x,\mathscr{Y}}|\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}})$ can be described by its moments. Here, the k-th raw moment is

$$\mathbb{E}\left[(p_{x,y})^k|\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}\right] = \frac{d(\alpha,\beta,\gamma,\boldsymbol{n}^{+k(x,y)}_{\mathscr{X},\mathscr{Y}})}{d(\alpha,\beta,\gamma,\boldsymbol{n}_{\mathscr{X},\mathscr{Y}})},$$

where we add k events $\{X = x,\ Y = y\}$ to the count statistic $\boldsymbol{n}_{x,y}$. We can also compute the posterior for the multibins (13)

$$\mathbf{P}\left[B|\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}\right] = \frac{\beta_{|B|}\prod_{b\in B}\gamma_b \frac{\mathrm{Beta}(\boldsymbol{n}_{b,\mathscr{Y}}+\alpha_{b,\mathscr{Y}})}{\mathrm{Beta}(\alpha_{b,\mathscr{Y}})}}{d(\alpha,\beta,\gamma,\boldsymbol{n}_{\mathscr{X},\mathscr{Y}})},$$

and by introducing a variable $M \in \{1, \ldots, L\}$ that restricts the model to multibins from $\mathcal{P}_m(\mathcal{X})$, we can compute

$$\mathbf{P}\left[M=m|\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}\right] = \frac{\sum_{B\in\mathscr{P}_m(\mathscr{X})}\beta_m\prod_{b\in B}\gamma_b \frac{\mathrm{Beta}(\boldsymbol{n}_{b,\mathscr{Y}}+\alpha_{b,\mathscr{Y}})}{\mathrm{Beta}(\alpha_{b,\mathscr{Y}})}}{d(\alpha,\beta,\gamma,\boldsymbol{n}_{\mathscr{X},\mathscr{Y}})}$$

which is the posterior assessment that $X \rightarrow Y$ is described by a multibin model with m bins. Since neighboring stimuli potentially share strength, it is of interest to quantify to which extent this is happening. Such an informative statistic is the *effective count* $\bar{n}_x$, which we define as the expectation value

$$\bar{n}_x(\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}) = \sum_{B\in\mathscr{P}(\mathscr{X})} n_{I_B(x)}\mathbf{P}\left[B|\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}\right], \tag{16}$$

where $I_B(x) = b$ if $x \in b$ and $b \in B$. An efficient evaluation is possible because

$$\bar{n}_x(\boldsymbol{Y}^{\mathscr{X}}_{\boldsymbol{n}_{\mathscr{X}}}) = \frac{d(\alpha,\beta,\tilde{\gamma},\boldsymbol{n}_{\mathscr{X},\mathscr{Y}})}{d(\alpha,\beta,\gamma,\boldsymbol{n}_{\mathscr{X},\mathscr{Y}})},$$

where

$$\tilde{\gamma}_b = \begin{cases} \gamma_b n_b & \text{if } x \in b \\ \gamma_b & \text{otherwise} \end{cases} .$$

3.6. Prior Selection

For the multibin Pólya mixture process we need to select parameters ($\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$). With β_m we specify the importance of models consisting of exactly m bins. For instance, if we want to fall back to a simpler problem with at most $l < L$ bins we can set $\beta_l = 1$ and $\beta_j = 0$ for $j \neq l$. A reasonable choice is often given by

$$\beta_m = \left(\begin{matrix} L-1 \\ m-1 \end{matrix} \right)^{-1}, \text{ for } m = 1, 2, \ldots, L.$$

The binomial coefficient gives the number of multibin models that constist of m bins. Hence, if for some constant c all $\boldsymbol{\gamma}_b = c$, $b \in \mathcal{C}(\mathcal{X})$, then the above prior assigns a uniform distribution to $\mathbf{P}[M = m]$, the a priori probability of multibins with m bins. In any case, there are only L values that we have to specify for $\boldsymbol{\beta}$. However, for $\boldsymbol{\alpha}_{b,\mathcal{Y}}$ and $\boldsymbol{\gamma}_b$ we need to select values for each $b \in \mathcal{C}(\mathcal{X})$. If we are dealing with many stimuli the assignment of parameters can get a very extensive task. A first and very convenient choice is to set all $\boldsymbol{\alpha}_{b,\mathcal{Y}}$ and $\boldsymbol{\gamma}_b$ to one, which means that we use an uninformative prior for the pseudo counts and we do not have any preference for specific bins. A more elaborate and natural way is to compute $\boldsymbol{\alpha}_{b,\mathcal{Y}}$ in a hierarchical manner, i.e. if we require that

$$\alpha_{b,\mathcal{Y}} = \frac{1}{|b|} \sum_{x \in b} \alpha_{x,\mathcal{Y}},$$

then we only need to specify $\boldsymbol{\alpha}_{x,\mathcal{Y}}$ for each $x \in \mathcal{X}$.

3.7. Placement Rules for Adaptive Sampling

We discussed in section 2.1 a utility based approach concerning the choice of a suitable action policy for choosing the placement rule π_n^X. We shall assume that our objective is to become most informed about the stimulus-response relation modeled by the multibin Pólya mixture process. We present here several proposals for how this might be achieved.

As already mention earlier, a most trivial sampling scheme is to distribute measurements uniformly, where the respective utility is

$$U_{n+1}(x, \boldsymbol{E}_n) = -n_x, \tag{17}$$

which guarantees at least some homogeneity in the data, but does not take any properties of the underlying model into account. A more sensible utility is given by

$$U_{n+1}(x, \boldsymbol{E}_n) = -\overline{n}_x,$$

where $\bar{n}_X$ is the effective count. The resulting placement rule essentially follows the logic of random uniform sampling, but takes the sharing of strength between neighboring stimuli into account. The attractive feature of this adaptive scheme is of course its conceptual and computational simplicity.

More elaborate adaptive strategies can be obtained by considering utilities based on information-theoretical quantities. The idea of using information theoretic utility measures for experimental designs was probably first considered by Cronbach (1953). A more detailed study of their application in experimental designs was later given by Lindley (1956), whereas explications of Lindley's ideas within the context of Bayesian experimental design can be found in Bernardo (1979); Chaloner and Verdinelli (1995), but see also the application for optimizing sequential experimental designs in DeGroot (1962).

