BRILL

Seeing and Perceiving 23 (2010) 483–515

brill.nl/sp

Functional Adaptive Sequential Testing

Edward Vul *, **Jacob Bergsma and Donald I. A. MacLeod**

Department of Psychology, University of California, San Diego, USA

Received 10 September 2009; accepted 4 August 2010

Abstract

The study of cognition, perception, and behavior often requires the estimation of thresholds as a function of continuous independent variables (e.g., contrast threshold as a function of spatial frequency, subjective value as a function of reward delay, tracking speed as a function of the number of objects tracked). Unidimensional adaptive testing methods make estimation of single threshold values faster and more efficient, but substantial efficiency can be further gained by taking into account the relationship between thresholds at different values of an independent variable. Here we present a generic method — functional adaptive sequential testing (FAST) — for estimating thresholds as a function of another variable. This method allows efficient estimation of parameters relating an independent variable (e.g., stimulus spatial frequency; or reward delay) to the measured threshold along a stimulus strength dimension (e.g., contrast; or present monetary value). We formally describe the FAST algorithm and introduce a Matlab toolbox implementation thereof; we then evaluate several possible sampling and estimation algorithms for such two-dimensional functions. Our results demonstrate that efficiency can be substantially increased by considering the functional relationship between thresholds at different values of the independent variable of interest.

Keywords

Psychophysical methods, adaptive testing, Bayesian inference

1. Introduction

In many psychophysical and cognitive experiments, stimuli are presented in multiple trials, and the subject makes a response that is classified into a binary alternative (e.g., seen/not seen, correct/incorrect, or too much/too little) after each trial. For instance, in a taste detection experiment, one might aim to estimate how many grams of sugar should be dissolved in liter of water to be detected 90% of the time; in contrast sensitivity experiments: the contrast at which subjects respond with 75%

* To whom correspondence should be addressed. E-mail: evul@ucsd.edu

This article is part of the 2010 Fechner collection, guest edited by J. Solomon.

DOI:10.1163/187847510X532694

accuracy; or in delay discounting experiments: the immediate value of a delayed reward corresponds to a 50% chance of preferring a smaller immediate reward to the delayed reward, etc. In general, the 'psychometric function' relates the proportion of positive responses to stimulus strength (e.g., contrast, chromaticity, or immediate reward value), and the results of such experiments are often summarized as the stimulus strength necessary for a certain proportion of positive responses (e.g., a threshold for detection, a threshold level accuracy for discrimination, or a match point giving equal proportions of positive and negative responses for matching).

Fechner's pioneering discussion of psychophysical methods (Fechner, 1860) introduced three methods to estimate these threshold: the Method of Constant Stimuli, the Method of Limits, and the Method of Adjustment. Because the method of limits tends to produce bias through hysteresis effects and the method of adjustment yields slow, and potentially over-thought, responses, the method of constant stimuli has been more often used as a general procedure for experiments of this type. In the method of constant stimuli, the stimulus strength on any trial is selected randomly from a fixed set. Due to the importance and prevalence of threshold estimation, and the growing need to test more conditions faster, many adaptive testing techniques have been developed to expedite the estimation process over and above the original method of constant stimuli. The initial development of these techniques focused on unidimensional threshold estimation: the parameters of a psychometric function, or just a single point along the psychometric function.

However, modern psychophysical and cognitive experiments almost always measure not just a single threshold or match point, but how the threshold or match point changes depending on some independent stimulus or task parameter of interest: for instance, the contrast sensitivity function, which is defined by the change of contrast threshold (the reciprocal of sensitivity) with spatial frequency. A single experiment typically has the goal of estimating such a function, relating (usually) threshold stimulus strengths or (less often) match points or null settings (we can think of these measures as dependent variables in an experiment) to other stimulus parameters (independent variables). We will refer to such a function, that a particular experiment is intended to estimate, as the '*threshold function*', because the threshold case is the most common. However, our discussion and proposals apply equally to matching or null measurements that are based on responses that may be classified into binary outcomes.

The threshold function in the sense used here is quite different from the psychometric function that relates the proportion of positive responses to stimulus strength. The threshold function specifies how the psychometric function as a whole depends on some independent variable — for example, spatial frequency in the case of the contrast sensitivity function. Typically the independent variable translates the psychometric function on a suitably chosen axis. For example the change in contrast sensitivity with changing frequency can be considered as a translation of the psychometric function on a scale of log contrast. Likewise the threshold *vs* contrast (TvC) function describes the just detectable contrast increment as a function of the

contrast to which the increment is added; cortical magnification may be measured as the threshold size necessary for some probability of detection (e.g., gap size for Vernier acuity) as a function of eccentricity; the time-course of adaptation may be expressed by the nulling stimulus strength as a function of adaptation time; and the delay-discounting function may be expressed by the subjectively equivalent proportion of immediate to delayed reward amount.

In all these cases and indeed in most psychophysical experiments, the estimation problem is two dimensional, in the sense that the goal is the estimation of threshold as a function of another variable. However, unidimensional adaptive methods for assessing thresholds are less efficient for such problems, and there has been relatively little work on general formalisms and algorithms for two-dimensional adaptive testing. This is the problem we aim to address with Functional Adaptive Sequential Testing: we seek to make estimation of threshold functions more efficient, just as estimation of unidimensional thresholds has been made more efficient with the development of adaptive testing methods since Fechner introduced psychophysics.

2. Unidimensional Adaptive Testing

While there is no need for a thorough review of unidimensional adaptive testing here (for useful reviews see: King-Smith and Rose, 1997; Klein, 2001; Leek, 2001; McKee *et al.*, 1985; Treutwein, 1995) it is useful to note where and how progress has been made on this problem, as efficient estimation of the two-dimensional threshold function must begin with efficient methods for the one-dimensional case.

Early methods for threshold estimation (Fechner's method of constant stimuli; Fechner, 1860) were inefficient (see Note 1). Many trials were wasted: stimuli were presented at very high or very low stimulus strengths, at which subjects will nearly always be correct or incorrect, and thus responses are minimally informative. This was improved with adaptive reversal staircases, which reduced the frequency of testing at the less informative stimulus strengths (Dixon and Mood, 1948; Wetherill and Levitt, 1965).

Later, by explicitly formalizing the underlying psychometric function, procedures like QUEST (Taylor and Creelman, 1967; Watson and Pelli, 1983) could achieve even greater efficiency. By fixing the slope parameter of the underlying function *a priori*, Watson and Pelli could use data from all of the trials to find the most informative stimulus placement to estimate the threshold. Others (Cobo-Lewis, 1997; King-Smith and Rose, 1997; Kontsevich and Tyler, 1999) extended this procedure to estimate the slope parameter concurrently with the threshold, thus rendering the estimation procedure more efficient and robust under variable slopes. These are the current state-of-the-art adaptive, parametric threshold estimation procedures (see Note 2).

3. Multidimensional Adaptive Testing

Despite this progress in estimating single thresholds, estimation of threshold functions has seldom been considered. A straightforward approach has been called the Method of 1000 Staircases (Cornsweet and Teller, 1965; Mollon *et al.*, 1987): To estimate the threshold function, estimate a threshold at many points along the function independently, and then fit a function to those points. The threshold at each point of the function may be estimated efficiently (for instance with QUEST or a reversal staircase), but because each threshold is estimated independently, such a procedure encounters inefficiencies similar to those that arise in the method of constant stimuli. Because each staircase is independent, it must start with assumptions of ignorance, leading to several trials at minimally informative stimulus strengths — across all of the independent staircases, many trials are wasted in this manner.

The QUEST (Watson and Pelli, 1983) and the Psi (Kontsevich and Tyler, 1999) procedures allow more efficient unidimensional threshold estimation by choosing the presented stimulus values that best constrain the parameters of the psychometric function. Similarly, estimation of the threshold function can be made efficient by specifying that function formally in terms of parameters whose values are to be experimentally estimated. With the assumption that the slope of the psychometric function is the same at points along the underlying threshold function, one may define a probability surface (Fig. 1) with a relatively parsimonious parameterization (Fig. 2).

This idea has previously been implemented for the cases of estimating the TvC function (Lesmes *et al.*, 2006), the CSF function (Lesmes *et al.*, 2010), and the somewhat more general case of ellipsoidal threshold functions (Kujala and Lukka, 2006). Although these examples use specific parametric models for specific applications, the assumption of independent translation of the psychometric function across the threshold function can be fruitfully extended to a very broad family of parametric models that cover most experiments aiming to estimate parametric threshold

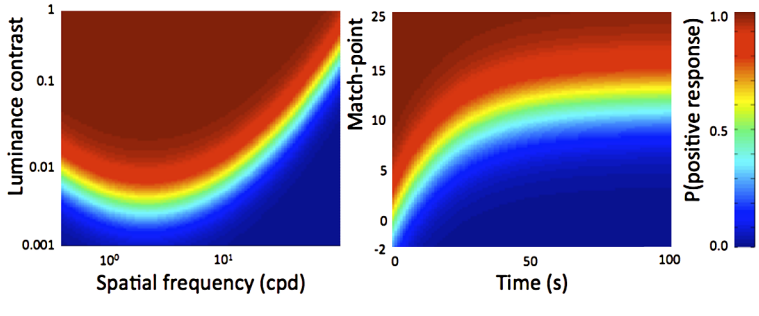

Figure 1. (Left panel) The two dimensional response probability surface corresponding to a contrast sensitivity function (probability of detection as a function of grating spatial frequency and contrast; from equation (3)). (Right panel) The same surface for a decaying exponential curve, as would be seen in the timecourse of aftereffects. This figure is published in colour on http://brill.publisher.ingentaconnect.com/content/vsp/spv

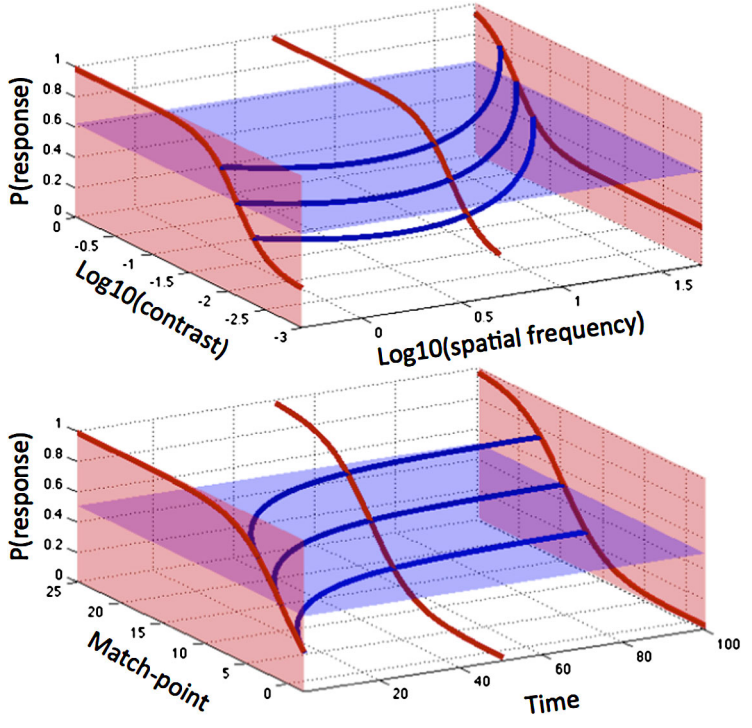

Figure 2. Like many two-dimensional functions of psychophysical interest, the contrast sensitivity and decayed exponential response probability surfaces can be defined by factoring into a threshold function (relating threshold contrast to spatial frequency, and match-point to time) and an invariant psychometric function (relating probability of detection to stimulus contrast or match-point). This figure is published in colour on http://brill.publisher.ingentaconnect.com/content/vsp/spv

functions, as suggested by Lesmes *et al.* (2006, 2010). Here, we write down the formalization of such a generalized two-dimensional testing procedure and provide a generic implementation of such an algorithm.

There are two primary advantages to characterizing the threshold function using parametric optimization, rather than independently measuring points along the threshold function. First and foremost, each point along the threshold function is informed by every other one: each trial contributes information about the psychometric functions at all values of the independent variable by changing the most likely values of the threshold function parameters (see Note 3). The utilization of data across the threshold function speeds up the selection of optimally informative stimulus parameters at each point as the sequence of trials progresses (Fig. 3). Thus, estimation of all thresholds is more efficient and fewer trials are necessary to converge on an estimate of the underlying function. Second, because any point along this function can be used to inform the overall estimate of the function parameters, estimation need not be restricted to several discrete points: instead one can sample any point along the function. Such continuous sampling not only produces prettier

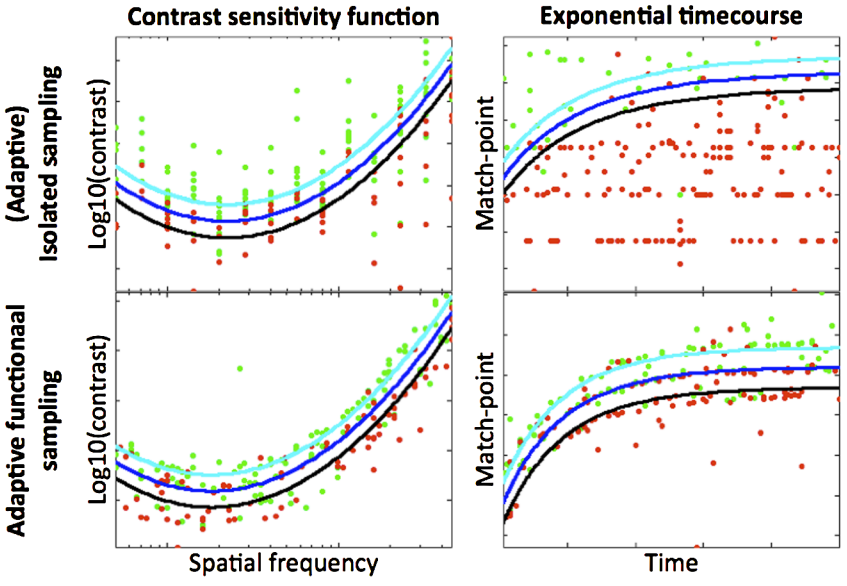

Figure 3. 200 trials of a simulated experiment on the contrast sensitivity function and exponential time-course using isolated sampling (independent adaptive threshold estimation at each point; upper graphs), and adaptive functional sampling (lower graphs). Adaptive functional sampling can sample x values randomly, and samples fewer trials at uninformative points of the stimulus space, where the y value is far from its corresponding threshold. Crosses (red) correspond to negative responses; circles (green): positive responses; the curves correspond to response probability quantiles of the final function fit. This figure is published in colour on http://brill.publisher.ingentaconnect.com/content/vsp/spv

graphs (Fig. 3), but may also indicate systematic deviations from the assumed functional form that may otherwise be missed if the function is sampled at only a few discrete points.

We will begin with a formal description of functional adaptive sequential testing, then briefly describe the Matlab toolbox where we have implemented this method, and present simulations and experiments demonstrating the reliability and efficacy of this tool.

4. Theoretical Framework

The goal of functional adaptive sequential testing (FAST) is to estimate efficiently the response probability as a function of stimulus strength (y) and an independent mediating variable (x). As described previously for the TvC function (Lesmes *et al.*, 2006) and an ellipsoid function (Kujala and Lukka, 2006), this two-dimensional probability surface is parsimoniously parametrized by partitioning it into two functions: a psychometric function, and a threshold function that describes the translation of the psychometric function along the stimulus strength axis.

First, the probability of a response is described by a psychometric link function:

$$p_r = \Psi(y_p, y_c, S), \tag{1}$$

where p_r is the probability of a particular response which we will refer to as the positive response (for instance 'yes' in a detection task, or the correct alternative in a forced choice task, or 'too much' in a matching task). Ψ is the psychometric function, which may be of any formulation desired (logistic, cumulative normal, etc.). y_p is the presented stimulus strength — for instance, if estimating a contrast sensitivity function the y dimension would typically be the logarithm of the test grating contrast. y_c is the 'threshold' parameter that determines the translation of the psychometric function along the y dimension. S is the width (or inverse slope) parameter of the psychometric function — while psychometric functions may differ in what this slope parameter means, all have a parameter that corresponds to the steepness of the function. Other parameters of the psychometric function (such as lapse probability, or guessing probability) are typically fixed, and need not be estimated through testing.

The threshold function relates the critical value, or translation parameter, of the psychometric function described above to the independent variable of interest (for example, the spatial frequency of a grating in the case of the contrast sensitivity function):

$$y_c = F(x, \Theta), \tag{2}$$

where F is the 'threshold function': a function that relates the independent variable, x, to the translation of the psychometric function, y_c. Θ is a vector of parameters of the threshold function, and the ultimate goal of the experiment is to estimate those parameters.

To make this abstract formulation more specific, we continue with the contrast sensitivity example and assume that we are estimating the probability of detection as a function of grating contrast and spatial frequency. We employ the logistic psychometric function for Ψ as a function of the log contrast y_p (lapse probability parameters are excluded in this example). For the CSF we adopt an intuitive parameterization of the CSF as a log parabola (Pelli *et al.*, 1986) for F, that allows sufficient flexibility in specifying y_c as a function of the spatial frequency x (see Figs 1, 2 and 3 for plots of this function). Note that this is a simpler function than that of Watson and Ahumada (2005) as used in Lesmes *et al.* (2010), although it appears sufficient when no particularly low spatial frequencies are being tested:

$$\begin{aligned} p_r &= 1/(1 + \exp((y_p - y_c)/S)), \\ y_c &= \Theta_1 + 10^{\Theta_3}(\log_{10}(x) - \Theta_2)^2. \end{aligned} \tag{3}$$

Here y_c corresponds to the decimal log of contrast threshold. The parameter Θ_1 is the decimal log of the contrast threshold at peak sensitivity, Θ_2 is the decimal log of the spatial frequency that yields peak sensitivity, and Θ_3 is the decimal log of the inverse bandwidth of the CSF.