Consider that we want to learn as much as possible about which multibin model $B \in \mathcal{P}(\mathcal{X})$ describes the data best. Our a priori expectation is given by the probability $\mathbf{P}[B]$, whereas our a posteriori assessment is expressed by $\mathbf{P}\left[B|\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}=\boldsymbol{y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}\right]$. The *information* we would gain about B if we learn that $\{\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}=\boldsymbol{y}_{\boldsymbol{n}_{\mathcal{X}}}\}$ is then quantified by the Kullback-Leiber divergence[4]

$$\mathrm{D}_{\mathrm{KL}}\left(\mathbf{P}\left[B|\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}=\boldsymbol{y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}\right]\|\,\mathbf{P}[B]\right)=-\sum_{B\in\mathcal{P}(\mathcal{X})}\mathbf{P}\left[B|\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}=\boldsymbol{y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}\right]\log\left(\frac{\mathbf{P}[B]}{\mathbf{P}\left[B|\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}=\boldsymbol{y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}\right]}\right),$$

which measures the deviation between both measures. Hence, if we want to learn as much as possible about B then it seems reasonable to adopt the Kullback-Leibler divergence as a global utility measure. This, however, leads to almost intractable computations, as mentioned already in section 2.1. Nonetheless, it motivates the following myopic adaptive sampling strategy with local utility measure

$$u^{\mathrm{MB}}_{n+1}(x,y,\boldsymbol{E}_n)=\mathrm{D}_{\mathrm{KL}}(\mathbf{P}\left[B|\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}},Y^{x}_{n_x+1}=y\right]\|\,\mathbf{P}\left[B|\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}\right]),$$

i.e. we try to improve the incremental information in every sampling step. This divergence can be computed with

$$u^{\mathrm{MB}}_{n+1}(x,y,\boldsymbol{E}_n)=\frac{d(\alpha,\beta,\tilde{\gamma},\boldsymbol{n}_{\mathcal{X},\mathcal{Y}})}{d(\alpha,\beta,\gamma,\boldsymbol{n}_{\mathcal{X},\mathcal{Y}})},$$

where

[4]The Kullback-Leibler divergence is formally defined as follows: If μ, ν are two measures on a measurable space $[S,\mathcal{S}]$, such that ν is absolute continuous w.r.t. μ, then $\mathrm{D}_{\mathrm{KL}}(\mu\|\nu)=-\int_S \ln\left(\frac{d\nu}{d\mu}(s)\right)\mu(\mathrm{d}s)$ is the Kullback-Leibler divergence from μ to ν, where $\frac{d\nu}{d\mu}$ is the respective Radon-Nykodým derivate of ν w.r.t. μ.

$$\tilde{\gamma}_b = \begin{cases} \gamma_b \ln\left(\frac{n_{b,y}+\alpha_{b,y}}{n_b+\alpha_b}\frac{d(\alpha,\beta,\gamma,\boldsymbol{n}_{\mathscr{X},\mathscr{Y}})}{d(\alpha,\beta,\gamma,\boldsymbol{n}^{+(x,y)}_{\mathscr{X},\mathscr{Y}})}\right) & \text{if } x \in b \\ \gamma_b, & \text{otherwise .} \end{cases}$$

The respective expected local utility is

$$U^{\mathrm{MB}}_{n+1}(x, \boldsymbol{E}_n) = \sum_{y\in\mathscr{Y}} u^{\mathrm{MB}}_{n+1}(x, y, \boldsymbol{E}_n)\pi^{Y}_{n+1}(Y_{n+1}=y|X_{n+1}=x, \boldsymbol{E}_n),$$

such that the optimal stimulus is determined by

$$x^{*}_{n+1}(\boldsymbol{E}_n) = \arg\max_{x\in\mathscr{X}} U^{\mathrm{MB}}_{n+1}(x, \boldsymbol{E}_n).$$

It is noteworthy that the expected gain can be rewritten in terms of a mutual information[5], namely

$$U^{\mathbf{MB}}_{n+1}(x, \boldsymbol{E}_n) = \mathrm{MI}(B{:}Y^{x}_{n_x+1}|Y^{\mathscr{X}}_{n_{\mathscr{X}}}),$$

such that by maximizing the expected local utility we do in fact maximize the mutual information between B and $Y^{x}_{n_x+1}$ conditional on $\boldsymbol{Y}^{\mathscr{X}}_{n_{\mathscr{X}}}$. The so obtained adaptive sampling scheme, shortly denoted u^{MB}, is fully equivalent to the Bayesian framework of adaptive design optimization (ADO) for model discrimination (cf. Cavagnaro et al. (2010)).

The very same line of reasoning applies to the case where we do have strong evidence for a particular multibin model, say $B \in \mathcal{P}(\mathcal{X})$, but want to optimally learn the psychometric rates $\boldsymbol{p}_{x,y}$. The respective local utility measures are given by

$$u^{\Psi|\mathrm{B}}_{n+1}(x, y, \boldsymbol{E}_n) = \mathrm{D}_{\mathrm{KL}}(\mu[\cdot|\boldsymbol{Y}^{\mathscr{X}}_{n_{\mathscr{X}}}, Y^{x}_{n_x+1}=y, B] \,\|\, \mu[\cdot|\boldsymbol{Y}^{\mathscr{X}}_{n_{\mathscr{X}}}, B]),$$

where $\mu[\cdot|\boldsymbol{Y}^{\mathscr{X}}_{n_{\mathscr{X}}}, B]$ and $\mu[\cdot|\boldsymbol{Y}^{\mathscr{X}}_{n_{\mathscr{X}}}, Y^{x}_{n_x+1}=y, B]$ are the respective successive posterior measures for the psychometric rates $\boldsymbol{p}_{x,y}$ conditional on B. These utility measures can be computed with

[5]The mutual information between two variables X and Y conditional on a third variable Z is formally defined as follows: If $\mu_{Y,\,X|\,Z}$ is the joint measure of X, Y conditional on Z and $\mu_{X|\,Z}, \mu_{Y|\,Z}$ are the respective conditional marginal measures, such that $\mu_{X\times Y|\,Z}$ is the product measure constructed from both marginal measures, then the mutual information between X and Y conditional on Z is defined as $\mathrm{MI}(X{:}\,Y|Z) := \mathrm{D}_{\mathrm{KL}}(\mu_{X,\,Y|\,Z} \| \mu_{X\times Y|\,Z})$, i.e. the Kullback-Leiber divergence from the conditional product measure to the conditional joint measure.

$$u_{n+1}^{\Psi|\mathrm{B}}(x, y, \boldsymbol{E}_n)=\frac{d(\alpha,\beta,\tilde{\gamma},\boldsymbol{n}_{\mathscr{X},\mathscr{Y}})}{d(\alpha,\beta,\gamma,\boldsymbol{n}_{\mathscr{X},\mathscr{Y}})},$$

where

$$\tilde{\gamma}_b=\begin{cases}\gamma_b\left(-\ln\left(\frac{n_{b,y}+\alpha_{b,y}}{n_b+\alpha_b}\right)+\psi(n_{b,x}+\alpha_{b,x}+1)-\psi(n_b+\alpha_b+1)\right) & \text{if } x\in b\\ \gamma_b & \text{otherwise}\end{cases},$$

and ψ is the psigamma function. We expect the resulting adaptive sampling scheme, denoted $u^{\Psi|B}$, to optimize the inferential task for the psychometric rate $\boldsymbol{p}_{x,y}$ given that B is the 'true' underlying model.