For any trial with presented stimulus values x_p and y_p the theoretical probability of either possible response (r, quantified as 1 for the 'positive' response or 0 for the

other response option, corresponding to yes/no or correct/incorrect), can be stated for any set of model parameters (Θ and S) as:

$$P(r|x_{\mathrm{p}}, y_{\mathrm{p}}, \Theta, S) = \begin{cases} p_{\mathrm{r}}, & \text{for } r = 1, \\ (1 - p_{\mathrm{r}}), & \text{for } r = 0. \end{cases} \tag{4}$$

Multiplying this probability by the prior likelihood function in the model parameter space in accordance with Bayes theorem, we obtain the posterior probability of a given set of parameters (Θ and S) as:

$$P_{\mathrm{post}}(\Theta, S|x_{\mathrm{p}}, y_{\mathrm{p}}, r) \sim P(r|x_{\mathrm{p}}, y_{\mathrm{p}}, \Theta, S) P_{\mathrm{prior}}(\Theta, S), \tag{5}$$

where $P_{\mathrm{prior}}(\Theta, S)$ corresponds to the prior probability of a given set of Θ and S parameters. The prior can be uninformative at the onset of the experiment, but after the first trial will reflect any information gained from previous trials — that is, the posterior from the current trial will be the prior for the subsequent trial. With this equation, all of the elements necessary for Bayesian estimation of the posterior probability of the parameters of Ψ and F are in place, and we must next decide how to sample stimuli for efficient estimation of these parameters.

4.1. Stimulus Placement

Placement of a stimulus on any given trial determines the information we should expect to obtain on that trial. There are several available placement strategies that may be roughly divided along two dimensions. First, should we choose a *global* 'optimal' position in both dimensions (x and y), or a *local* optimum by picking the best y for a given x? Second, what defines 'optimality'? What follows is a description and discussion of this space of placement strategies as implemented in FAST.

4.2. Global Placement Strategies

One may decide to choose x and y values simultaneously to find the globally most informative stimulus parameters (Lesmes *et al.*, 2006). The obvious advantage of choosing the globally most informative point is that the trial is guaranteed to provide the most useful information, given the assumed parameterization of the response probability surface. One drawback of global placement strategies is that if the form of the assumed threshold function is wrong, global minimization may choose (x, y) locations that would be most informative if the function were correct, but are not efficient for detecting deviations from this hypothesis and may not be efficient for estimating parameters of an alternative function, once the deviation is established. Moreover, the computations required for choosing the globally optimal stimulus may be slower than is desirable.

4.3. Local Placement Strategies

Alternatively, one may choose the locally optimal y value given a fixed x (choosing the most informative contrast given a pre-determined spatial frequency), while x is either sampled randomly, or according to some fixed scheme (for instance,

decay-time in adaptation experiments). Although local placement strategies are theoretically less efficient than choosing the global optimum, the lost efficiency is traded for increased robustness.

4.4. 'Optimality': Minimizing Expected Posterior Entropy

The principle of minimizing expected posterior entropy of the multimensional probability density function for the parameter values has been recently advocated as a method for stimulus placement (Cobo-Lewis, 1997; Kontsevich and Tyler, 1999; Kujala and Lukka, 2006, Lesmes *et al*., 2006). This method is motivated by information theory and selects stimuli so as to maximize the certainty about parameter values one may have after the trial, thus maximizing the information gained from that trial. This method has been described in sufficient detail in the papers cited above, and our procedure is broadly similar to theirs (see Appendix B). This method has the advantage of providing the theoretically most informative point, given the goal of increasing confidence about parameter values; however, the tradeoff between efficiency of estimating one or another parameter is made arbitrarily based on the scale and resolution of those parameter values. In practice, small changes along one parameter might make a much larger difference to the response probability surface than large changes along another parameter, and expected posterior entropy of the distribution of parameter values cannot take these factors into account.

4.5. 'Optimality': Minimizing Average Expected Posterior Predictive Variance

One principled method for trading off efficiency in estimating one parameter or another would be to explicitly take into account the impact of these parameters on the response probability surface. To do so, we define a new optimality criterion: Minimizing expected average posterior predictive variance (or 'minimizing predictive variance' for short). Essentially, this amounts to finding a stimulus value which is expected to alter the posterior distribution in a manner that decreases the uncertainty about the predicted threshold (y^*) values across a range of x values deemed relevant (see Appendix C for a mathematical treatment). On this optimality criterion, the tradeoff between different parameters is made implicitly through their impact on the predicted threshold values: focusing on the statistics of the expected threshold distribution ensures that trials will not be wasted to efficiently estimate parameters that have little impact on the predicted thresholds. One could choose to minimize the entropy, rather than the variance, of the posterior predictive pdf of the threshold. But variance minimization is appropriate if larger errors carry a greater cost than small ones. Variance minimization is optimal when cost increases as the square of the error, while entropy minimization is not supported by any such increasing cost function, instead minimizing errors without regard to their size (Twer and MacLeod, 2001). The minimum expected variance criterion provides a principled method for trading off the precision of different parameters, that can be generally applied to all threshold functions. But it has some drawbacks. It provides no incentive for ef-

ficient estimation of the slope parameter of the psychometric function. It is slower to compute than minimum expected posterior entropy, and most important, in our hands (see simulation section) it does not in practice provide a substantial benefit.

4.6. 'Optimality': Obtaining Constant Response Probabilities

A classical, and intuitive, stimulus selection procedure favors sampling a specific response probability quantile: that is, choosing stimulus parameters that will yield a particular p_r. This is done in many staircasing procedures transformed or weighted to converge to a specific quantile (Kaernbach, 2001). If the data are being used to constrain the slope of the psychometric function, multiple response probabilities must be tracked, much as in multiple staircase procedures. For the case of two-dimensional estimation, only one y value will produce a particular probability of response at a given x value. This y value can be found by adopting the current best estimate of the parameters Θ and S; computing y_c for the chosen x value from the threshold function and the current best estimate of Θ, and then computing the inverse of the psychometric function to find the y value at quantile p_r. But it is not logical to condition one's estimates of one parameter on definite assumed values for other parameters that are in fact uncertain. So in FAST, we instead combine all possible values in the parameter space in proportion to their posterior probability, thus avoiding premature commitment to a currently plausible estimate of the model parameters.

This space of placement strategies yields several plausible combinations: global and local entropy minimization, global and local minimization of predictive variance, and local quantile sampling. These alternatives are implemented in FAST, and their relative advantages are discussed in later sections.

5. Implementation

5.1. Estimating the Posterior Likelihoods

Functional adaptive sequential testing is designed to be a general procedure for estimating threshold functions. The estimation approach we have described may be implemented in a number of different ways, each with certain advantages, and each with certain drawbacks. A number of probabilistic sampling methods are guaranteed to converge to the true posterior probability defined over the entire space of Θ and S (e.g., MCMC; Kuss *et al.*, 2005); however, these methods may take thousands of iterations to converge. For adaptive testing in psychophysical experiments, the posterior over Θ and S should ideally be recomputed before each trial so that the stimulus for the next trial can be chosen most appropriately, thus MCMC strategies tend to be prohibitively slow.

An alternative to running a full MCMC search is to employ sequential monte carlo (particle filtering). This method will also converge on the true posterior, and is efficient for online estimation problems. However, the optimal proposal distribution for the essential MCMC rejuvenation step (Kujala and Lukka, 2006) is usually

specific to particular parameters, and as such, this step may be ineffective or slow in the generic case. Thus, to make this computation speedy for any threshold function, we opted to do a simple grid search (Kontsevich and Tyler, 1999; Lesmes *et al*., 2006).

At initialization of the FAST structure, we define a lattice over parameter space. Each parameter is sampled within user-defined bounds (either logarithmically or linearly), with a particular uniform sampling density (this sampling density varies depending on the number of parameters, since for each added parameter, the total size of the parameter grid is multiplied by the number of sampled values).

Each point in the lattice defined over parameter space corresponds to a certain vector of values of psychophysical function parameters (Θ) and the psychometric slope parameter (S). Each point is also initialized with a prior probability (which may be informative, uninformative, or uniform, as deemed appropriate). After each trial, the log-posterior probability of each parameter lattice point is updated by adding the log of the likelihood of the new response under the parameter values represented by that lattice point. Thus, the un-normalized log-posterior is computed after each trial.