In practice, we usually have only vague knowledge about the underlying multi-bin model. Henceforth, it seems reasonable to consider the local utility measures

$$u_{n+1}^{\mathrm{Total}}(x, y, \boldsymbol{E}_n)=\mathrm{D}_{\mathrm{KL}}(\mu[\cdot|\boldsymbol{Y}_{\boldsymbol{n}_{\mathscr{X}}}^{\mathscr{X}}, Y_{n_x+1}^{x}=y]\,||\,\mu[\cdot|\boldsymbol{Y}_{\boldsymbol{n}_{\mathscr{X}}}^{\mathscr{X}}]),$$

where $\mu[\cdot|\boldsymbol{Y}_{\boldsymbol{n}_{\mathscr{X}}}^{\mathscr{X}}]$ and $\mu[\cdot|\boldsymbol{Y}_{\boldsymbol{n}_{\mathscr{X}}}^{\mathscr{X}}, Y_{n_x+1}^{x}=y]$ are the respective successive posterior measures for the psychometric rates $\boldsymbol{p}_{x,y}$. This measure decomposes as

$$u_{n+1}^{\mathrm{Total}}(x, y, \boldsymbol{E}_n)=u_{n+1}^{\mathrm{MB}}(x, y, \boldsymbol{E}_n)+u_{n+1}^{\Psi}(x, y, \boldsymbol{E}_n),$$

where

$$u_{n+1}^{\Psi}(x, y, \boldsymbol{E}_n):=\sum_{E\in\mathscr{P}(\mathscr{X})}P[B|\boldsymbol{Y}_{\boldsymbol{n}_{\mathscr{X}}}^{\mathscr{X}}, Y_{n_x+1}^{x}=y]u_{n+1}^{\Psi|\mathrm{B}}(x, y, \boldsymbol{E}_n)$$

is the model-averaged local utility for optimizing the inference of the psychometric rates given a multibin model. Hence, if we want to learn about $\boldsymbol{p}_{x,y}$ given that B is uncertain, then we also have to make inference about B. That is the uncertainty about B also influences our uncertainty about $\boldsymbol{p}_{x,y}$. We thus expect this strategy, called u^{Total}, to optimize the inference of B and $\boldsymbol{p}_{x,y}$. It might also be worthwile to base the sampling process solely on u^{Ψ} because it allows us to optimize the inference of the psychometric rates regardless of the underlying multibin model.

The actual effect of the proposed adaptive sampling strategies are difficult to describe and we shall proceed by discussing a simple example. In general, it can be said that there is no such thing as a universal criterium for optimal adaptive sampling as long as we do not clearly formulate what we want to achieve. Whether or not a chosen strategy is appropriate for the experiment at hand must be carefully assessed from case to case, for example with simulation studies or experimental pre-studies.

4. A Practical Demonstration

In order to demonstrate our framework we consider a hypothetical 2AFC experiment with 35 stimuli $\mathcal{X} = \{x_1, x_2, \ldots, x_{35}\}$ and a dichotomous outcome $\mathcal{Y} = \{s, f\}$. The stimulus-response relation (ground truth) is given by an asymmetric U-shaped function with a small irregularity on the right side (see Figure 2). We assume that we have no a priori knowledge about the curve, except that proximal stimuli are likely to cause similar responses. In a naïve approach we would simply distribute N measurements random uniformly over $\mathcal{X}$, see (17), and infer the psychometric rates for each $x \in \mathcal{X}$ independently. Figure 3 shows the result of such an experiment with 200 and 400 samples, where we chose pseudo counts $\boldsymbol{\alpha}_{x,\mathcal{Y}} = (1, 1)$, $y \in \mathcal{Y}$. The variance of the estimate is large because there are only very few samples for each stimulus x. A highly irregular curve as an estimate for the stimulus-response relation is therefore obtained and many measurements are taken at regions that are quite uninformative. This motivates two advantages of our proposed framework: First, we can use samples from neighboring stimuli to share statistical strength. Second, we want to distribute the samples such that we maximize the information that we gain with each measurement.

With our framework we can optimize the experimental process. We choose the following parameters:

$$\alpha_{x,\mathcal{Y}} = (1, 1), \quad \beta_m \binom{L-1}{m-1}^{-1}, \quad \gamma_b = 1 \text{ for } b \in \mathcal{C}(\mathcal{X}) \text{ and } m = 1, 2, \ldots, L$$

With $\boldsymbol{\alpha}_{x,\mathcal{Y}} = (1, 1)$ we assign equal pseudo counts to all stimuli and responses, since we have no a priori knowledge about the shape of the curve. The particular choice of β_m and γ_b assigns a uniform distributions to $\mathbf{P}[M = m]$, which is the a priori probability of multibins with m bins (see section 3.6). Finally, we set all $\gamma_b = 1$ since all bins should receive an equal weight.

With this prior setting we observe a smoothening of the inferred stimulus-response curve (see Figure 4). However, in contrast to common kernel smoothing approaches, our method does not smear sharp transitions but represents a higher degree of uncertainty whenever necessary.

In order to distribute samples more efficiently, we can utilize the adaptive sampling strategies described in section 3.7. Unfortunately, it is difficult to compare the performance of different strategies since there is no general criterion for optimality. A good intuition can however be gained from the distribution of measurements.

The Figures 5, 6, 7, and 8 show the time course of the experiment for the adaptive sampling schemes u^{Ψ}, u^{MB}, u^{Total}, and the strategy based on the effective count $\bar{n}_x$. All simulations were initialized with the same random seeds for each stimulus $x \in \mathcal{X}$, such that results can be better compared. The uninformative parameter setting lead to a uniformly distributed expected local utility for all four strategies before the first experiment. Therefore, the first measurement was always taken at $x = 20$.

Already after the first sample one can observe a striking difference between u^{Ψ} and u^{MB}. Whereas the expected utility for u^{Ψ} is maximal at the very left stimulus $x = 1$, the expected utility for u^{MB} shows two maxima directly next to the previous measurement at $x = 20$. This reveals the very distinct properties of both schemes, which are better seen after 10 and 50 samples. u^{Ψ} causes a uniform distribution of the first samples on $\mathcal{X}$. On the other hand, the scheme u^{MB} places all measurements around the initial sample and only gradually moves

further away from $x = 20$. The U-shape of the stimulus-response function is already very well established after 100 samples and after 200 samples one can see that both measures place most samples at positions where the psychometric function is highly sloped. This behavior is expected since those regions allow less sharing of statistical strength. For the scheme u^{Total} we observe the same initial behavior as for u^{Ψ}, but the count statistic after 200 samples shows characteristics of both u^{Ψ} and u^{MB}, which is expected from its definition. It is however quite noteworthy that the effective counts show a very similar behavior as u^{Ψ}. Figure 9 summarizes the differences between the four adaptive sampling strategies by showing the experiment after 200 samples.

Note also that many measurements are taken at the boundaries. This is because these stimuli have only one neighbor which limits the extent to which statistical strength can be shared. We now assume that we have prior knowledge about the ground truth, i.e. we have a strong belief about its value at x_1 and x_{35}. We incorporate this information into our model and thereby optimize the sampling process further. That is, we set $\boldsymbol{\alpha}_{1,\,y} = \boldsymbol{\alpha}_{35,\,y} = (100, 1)$. This parameter setting alters our prior expectation about the stimulus-response relation, i.e. $p_{x,s}$ is expected to be close to one at x_1 and x_{35}, see Figure 10(a). The expected utility to sample at these points is substantially reduced, since we have a strong belief about the response. Figure 10(b) shows the experiment after 200 samples. One can see that all measurements that were previously allocated at the boundaries are now distributed elsewhere. Further optimizations of the sampling process are possible if more prior knowledge is available.