5.2. Dynamic Lattice Resampling

With only a finite number of points in the parameter lattice, there is an unavoidable tradeoff between the range and density of sampled parameters. From one experiment to the next, one might opt for a larger range, or a finer sampling grain, but inevitably, a situation will arise wherein both a large range and a fine sampling grain are required. In such situations, if one samples too narrow a range (but with high enough density) one may find that the maximum *a posteriori* parameter values are not contained in the parameter lattice, and testing will be inefficient, if not useless; similarly, if one samples too coarsely (but with a sufficiently wide range), steps between lattice points may be too large for efficient testing.

Although Lesmes *et al*. (2006) reported that the sampling density over parameter space did not substantially alter the efficiency of the qTvC algorithm, we find that in cases of more extreme undersampling these problems do arise. In the qTvC case, such problems can, for the most part, be avoided due to the small number of parameters required for the linear TvC function. However, for functions like the CSF, where the total number of parameters (including slope) is 5 or higher, the number of lattice points rises with n^5 (or even steeper), where n is the number of sampled points along each parameter. In such cases, the computational costs associated with high sampling densities become prohibitive.

Since FAST is designed to operate in the general case, with any threshold function, this is a considerable problem. Our solution is to adopt dynamic resampling, or a ‘roving range’. Once some considerable amount of data has been gathered (enough to provide a reasonable estimate of the probability surface over the current parameter lattice), one may resample the parameter lattice such that it either shifts or shrinks appropriately to converge on the global maximum.

Specifically, we obtain an estimate of the parameter values, and uncertainty over those values (the marginal mean and standard deviation; see next section). Using these estimates, we pick a new range that is centered on the marginal mean, and spans several standard deviations (2 is used in the subsequent simulations) on either side of the mean. With this method, an initial coarse range will shrink as additional information is gleaned about the underlying parameter values — similarly, an initially misplaced parameter range will shift to higher probability regions of parameter space.

Although the details of this algorithm are rather unprincipled, we find that in practice, dynamic resampling is quite helpful (see simulations in later sections).

5.3. Estimating Parameter Values

From the computed joint posterior probability density (pdf) over the parameter lattice, there are several possible ways to derive a point estimate and corresponding confidence measure for each parameter value. We focus mainly on the 'marginal' probability density function, in which the N-dimensional pdf in the space of parameter values is integrated over all parameters except the one of interest, to yield 1-dimensional pdfs for each parameter in turn (this may be loosely thought of as the profile of the N-dimensional joint probability density as projected onto the axis of interest). The peak of the marginal pdf does not generally agree with the peak of the N-dimensional pdf. The latter estimate, from the 'maximum *a posteriori*' (MAP) point in the N-dimensional space of parameter values, would be appropriate if we knew that the values of all other parameters were the ones associated with the MAP point. But in reality, of course, the values of all parameters are uncertain, and the probability of any given value for a single parameter of interest has to include all the cases generated by variation in the other parameters. This is what the marginal pdf represents (illustrated in Fig. 4).

In specifying the confidence interval or standard error for the estimates, similar considerations arise. The most likely value for one parameter generally depends, often strongly, on the values assumed for the other parameters. This creates a ridge of high probability in parameter space. The standard error for one parameter is often calculated on the assumption that the parameter of interest is varied in isolation, with other parameters pinned at fixed values. The resulting error measure reflects the width of that ridge (usually, the cross-section through the MAP point in the relevant direction). But when the concurrent uncertainty about other parameters, as reflected in the N-dimensional pdf, is taken into account, the appropriate 'integrated' uncertainty measure is determined from the marginal pdf. FAST reports this 'integrated standard error', which is commonly much greater than the 'isolated' standard error obtained from the cross-section through the MAP point.

In practice, FAST samples the continuous marginal pdf for any parameter at lattice points equally spaced over a finite range. The lattice can be recentered and rescaled by resampling, as described above, but the estimate of the mean and standard error of the pdf are always prone to distortion by truncation and sampling error.

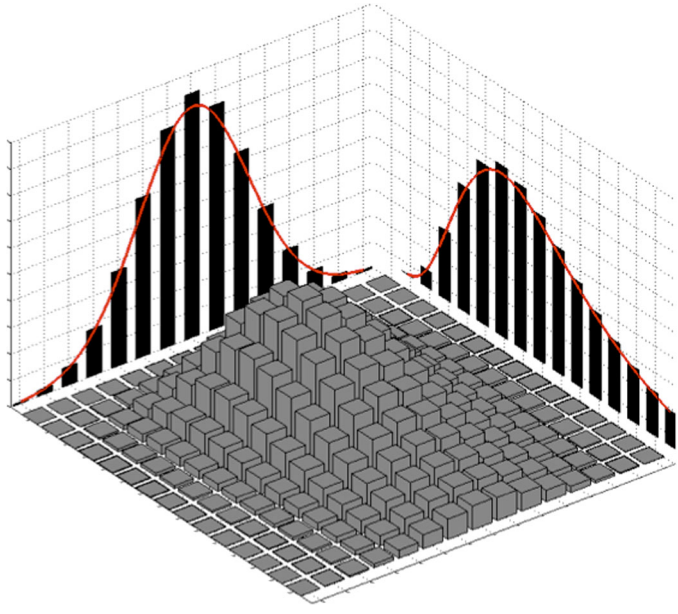

Figure 4. Parameter estimation procedure. We begin with a multidimensional probability density function (pdf) sampled at a few discrete lattice points (grey bars). By summing over all other parameters, we obtain a marginal probability distribution for a parameter of our choosing (two sets of black bars correspond to these marginal distributions). We then fit a Gaussian distribution to these marginals to obtain an estimate of the marginal mean and standard deviation that is less susceptible to truncation (red lines). This figure is published in colour on http://brill.publisher.ingentaconnect.com/content/vsp/spv

Notably, when the parameter lattice is too coarse and concentrates the probability mass at one sample point, the mean regresses toward that point and the standard deviation based on the sampled points is an underestimate of the standard deviation of the complete pdf. Both in FAST's intermediate computations and in its final estimates, these problems are addressed by fitting a Gaussian to the lattice point probabilities (see Fig. 4), and using the mean and standard deviation of the fitted Gaussian, rather than the sample values, to represent the mean and standard deviation of the continuous pdf.

By focusing on the 1-dimensional pdfs during intermediate computations, FAST initially neglects the correlational structure of the N-dimensional pdf. That structure might seem important not only for a complete characterization of the uncertainty in parameter values, but also for directing the resampling process described above along the ridge of high probability in parameter space. The resampling procedure in FAST addresses this by always recentering the sample points for each parameter in turn. In this way the sequence of N 1-dimensional translations moves the lattice efficiently along the ridge.

The Gaussian description may be poor in cases where the contours of the posterior probability function in parameter space deviate markedly from symmetry

around the peak, as commonly happens when the slope parameter S of the psychometric function must be estimated from the data). In these cases the mean of the Gaussian fit to the maringal pdf differs from the mode of the marginal pdf as well as (generally) from the MAP. But even here the mean remains the optimum estimate in the least squares sense, and minimizes the cost of errors if cost increases as the square of the error (see Note 4).

5.4. Choosing Stimuli

Choosing informative stimuli by minimizing the entropy of the expected posterior probability distribution requires that the expected posterior probabilities be evaluated for every stimulus of interest. This is impossible in practice, and numerical approximations must be used.

To carry out these calculations for global stimulus placement strategies, we define conditional probability look-up tables for a pre-defined range of stimulus values. The conditional probability look-up table specifies the probability of a positive or of a negative response under each set of parameters in the lattice for a given stimulus. These conditional probabilities are the factors by which the relative likelihood of each set of parameter values is changed when the response is known. Using this look-up table, one can compute, for any proposed stimulus placement, the expected posterior entropy for a trial with that stimulus placement, and choose the stimulus placement that minimizes this value. In this manner, the expected posterior probability over the space of x and y stimulus values may be computed with reasonable efficiency. To circumvent the loss of resolution associated with discrete sampling of the stimulus space in the pre-defined look-up table, we also implement quadratic interpolation to estimate the optimal stimulus values, as we did for the interpolated MAP estimator.

Carrying out the conditional probability calculations for local stimulus placement is easier because fixing x greatly reduces the number of alternative stimulus placements that need to be considered. Thus, the optimal stimulus strength y for a particular x value is a relatively fast computation that can usually be made on a trial-by-trial basis without constructing a conditional probability look-up table for all x values in advance. Our implementation of this search is recursive: we coarsely sample a range of y values, and compute the expected posterior entropy for each. We then choose the y value that minimizes this quantity, and resample more finely around this y value. A few iterations quickly produce a fine-grained optimal estimate that is typically much faster than finely sampling the full initial range of y values.

6. Simulation Results

In this section we present computer simulations that illustrate the efficiency and flexibility of FAST. In these simulations, all parameter estimates are obtained using the mean obtained from a Gaussian fit to the marginal pdf (see section on Estimation

for details and *a priori* justification; see Fig. 9 and accompanying discussion for empirical justification).