5. Conclusion

We have introduced a framework for adaptive sequential sampling which helps to optimize the measurement process in a wide range of psychophysical experiments, especially when there is only vague prior knowledge about the relation $X \rightarrow Y$. Our framework consists of two major components. The first is a response process that we use to make predictions based on a finite number of observations. We termed it the multibin Pólya mixture process as it consists of a mixture of binned Pólya urns. On top of the response process we defined various adaptive sampling process, which are equipped with utility measures to actively guide the course of an experiment. We also demonstrate the effect of several sampling strategies on a hypothetical experiment and how prior knowledge can be used to further optimize the allocation of measurements. During experiments it is of great importance that decisions for the next stimulus are computed fast. Although our model is computationally demanding, we provide an algorithm that makes it applicable in typical psychophysics experiments.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

All three authors were supported by Max Planck Society. T. E. has been supported by NIH grant R01 EY018664. We thank Claudia Freigang, Pierre-Yves Bourguignon, and Wiktor Mlynarski for most helpful suggestions. We would also like to thank the anonymous reviewers whose suggestions have led to a considerable improvement of the manuscript.

References

Barry D, Hartigan JA. Product partition models for change point problems. The Annals of Statistics. 1992; 20 (1):260–279.

Berger, JO. Statistical decision theory and Bayesian analysis. 2. Springer-Verlag; New York: 1993.

Bernardo JM. Expected information as expected utility. The Annals of Statistics. 1979; 7 (3):686–690.

Bernardo, JM.; Smith, AFM. Bayesian theory. Wiley, Chichester; 1995.

Cavagnaro DR, Myung JI, Pitt MA, Kujala JV. Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. Neural Computation. 2010; 22:887–905. [PubMed: 20028226]

Chaloner K, Verdinelli I. Bayesian experimental design: A review. Statistical Science. 1995; 10 (3): 273–304.

Cronbach, LJ. Tech Rep. Vol. 1. Illinois University; Urbana: Bureau of Research and Service; 1953. A consideration of information theory and utility theory as tools for psychometric problems.

de Finetti B. La prévision: ses lois logiques, ses sources subjectives. Ann Inst Poincare. 1937; 7 (2):1–68.

de Finetti, B. Theory of probability: A critical introductory treatment. Vol. 1. Wiley; London;, New York, N.Y: 1974.

de Finetti, B. On the condition of partial exchangeability. In: Jeffrey, RC., editor. Studies in inductive logic and probability. University of California Press; Berkeley: 1980. p. 193-205.

DeGroot MH. Uncertainty, information, and sequential experiments. The Annals of Mathematical Statistics. 1962; 33 (2):404–419.

DeGroot, MH. Optimal statistical decisions, wiley classics library. Wiley-Interscience; Hoboken and N.J: 2004.

Eggenberger F, Pólya G. Über die statistik verketteter vorgänge. ZAMM - Zeitschrift für Angewandte Mathematik und Mechanik. 1923; 3 (4):279–289.

Elze T, Song C, Stollhoff R, Jost J. Chinese characters reveal impacts of prior experience on very early stages of perception. BMC Neuroscience. 2011; 12:14. [PubMed: 21269486]

Endres D, Földiák P. Bayesian bin distribution inference and mutual information. IEEE Transactions on Information Theory. 2005; 51 (11):3766–3779.

Endres, D.; Oram, M.; Schindelin, J.; Foldiak, P. Bayesian binning beats approximate alternatives: estimating peri-stimulus time histograms. In: Platt, J.; Koller, D.; Singer, Y.; Roweis, S., editors. Advances in Neural Information Processing Systems 20. MIT Press; Cambridge, MA: 2008. p. 393-400.

Fernhead P. Exact and efficient bayesian inference for multiple change-point problems. Statistics and Computing. 2006; 16 (2):203–213.

Geisser, S. Predictive inference: An introduction. Chapman & Hall; London: 1993.

Hartigan JA. Partition models. Communications in statistics. Theory and methods. 1990; 19 (8):2745–2756.

Hutter M. Exact bayesian regression of piecewise constant functions. Bayesian Analysis. 2007; 2 (4): 635–664.

Kontsevich LL, Tyler CW. Bayesian adaptive estimation of psychometric slope and threshold. Vision Research. 1999; 39 (16):2729–2737. [PubMed: 10492833]

Kujala JV. Obtaining the best value for money in adaptive sequential estimation. Journal of Mathematical Psychology. 2010; 54:475–480.

Kujala JV, Lukka TJ. Bayesian adaptive estimation: The next dimension. Journal of Mathematical Psychology. 2006; 50:369–389.

Kuss M, Jakel F, Wichmann FA. Bayesian inference for psychometric functions. Journal of Vision. 2005; 5 (5):8.

Lauritzen, SL. Tech Rep. Vol. 18. Stanford University, Department of Statistics; 1974. On the interrelationships among sufficiency, total sufficiency and some related concepts.

Leek M. Adaptive procedures in psychophysical research. Attention, Perception, Psychophysics. 2001; 63:1279–1292.

Levitt H. Transformed up-down methods in psychoacoustics. The Journal of the Acoustical Society of America. 1971; 49 (2B):467–477. [PubMed: 5541744]

Lindley DV. On a measure of the information provided by an experiment. The Annals of Mathematical Statistics. 1956; 27 (4):986–1005.

Link, G. Representation theorems of the de finetti type for (partially) symmetric probability measures. In: Jeffrey, RC., editor. Studies in inductive logic and probability. University of California Press; Berkeley: 1980. p. 207-231.

MacKay DJC. Information-based objective functions for active data selection. Neural Computation. 1992; 4:590–604.

McKee S, Klein S, Teller D. Statistical properties of forced-choice psychometric functions: Implications of probit analysis. Attention, Perception, & Psychophysics. 1985; 37 (4):286–298.

Müller P, Berry DA, Grieve AP, Smith M, Krams M. Simulation-based sequential bayesian design: Special issue: Bayesian inference for stochastic processes. Journal of Statistical Planning and Inference. 2007; 137 (10):3140–3150.

Roberts HV. Probabilistic prediction. Journal of the American Statistical Association. 1965; 60 (306): 50–62.

Taylor MM, Creelman CD. Pest: Efficient estimates on probability functions. The Journal of the Acoustical Society of America. 1967; 41 (4A):782–787.

Watson A, Fitzhugh A. The method of constant stimuli is inefficient. Attention, Perception, & Psychophysics. 1990; 47 (1):87–91.

Watson A, Pelli D. Quest: A bayesian adaptive psychometric method. Attention, Perception, & Psychophysics. 1983; 33:113–120.

Wichmann F, Hill N. The psychometric function: I. fitting, sampling, and goodness of fit. Attention, Perception, & Psychophysics. 2001; 63:1293–1313.

Yao YC. Estimation of a noisy discrete-time step function: Bayes and empirical bayes approaches. The Annals of Statistics. 1984; 12 (4):1434–1447.

Appendix

In Yao (1984); Barry and Hartigan (1992); Endres and Földiák (2005); Fernhead (2006); Hutter (2007) various algorithms are presented, which in their given context address all the very same computational problem. Within our terminology of proximal multibins the general problem can be described as follows:

Let $f\colon C(\mathcal{X}) \to \mathbb{R}$ be any function from consecutive proximal bins to real numbers and $\boldsymbol{g} = (g_1, \ldots, g_L) \in \mathbb{R}^L$ be any real valued vector. Define the sum-products

$$S_m[f] := \sum_{B \in \mathcal{P}_m(\mathcal{X})} \prod_{b \in B} f(b), m = 1, \ldots, L,$$

and

$$S[f, \boldsymbol{g}] := \sum_{B \in \mathcal{P}(\mathcal{X})} g_{|B|} \prod_{b \in B} f(b).$$

Each $S_m[f]$ consists of $\binom{L-1}{m-1}$ terms, whereas the weighted total sum-product $S[f, \boldsymbol{g}]$ consists of 2^{L-1} terms. For large L a computationally intractable effort of $\mathcal{O}(2^{L-1})$ is expected. However, the following lemma provides a simple and efficient algorithm, which computes the sum-product in $\mathcal{O}(L^3)$.

Lemma 1 (The Proximal Multibin Summation (ProMBS) Algorithm)

Define the upper triangular matrices $A_l=(a^l_{i,j})_{L\times L}$, $l = 1, \ldots, L$, recursively by

$$a^1_{i,j}=\begin{cases} f(b_{i,j}) & \text{if } i \le j; \\ 0 & \text{otherwise}; \end{cases} \quad \text{and} \quad a^l_{i,j}=\begin{cases} a^{l-1}_{i-1,j-1} & i, j>1; \\ 1 & i=j=1; \\ 0 & \text{else}; \end{cases}$$

and define recursively the matrix-vector product

$$\boldsymbol{v}^{l+1}=(v^{l+1}_1,\ldots,v^{l+1}_L)^T=A_l\times\boldsymbol{v}^l,\ l=1,\ldots L,$$

where the initial vector is $\boldsymbol{v}^1 = (0, \ldots, 0, 1)^T$. For every $m, l \in \mathbb{N}$ with $m \le l \le L$ it holds that

$$v^{l+1}_m=S_m[f],$$

such that

$$S[f,\boldsymbol{g}]=\sum_{m=1}^{L} g_m S_m[f].$$

A simple example might help to illustrate the abstract formulation of the ProMBS algorithm. Consider the case with $|\mathcal{X}| = 3$ and let us abbreviate $f(b_{i,j})$ simply by f_{ij}. In the first step we compute

$$\boldsymbol{v}^2=A_1\boldsymbol{v}^1=\begin{pmatrix} f_{11} & f_{12} & f_{13} \\ 0 & f_{22} & f_{23} \\ 0 & 0 & f_{33} \end{pmatrix}\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}=\begin{pmatrix} f_{13} \\ f_{23} \\ f_{33} \end{pmatrix}.$$

Note that $v^2_1=f_{13}=S_1[f]$. In the next step we compute

$$\boldsymbol{v}^3=A_2\boldsymbol{v}^2=\begin{pmatrix} 1 & 0 & 0 \\ 0 & f_{11} & f_{12} \\ 0 & 0 & f_{22} \end{pmatrix}\begin{pmatrix} f_{13} \\ f_{23} \\ f_{33} \end{pmatrix}=\begin{pmatrix} f_{13} \\ f_{11}f_{23}+f_{12}f_{33} \\ f_{22}f_{33}, \end{pmatrix}$$

where $v^3_1=f_{13}=S_1[f]$ and $v^3_2=f_{11}f_{23}+f_{12}f_{33}=S_2[f]$. In the last step we find that

$$\boldsymbol{v}^4 = A_3\boldsymbol{v}^3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & f_{11} \end{pmatrix} \begin{pmatrix} f_{13} \\ f_{11}f_{23}+f_{12}f_{33} \\ f_{22}f_{33} \end{pmatrix} = \begin{pmatrix} f_{13} \\ f_{11}f_{23}+f_{12}f_{33} \\ f_{11}f_{22}f_{33} \end{pmatrix},$$

such that $(\boldsymbol{v}^4)^T = (S_1[f], S_2[f], S_3[f])$ and $S[f, \boldsymbol{g}] = g_1S_1[f] + g_2S_2[f] + g_3S_3[f]$.

Of course, the real computational power of the ProMBS algorithm becomes only evident when the set $\mathcal{X}$ is large, but the case $|\mathcal{X}| = 3$ shows the general logic behind the ProMBS algorithm sufficiently enough. We should finally mention that numerical imprecisions can occur if the values of f become small. In such cases it is advisable to implement the ProMBS algorithm on a logarithmic number scale.

Highlight

- We present a predictive account on adaptive sampling in psychophysical experiments.
- Our method applies to situations where there is only weak knowledge available.
- We demonstrate the advantages of our framework on a hypothetical sampling problem.

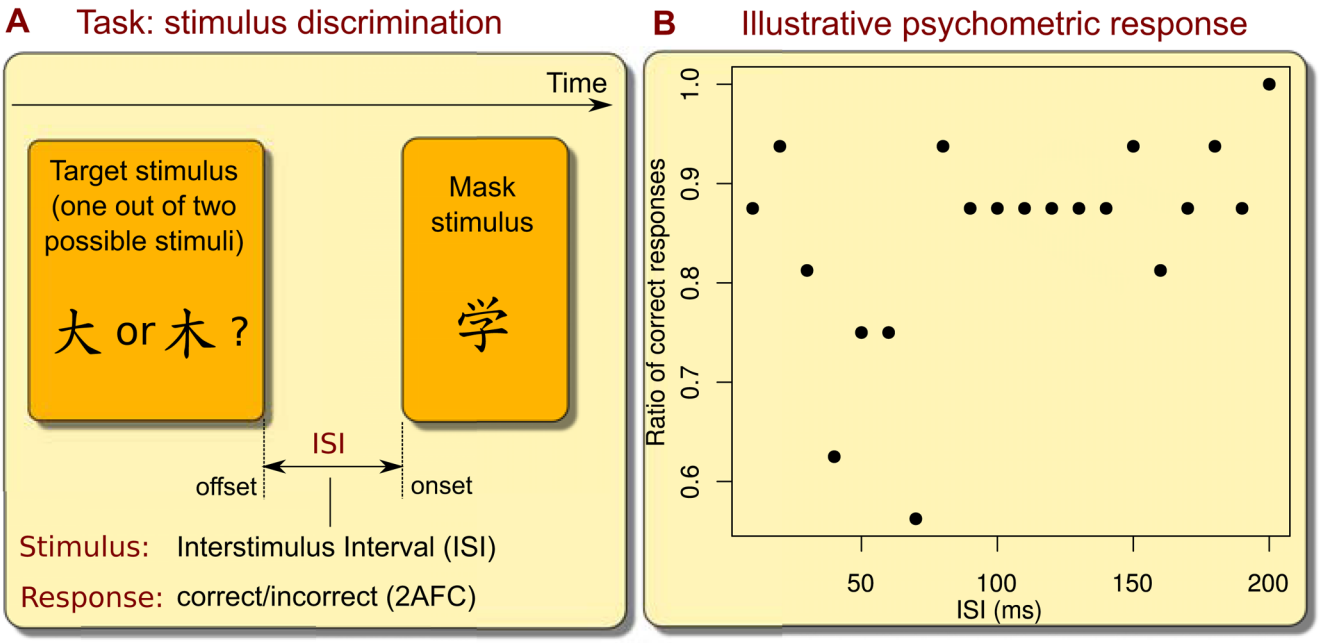

Figure 1.