We evaluate the parameter estimates obtained from a particular simulation in terms of their error:

$$\mathrm{MSE}(\Theta_j) = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\Theta}_{i,j} - \Theta_{i,j})^2, \tag{6}$$

where Θ is the true parameter vector, $\widehat{\Theta}$ is the parameter estimate, i is experiment number, and j is the parameter number. Thus, error $\mathrm{MSE}(\Theta_j)$ is the mean squared deviation, to which bias and variance both contribute.

6.1. Efficiency in an Extreme Case

The FAST algorithm is most effective when estimating functions with many sampled points. It was developed for the purpose of estimating time-courses of after-effects — a scenario in which the gains from functional estimation over isolated estimation are most dramatic.

We compare adaptive functional testing to adaptive testing at several discrete x values without assuming an underlying threshold function — 'the method of 1000 staircases' (Mollon *et al.*, 1987; or more precisely, 100 staircases to simulate a sensible time-course experiment, Vul *et al.*, 2008). Each independent staircase tests adaptively according to our implementation of the Psi testing algorithm described by Kontsevich and Tyler (1999).

For these simulations, we adopt a logistic psychometric function, as would be natural to use in a matching task, and assume that the underlying threshold function is an exponential decay:

$$y_c = \Theta_1 + (\Theta_2 - \Theta_1)(1 - \exp(-x/\Theta_3)), \tag{7}$$

where Θ_1 is the initial (pre-adaptation) threshold value; Θ_2 is the saturation point; and Θ_3 is the time-constant.

We compare the error in estimating these parameters over 1000 trials when those trials are sampled according to 100 independent staircases (at each of 100 time-points), to estimation over those same time-points using functional adaptive testing, where the stimulus placement at each point is informed by all other points. For each of the 250 simulated experiments, we randomly choose new parameter values for Θ and S. Θ_1 is drawn uniformly from the range of -4 to 4; Θ_2 is drawn uniformly from the range of -15 to 15; Θ_3 is drawn uniformly from the range of 1 to 50; and S is fixed to 0.4. The prior probabilities over these values for the estimation are: Θ_1: normal with $\mu = 0, \sigma = 3$; $\Theta_2 : \mu = 0, \sigma = 8$; Θ_3: $\log_{10}$ normal with $\mu = 0.75, \sigma = 0.5$; S: $\log_{10}$ normal with $\mu = -1, \sigma = 0.8$ (the parameters that have log-normal priors are ones that have a $[0, \infty)$ support).

Figure 5 shows the results of our simulations: Functional testing confers a substantial early advantage over independent testing at different points along the function, and this initial advantage is sustained over 1000 trials for nearly all (local)

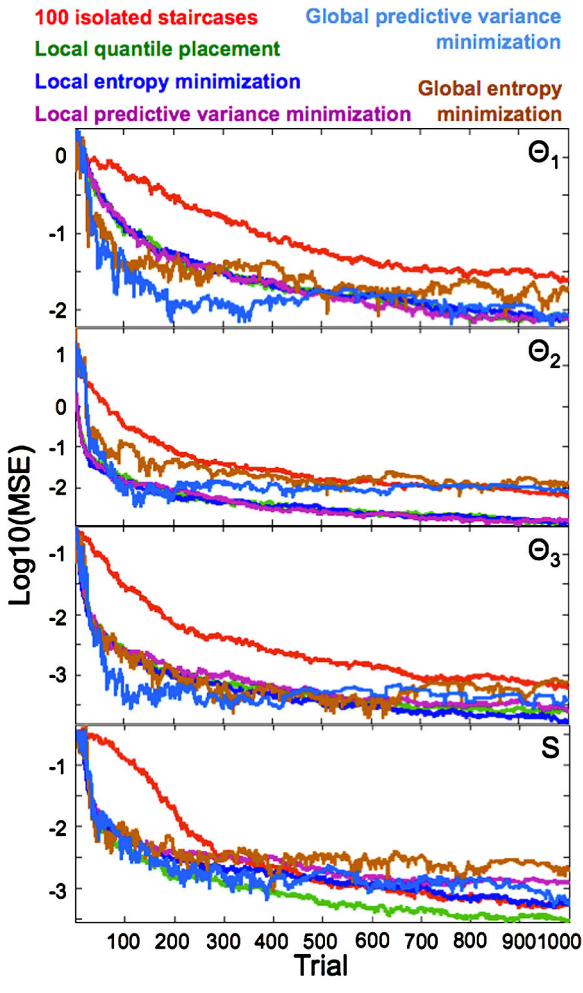

Figure 5. The mean squared error as a function of the number of trials for different stimulus placement strategies estimating an exponential decay curve (as often used to describe the time-course of aftereffects). The four panels represent different parameters, from top to bottom: initial null-setting value (at $\Theta = 0$), saturated null-setting value, time-constant, and logistic slope parameter. Mean squared error decreases substantially faster under all functional sampling schemes compared to independent adaptive estimation procedures at each timepoint (of 100), and this initial advantage persists throughout the course of many trials. This figure is published in colour on http://brill.publisher.ingentaconnect.com/content/vsp/spv

functional testing strategies outperform independent testing throughout the duration of the 'experiment'; thus validating the basic premise of FAST: substantial testing and estimation efficiency may be gained by explicitly considering the function relating variation in thresholds to an independent variable.

6.2. Efficiency in a Less Extreme Case (CSF)

The previous example was designed to illustrate an extreme situation in which one seeks to obtain threshold estimates at 100 different points along a function — in this

case, running 100 independent staircases is grossly inefficient compared to sampling those 100 points while considering the underlying function. However, few psychophysical experiments aim to assess a function with that level of precision; often, a much smaller set of points is used. In this section, we aim to both illustrate the generality of functional adaptive testing by applying it to the contrast sensitivity function, and also to quantify the reduced benefit of functional adaptive testing over independent testing when fewer points are considered.

In this case we adopt a logistic psychometric function with a 4AFC detection task (mimicking the psychophysical data presented later in this paper), and assume that the underlying threshold function is the log-parabola contrast sensitivity function from equation (3) and Figs 1–3. We compare the error in estimating the three parameters of this function over 800 trials when those trials are sampled according to 14 independent staircases (at each of 14 spatial frequencies, equally spaced on a logarithmic scale), to estimation over those same spatial frequencies where the stimulus placement at each point is informed by all other points *via* the commonly estimated function. For each of the 100 simulated experiments, we randomly choose new parameter values for Θ and S. Θ_1 is drawn uniformly from the range of -4 to -1; Θ_2 is drawn uniformly from the range of -1 to 1.5; Θ_3 is drawn uniformly from the range of -1.5 to 0.5; and S is drawn from the range of 0.01 to 0.4. The prior probabilities over these values for the estimation are: Θ_1: normal with $\mu = -2, \sigma = 2$; $\Theta_2 : \mu = -1, \sigma = 2$; Θ_3: normal with $\mu = -1, \sigma = 0.3$; S: $\log_{10}$ normal with $\mu = -1, \sigma = 1$.

Figure 6 reveals the predicted effect — with only 14 points being independently tested, the advantage of functional testing is much less dramatic than with 100 points. Nonetheless, a considerable advantage remains and lasts over the whole range of trials. Again, as can be seen in the results for the slope parameter, the optimality criterion has an important effect, wherein stimulus placement *via* posterior predictive variance minimization is particularly poor at estimating the slope parameter — this is to be expected, as the variance of the predicted threshold is completely independent of the slope, so stimuli are not chosen to estimate this parameter.

6.3. Dynamic Lattice Resampling

Although Lesmes *et al*. (2006) found no substantial effects of decreasing the parameter sampling grain for the qTvC method, all of the sampling grains tested were rather fine. This is practicable for the qTvC case, wherein there are only three parameters to be estimated. However, in the general case, where one might aspire to test different functions, with potentially higher-dimensional parameter spaces, the sampling density becomes an issue (indeed an exponential issue, as the total number of lattice points increases with the power of the number of dimensions). Thus we tested the effects of rather dramatic under-sampling of parameter space: 5 lattice points per parameter.