Illustrative example of a simple psychophysical experiment. **A**: The observer has to report which of two possible target stimuli were presented on a computer display. The detection of the target stimulus is impaired by a mask stimulus which is presented in close proximity to the position of the target after a variable interstimulus interval (ISI). **B**: The observed relative frequency of correct responses of a single observer for 20 different discrete ISIs ≤ 200 ms, which have been uniformly sampled. Each ISI occured 16 times.

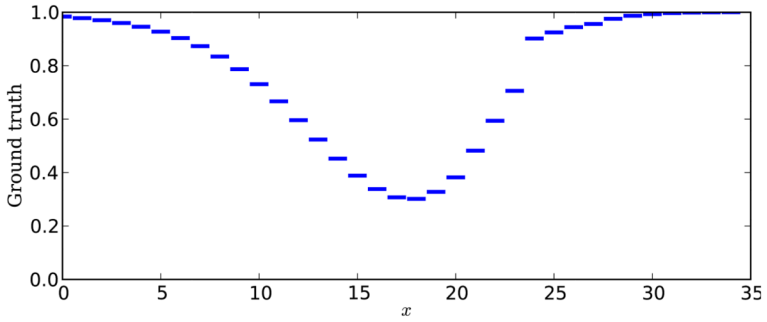

Figure 2.
Hypothetical stimulus-response curve used as ground truth for the practical demonstration. The curve shows the probability of the response s (success) for a given stimulus x.

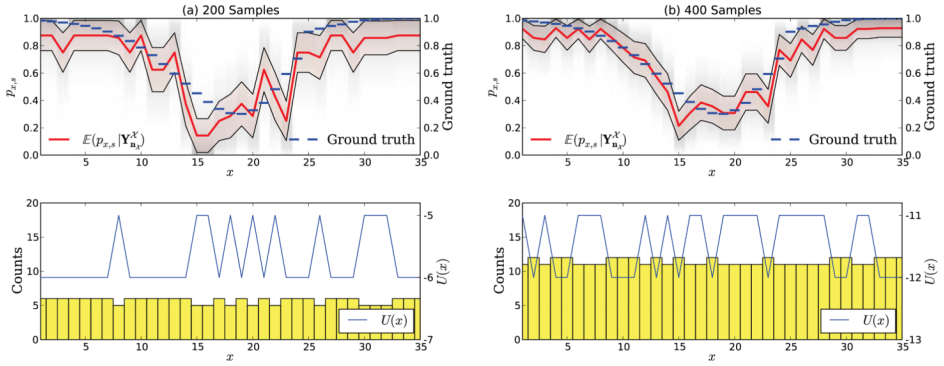

Figure 3.

A hypothetical experiment with (a) 200 and (b) 400 samples uniformly distributed over the stimulus space $\mathcal{X}$. The stimulus-response relation (ground truth), shown as dashed line, is inferred without taking information from proximal x into account. The thick continuous line shows the first moment $\mathbb{E}[p_{x,s}|\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}}]$ of the stimulus-response function given all outcomes of previous measurements. The standard deviation is shown as a thin continuous line around the expectation. The marginal posterior density $f(p_{x,s}|\boldsymbol{Y}^{\mathcal{X}}_{\boldsymbol{n}_{\mathcal{X}}})$ is plotted as shadings in the back of the figure. The number of measurements at each $x \in \mathcal{X}$ is shown as a bar plot in the lower plot with utility $U(x)$ (thin continuous line), which is the negative number of counts, see (17).

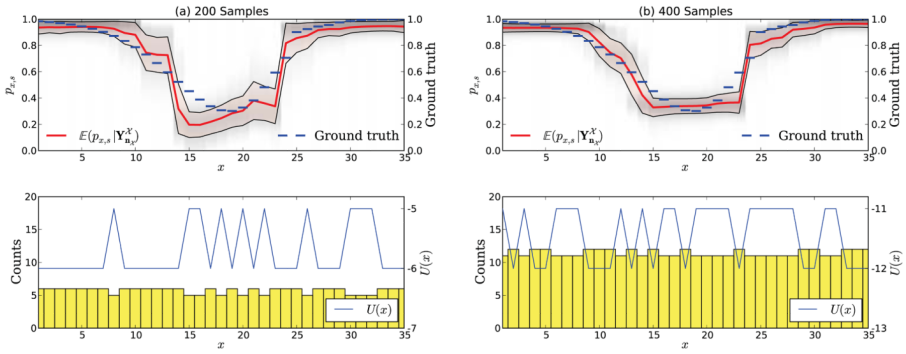

Figure 4.

A hypothetical experiment with (a) 200 and (b) 400 samples uniformly distributed. The ground truth is inferred by taking information from neighboring x into account. This sharing of statistical strength allows a much more accurate inference of the stimulus-response function as compared to Figure 2.

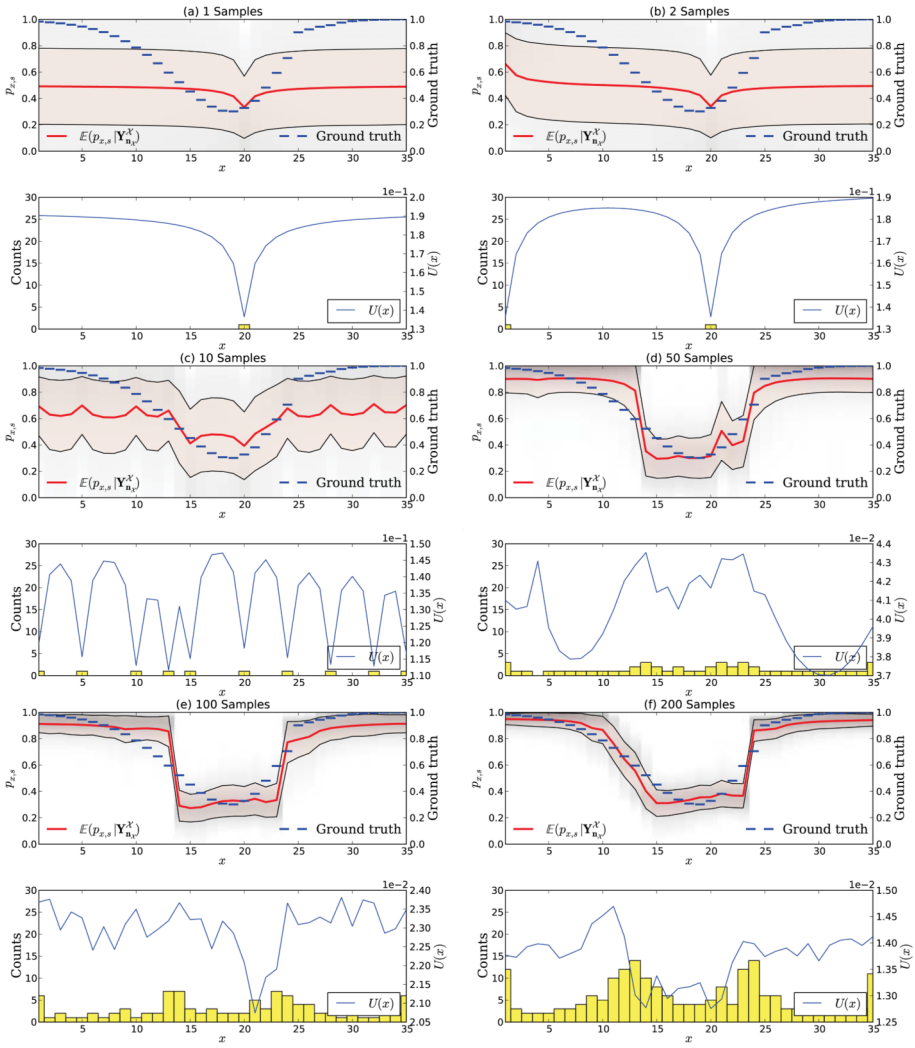

Figure 5.

Adaptive sampling in a hypothetical experiment with scheme u^{Ψ}. The figure shows the experiment after 1, 2, 10, 50, 100, and 200 samples. At first, samples are uniformly distributed. In (a) only one measurement was taken and the expected utility is largest at the left boundary. (d) shows the experiment after 50 samples where the algorithm starts to locate measurements at sloped regions. The general shape of the stimulus-response function is already well established after 100 samples (e). Many measurements are also taken at $x = 1$ and $x = 35$ since those x have only one neighbor.

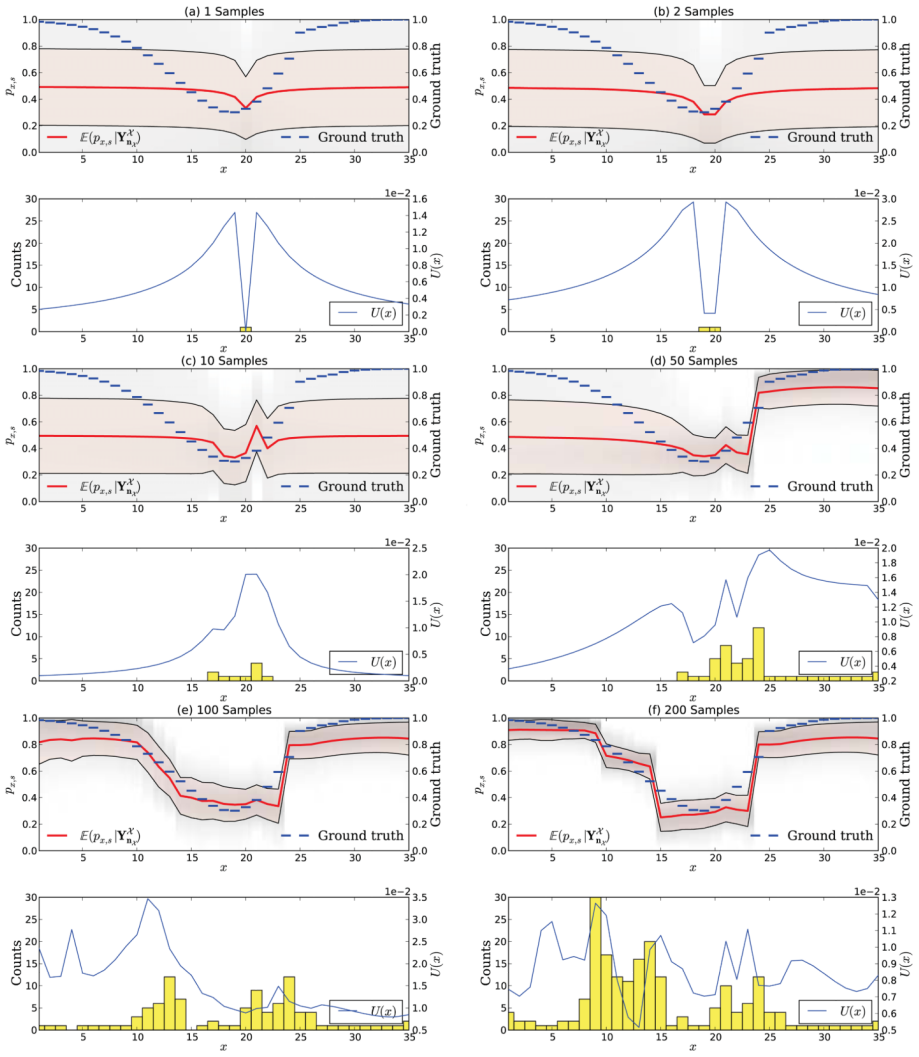

Figure 6.

Adaptive sampling in a hypothetical experiment with scheme u^{MB}. The figure shows the experiment after 1, 2, 10, 50, 100, and 200 samples. After the first sample (a) two maxima of the expected utility arise next to the measurement. The first 10 samples (c) are allocated near the initial measurement, whereas afterwards the algorithm starts move further right (d) until almost the full stimulus space has been explored (e–f).

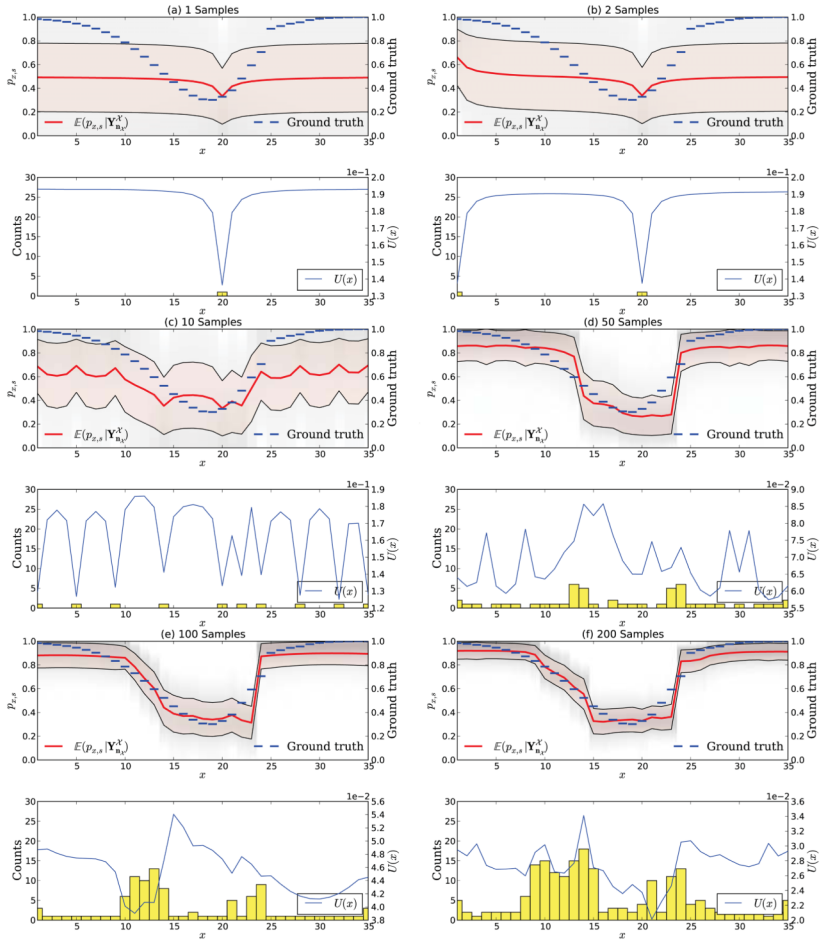

Figure 7.