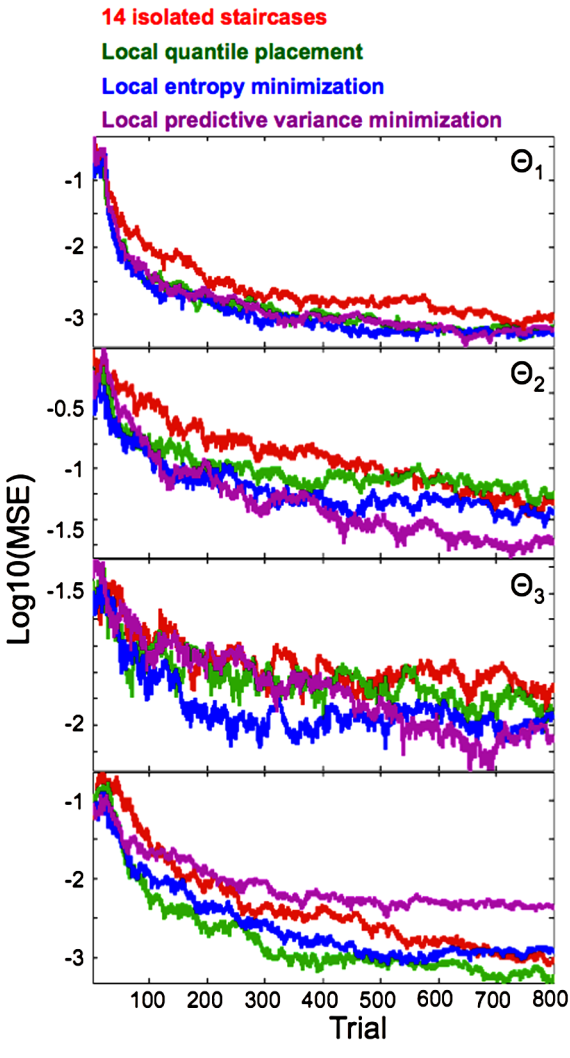

Figure 6. The mean squared error as a function of the number of trials for different stimulus placement strategies estimating a contrast sensitivity function. The four panels represent different parameters (see equation (3)). Here, the advantage of functional estimation over independent is smaller because fewer independent staircases are used (14). Moreover, a marked detriment is evident for the precision of estimating the slope parameter for predictive variance minimaztion. This figure is published in colour on http://brill.publisher.ingentaconnect.com/content/vsp/spv

Figure 7 shows the perhaps intuitive results of under-sampling: lower sampling density effectively caps the accuracy of parameter estimates. A coarser parameter sampling grain results in a higher asymptotic RMSE for all parameters. However, when we allow dynamic resampling of the 5^3 lattice (the 'roving range' we implemented), this problem is substantially reduced: we achieve much lower asymptotic errors by dynamically shrinking and resampling the (under-sampled) parameter lattice.

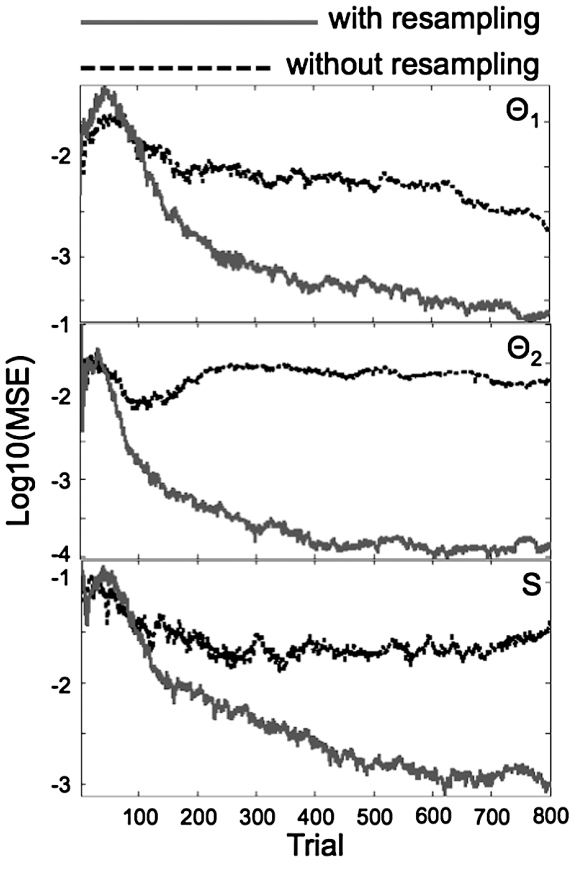

Figure 7. A 'roving range' (dynamic lattice resampling) mitigates the problems that arise from sparse sampling of the parameter lattice — by shifting and shrinking the lattice, we can achieve greater accuracy.

6.4. Effects of Assuming an Incorrect Function

Here we consider the effects of assuming an incorrect model of the data. Although global entropy minimization is theoretically most efficient for estimating the parameters of the assumed model, the data gathered by global stimulus placement may not be as effective at constraining the parameters of another model (which may turn out to be the reality underlying the data). To test this idea, we run 250 experiments, each with 800 trials, where we attempt to fit the correct (non-linear) TvC function to data that were gathered on the assumption of a linear TvC function, using two sampling schemes: (1) *global* entropy minimization, and (2) *local* entropy minimization.

Specifically, we conduct simulations in which the real model is a non-linear TvC function:

$$y_c = \begin{cases} \Theta_2 \left(\dfrac{x}{\Theta_1} \right)^{\Theta_3}, & \text{for } x > \Theta_1, \\ \Theta_2, & \text{for } x \leqslant \Theta_1, \end{cases} \tag{8}$$

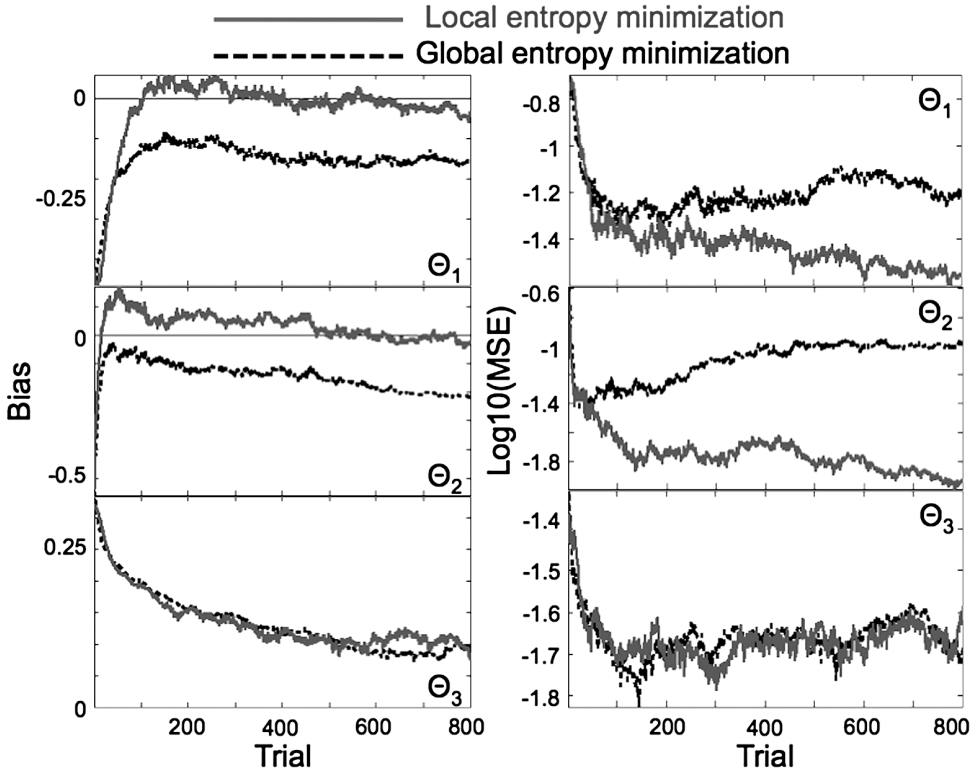

Figure 8. When stimulus sampling is done assuming a linear TvC function (equation (9)), but in fact, the correct model is a non-linear TvC curve (equation (8)), global stimulus placement schemes confer a substantial detriment, as seen in the final mean squared error when fitting the final, correct, function.

while functional testing assumes a simpler, linear TvC function:

$$y_c = \begin{cases} \Theta_2 \dfrac{x}{\Theta_1}, & \text{for } x > \Theta_1, \\ \Theta_2, & \text{for } x \leqslant \Theta_1. \end{cases} \tag{9}$$

Figure 8 shows that data gathered with the goal of *global* entropy minimization are far less effective at constraining parameters of an alternate (and only slightly different) model, yielding substantial bias, and greatly inflated mean squared error over parameter values estimated using *local* entropy minimization. For this reason, we advocate using fixed or random sampling of x values, and local, functional, stimulus selection.

6.5. Validation of Estimator and Estimation Procedure

Thus far, we have presented data describing the efficiency of different adaptive sampling schemes while the estimates themselves were obtained using a somewhat rarely adopted estimation procedure: fitting a Gaussian probability density function to the marginals of each parameter, and using the resulting mean and standard de-

viation. In this section, we demonstrate the validity of this estimator by empirically calibrating the confidence intervals obtained from it. We ask: does a $q\%$ confidence interval around the mean of this Gaussian have $q\%$ probability of containing the true value (in simulations)? This can be measured in two ways: Z-quantiles over time, and quantile–quantile plots.