Adaptive sampling in a hypothetical experiment with scheme u^{Total}. The figure shows the experiment after 1, 2, 10, 50, 100, and 200 samples. The sampling behavior clearly shows a mixture of both schemes u^{MB} and u^{Ψ}.

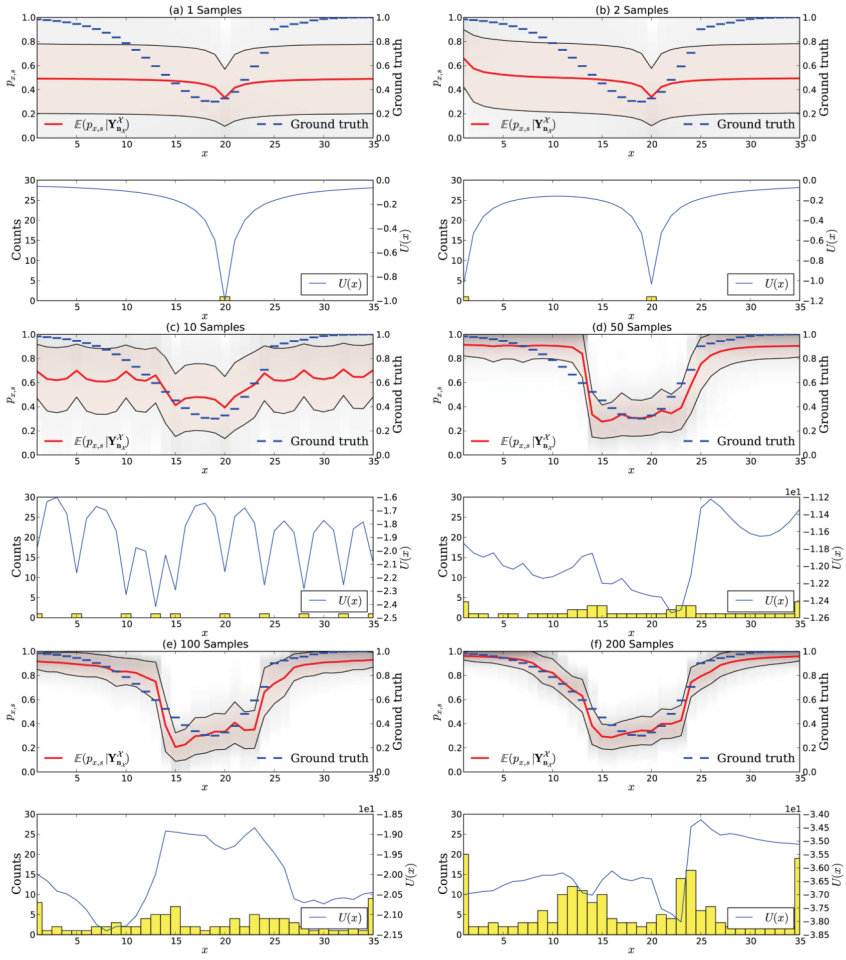

Figure 8.

Adaptive sampling in a hypothetical experiment with the strategy based on the effective count $\tilde{n}_x$. The figure shows the experiment after 1, 2, 10, 50, 100, and 200 samples. In (a) no measurements are taken and we can see our prior expectation. The respective utility is uniformly distributed on $\mathcal{X}$. The general shape of the stimulus-response function is already well established after 100 samples. Measurements are mostly allocated at regions where the slope of the stimulus-response function is high. Many measurements are also taken at $x = 1$ and $x = 35$ since those x have only one neighbor.

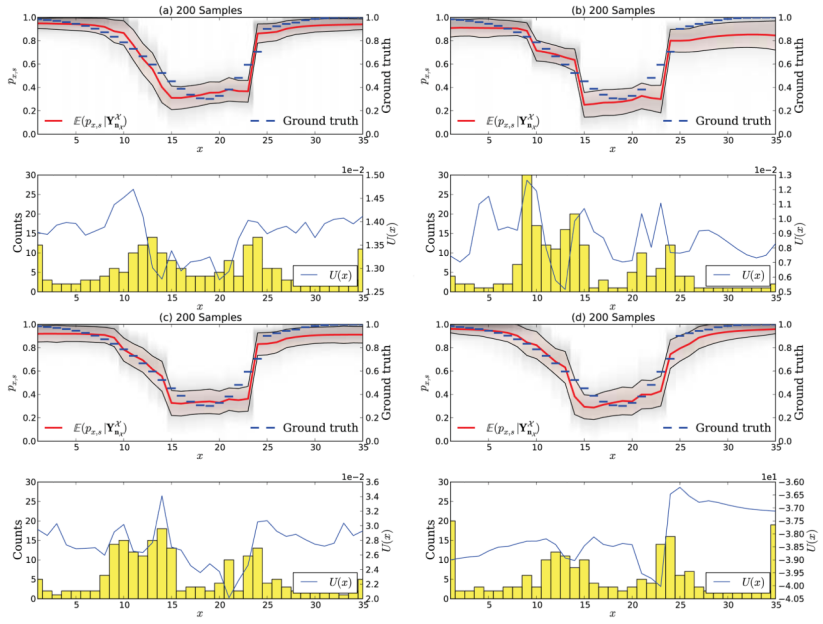

Figure 9.

Direct comparison of the four adaptive sampling strategies: (a) u^{Ψ}, (b) u^{MB}, (c) u^{Total}, and (d) effective counts. A relatively balanced distributions of measurements is observed with u^{Ψ} and the effective counts. On the other hand, u^{MB} and therefore also u^{Total} lead to a more peaked distribution.

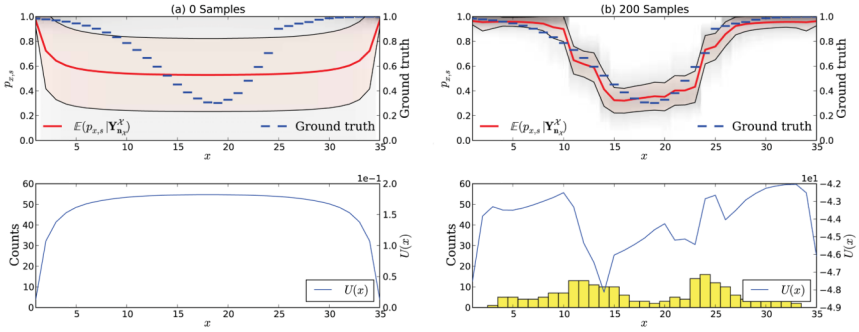

Figure 10.

Demonstration of using prior knowledge to limit sampling at the boundaries with scheme u^{Ψ}. (a) shows the prior belief before any measurements are made and (b) shows the posterior after 200 samples. The prior setting leads to a strong belief about the response at the boundaries which substantially reduces the expected utility at x_1 and x_{35}. The result is that all measurements that were previously allocated at the boundaries are now placed elsewhere.