To obtain Z-quantiles over time, for each simulated experiment, trial, and parameter, we compute the mean and standard deviation of the current Gaussian fit to the marginals. We compute $Z_{\text{err}} = (\Theta - \hat{\mu}_\Theta)/\hat{\sigma}_\Theta$, where Z_{err} is the Z-score distance of the true parameter value Θ, from the Gaussian mean ($\hat{\mu}_\Theta$) and standard deviation ($\hat{\sigma}_\Theta$ estimated for that parameter). By computing this quantity for each trial, each experiment, and each parameter, we obtain a large number of Z-scores. To check the calibration of the Gaussian fit to reality, we then compute the Z-quantiles, asking: what is the Z-score threshold such that $q\%$ of all computed Z_{err} values are below it? If the procedure is calibrated, we should expect these quantiles to remain constant over an experiment (despite the fluctuations in mean squared error seen in Fig. 5), and to correspond to the theoretically correct quantiles: $Z_{\text{thresh}}(q) = G^{-1}_{\text{CDF}}(q, \mu = 0, \sigma = 1)$, where G^{-1}_{CDF} is the inverse of the Gaussian cumulative density function. For example, $q = 0.025$ yields the familiar $Z_{\text{thresh}}(q) = -1.96$. Figure 9 shows that this plot for the exponential data from Fig. 5 reflects an ideal, stable calibration (similar results obtain for other simulations).

A quantile–quantile (QQ) plot is a fine-grained summary of the results presented above. To compute the QQ plot in Fig. 9, we calculate the Z-score boundary of

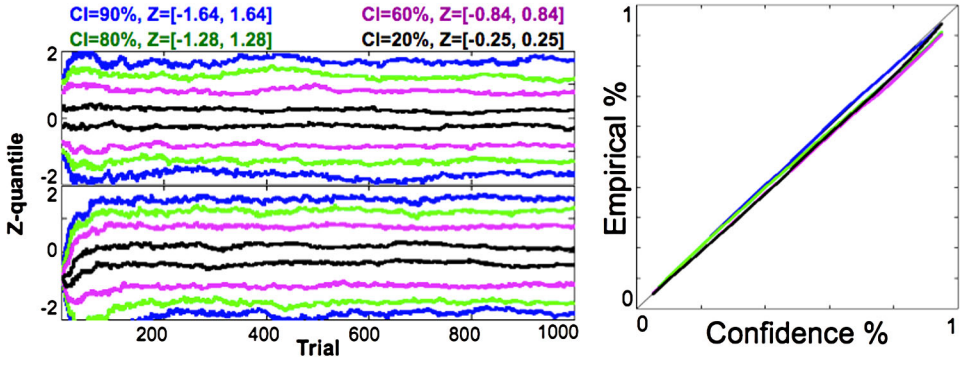

Figure 9. (Left panel) Z-quantile calibration, here we show a representative sample of two parameters (the initial setting in an exponential decay, and the slope parameter from an exponential decay — see equation (7)). Z-score quantiles of parameter errors are stable over time, and near their theoretical values. (Right panel) This calibration of the Gaussian fit to the marginal distribution can be seen in the QQ plot, where the confidence interval percentile predicts with near perfect identity the empirical quantile (e.g., 95% of true parameter values lie within a 95% confidence interval). Different lines here correspond to different parameters of the exponential, all are near the unity line indicating that our estimator is generally well calibrated. This figure is published in colour on http://brill.publisher.ingentaconnect.com/content/vsp/spv

a $q\%$ confidence interval around the mean: $CI = G^{-1}_{\mathrm{CDF}}(0.5 \pm q/2, 0, 1)$. Then we compute the proportion of trials in which the true parameter value lies within this confidence range p. Figure 9 plots p as a function of q. A perfectly calibrated plot will yield a perfect diagonal with unity slope, such that every $q\%$ interval contains $p\%$ of the true values. Our data are well aligned with this ideal, showing none of the over-confidence (a line below the diagonal) of truncated marginal estimators, or much of a miscalibration due to the true posterior not being Gaussian (often revealed as a crossing of the diagonal).

7. Psychophysical Results

FAST has been used to advantage in our investigations of topics including the time course of the generation and decay of afterimages; the dynamics of contrast gain control; and contrast sensitivity functions. We next report an experiment designed specifically to provide a comparison between the efficiency of FAST and other methods. This was a CSF measurement in which trials generated by FAST are interleaved with trials generated by a similar search algorithm operating independently at each spatial frequency.

We used a four-alternative forced choice procedure to measure contrast sensitivity as a function of spatial frequency. The stimulus was an annulus displayed for 200 ms. On each trial, one quadrant, randomly selected, contained a radial sinusoidal contrast grating, while the other three quadrants were of uniform luminance. Subjects were asked to indicate which quadrant was nonuniform. The luminance of each quadrant was independently offset by a small random amount each trial, to prevent detection of the nonuniform quadrant through a difference in average luminance (despite a gamma-corrected display). Eight spatial frequencies were presented in a cycle, and an equal number of trials were presented at each spatial frequency.

To compare the performance of FAST with the performance of independent staircases, we used two conditions, which ran simultaneously with trials randomly interleaved. The difference between the two conditions was in the method used to select the level of contrast for the stimulus grating. In both cases, the contrast level was chosen so as to be detectable at a rate between 0.65 and 0.85, based on data from previous trials. In the independent condition, the detection threshold was estimated for each spatial frequency based only on data collected from that spatial frequency. In the CSF condition, FAST estimated the detection threshold based on data previously collected from all spatial frequencies, under the assumption that the contrast sensitivity function would fit the form of a log parabola CSF (Pelli *et al.*, 1986). The log parabola CSF model fits a parabola to the logarithm of the spatial frequency (equation (3)), such that Θ_1 determines the level of contrast sensitivity at the peak (shifting the parabola up and down); Θ_2 determines at what spatial frequency this peak occurs (shifting the parabola left and right); and Θ_3 determines the rate at which sensitivity drops off (stretching the parabola horizontally). Data were

only used to inform the adaptive testing procedure in the same condition, so the CSF and independent conditions remained independent in their parameter estimates and choice of stimuli throughout the experiment.

After data were collected, they were pooled to calculate a final estimate of contrast sensitivity, both independently at each spatial frequency, and using the log parabola CSF (Fig. 10). We used the final *independent* estimates of contrast threshold at at spatial frequency x (y_x^*) as 'ground truth' to evaluate the error. Mean squared error for each trial, for each condition was computed using equation (6). The estimated contrast threshold, $\hat{y}_x^*$, for a given spatial frequency was calculated using the same method by which data were collected for that condition: For the independent condition, each spatial frequency's threshold was calculated independently, and for the CSF condition FAST estimated the threshold using the log-parabola CSF. For both conditions, the estimate at trial t was compared against the final independent estimate y_x^* based on all data from every condition at a given spatial frequency.

In the first 100 or 200 trials, the FAST condition obtains a more accurate estimate than the independent condition due to the advantages of pooling information across spatial frequencies and more effective stimulus placement. Thus, as our simulations suggested, FAST can be used to expedite the estimation of contrast sensitivity functions, even when a small number of spatial frequencies are tested (for a thorough treatment of expedited CSF estimation through functional adaptive testing, see Lesmes *et al*., 2010). However, after about 200 trials, the precision of the independent threshold estimates becomes sufficient to uncover systematic deviations from the model. If the model cannot fit the data perfectly (likely, most real-world cases), the error of a model-based estimate will have a nonzero asymptote because even after the parameters reach their optimal values, there will still be a difference between the model and reality. The independent estimate, providing more degrees of freedom, can be expected to perform better in the long run, and does so here; but, of course, data collected efficiently using FAST may later be fit with independent thresholds to take advantage of these additional degrees of freedom.

8. Discussion

Since Fechner (1860) suggested that perception can be measured against a physical standard, researchers have been estimating psychophysical thresholds and match points across the gamut of perceptual domains. As the sophistication of psychophysical experiments has increased, so too has the demand for efficient measurement of thresholds, and experimental methods have been gradually improving to meet this demand. However, as of yet, no general two-dimensional threshold estimation procedure has been proposed. We described a general and efficient algorithm for estimating two-dimensional response probability surfaces, and provide a Matlab implementation of the algorithm. While the FAST toolbox provides a variety of alternative strategies for stimulus placement and parameter estimation, here

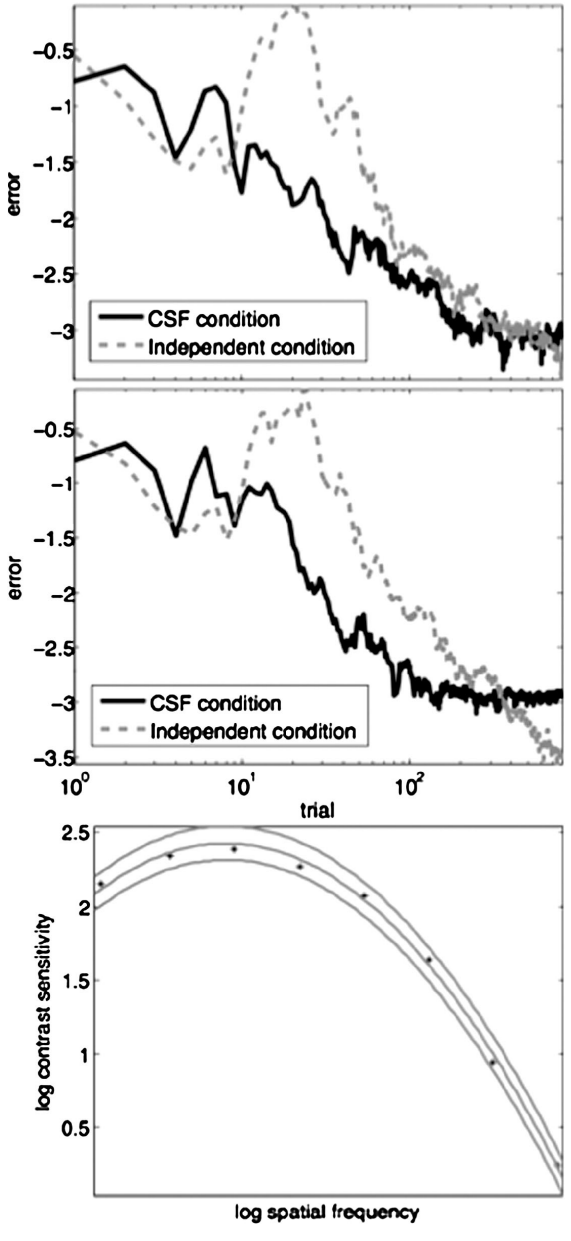

Figure 10. (Top panel) Mean squared error (log scale) for estimated threshold values as a function of the number of trials (log scale). Predictions using isolated staircases are less accurate as compared to functional adaptive testing. (See text for details.) (Center panel) Simulations using the final threshold estimates obtained from the psychophysical data and the same simulated testing procedure — these results indicate that given identical model mis-specification, our idealized simulations predict gains similar to those seen for real human observers. (Bottom panel) Final model fit to the data (lines correspond to different response probability quantiles — 40% 60% and 80%) and final independent threshold estimates of the 62.5% quantile (asterisks), in this case, the simple log-parabola CSF function appears to be a satisfactory (albeit imperfect) model of the data.

we want to briefly discuss the strategies that seem best, and potentially useful future extensions of this work.

8.1. Assumptions of Functional Adaptive Sequential Testing

As mentioned throughout the paper, the functional adaptive testing formulation does make several important assumptions, which may yield imprecise, and perhaps misleading, results if violated.

First, like nearly all psychophysical testing procedures and experimental designs, our formulation assumes that experimental trials are exchangeable — in other words, they are invariant to order. This assumption is crucial to essentially all cognitive and psychophysical experiments, but may often be violated due to a temporal correlation of errors (e.g., from phasic internal states) or learning effects. For basic threshold estimation, FAST may be used to account for online learning by modeling the threshold as an exponential learning curve, thus allowing for some robustness to these effects; however, if FAST is used to estimate a threshold function, like a CSF, learning effects will constitute unaccounted for variability. This is not a serious problem — in the worst case scenario such learning or temporal correlation effects will increase response variance and thus reduce the efficiency when testing human participants, but this loss of efficiency will apply to any common testing procedure.

Second and more critically, we must assume that the psychometric function is translation invariant in both stimulus strength and the independent variable of interest; in other words a single slope parameter captures the relationship between increasing stimulus strength and probability of positive response regardless of variations of the independent variable (e.g., spatial frequency) and critical value (e.g., 75% discrimination contrast threshold). Within the apparently wide jurisdiction of 'Crozier's Law' of proportionality between threshold stimulus magnitudes and their standard deviations (Treisman, 1965), the translation assumption holds provided that the psychometric function is expressed in terms of the logarithm of the stimulus magnitude. With that caveat, the translation assumption appears to be approximately valid in the literature (Watson and Pelli, 1983), but it may well be violated in particular psychophysical or cognitive paradigms.

Third, FAST requires an assumption of a particular parametric form of the threshold function — this assumption will generally be insecure and at best inexact (for instance, will delay discounting follow an exponential or a power law?) For this reason, we prefer local to global stimulus placement (see next section) thus trading off some theoretical efficiency in favor of robustness to an incorrect assumed functional form. We stress that FAST is not rendered useless by failure of the theoretical form assumed for the threshold function. On the contrary, it provides an efficient method for detecting small deviations from the assumed theoretical form. But the efficiencey of FAST for that purpose, and for threshold estimation in general, declines when the deviations of the theoretical model from reality exceed the current threshold uncerainty. This limitation is illustrated by the results of Fig. 10

(top panel) after large numbers of trials. We have also found that FAST can efficiently distinguish between models of different form: multiple FAST algorithms can be run concurrently, each using a different candidate threshold function: for each trial, a threshold function is randomly selected, and stimuli are chosen based on that threshold function, while the data are used to update the parameters of all functions. This procedure yields some confidence that no threshold function will benefit from a bias due to adaptive testing.

8.2. Global or Local Stimulus Placement?

Theoretically and empirically, selecting the point in stimulus space that is globally most informative generally results in faster convergence to the true function parameters than alternative methods. Indeed, so long as the psychometric function is monotonic, no disadvantages are incurred by choosing the optimal stimulus parameters in the unidimensional threshold estimation case. However, for two-dimensional estimation — particularly if the variation of response probability along one dimension is not monotonic — the stimulus that is globally most informative in theory may not be most informative in practice.

When employing global optimization for stimulus placement, there is no guarantee that the stimulus space will be sampled in an intuitively informative manner. Once a particular form of the threshold function (and hence, the two-dimensional probability surface) is assumed, there will often be particular locations in stimulus space that are always most informative about the parameters of the assumed function as the experiment progresses. For example, if one assumes that an aftereffect decays to zero, and attempts only to estimate the initial aftereffect strength and the rate of decay, the points that would constrain these two parameters best are in the initial period of rapid decay, and there is no need to sample stimuli when the aftereffect has decayed to the (assumed) limit of zero. In all such cases, stimulus selection based on global entropy minimization will choose trials isolated to a few particular regions of the stimulus space. This is optimal if one is correct about the form of the threshold function. However, the assumed functional form may not be accurate, for example, an aftereffect may not actually decay to zero, but may have a persistent component (Vul *et al.*, 2008). In this case, the stimulus parameters that seemed to be most informative under incorrect assumptions about the underlying function, will actually fail to inform about the deviation. Due to the possibility of these cases, it is generally useful to sample the stimulus space more uniformly, so as to detect systematic deviations from the assumed functional form. One may refuse on these grounds to entrust the sampling of the stimulus x axis to an entropy minimization algorithm, and instead select in advance the x values to sample, in the standard way. Another alternative is to sample the x axis randomly.

Another alternative (discussed in the 'Assumptions' section) is to choose stimuli based on multiple candidate threshold functions intermittently. If the goal of testing is to explicitly distinguish different threshold models, one might wish not to just optimize stimulus placement for the estimation of parameters in a specific model,

but instead to optimize stimulus placement for selection among alternative models and functional forms. This approach has been used to design experimental protocols in advance of experimentation with some success (Myung and Pitt, 2009), but not in online adaptive testing — this is a promising direction for future research. Nevertheless, if adaptively testing with the goal of model selection, it would not be possible to specify all possible alternative models, and the tradeoff between efficiency and robustness entailed in global *vs* local stimulus placement may still fall on the side of robustness.

8.3. What Optimality Criterion to Use?

We have considered several criteria for 'optimal' stimulus placement. According to information theory, the ideal is minimizing expected posterior entropy — when this is optimized one chooses the theoretically most informative stimulus. However, the goals of typical psychophysical experiments are not to minimize entropy over the joint probability distribution over all parameters, but rather to minimize expected variance (that is, decrease standard error and confidence intervals) of some or all parameters, or else of the threshold itself. In the latter case there is no principled way to combine variances across parameters at different scales. In future work it may be useful to consider explicit costs of errors in different parameters. This way, stimulus placement in an experiment primarily interested in the estimation of the peak frequency of the contrast sensitivity function would optimize this goal, while an experiment interested in the steepness of the sensitivity drop-off with increasing spatial frequencies could best estimate that parameter. Although such explicit formulations of the loss function may yield the most useful experiment, the efficiency gained may not outweigh the added complexity.

8.4. Estimation of Non-parametric Functions

Although in many psychophysical experiments, there is much to be gained from conducting functional adaptive testing with simple parametric functions, in some cases the parametric function is unknown, or very complex. In these cases, one may want to consider functions with many parameters, or even non-parametric estimation procedures like locally weighted regression. These might provide a basis for an alternative, multi-dimensional adaptive search technique requiring less strong assumptions, but to our knowledge such approaches has not yet been explored for adaptive testing.

9. Summary

We introduced and validated a method for efficient estimation of two-dimensional psychophysical functions and a Matlab toolbox implementation of it: Functional Adaptive Sequential Testing (see Appendix A). This toolbox (and corresponding manual) may be downloaded here: http://code.google.com/p/fast-toolbox/

This tool substantially improves efficiency of psychophysical testing under correct (or approximately correct) assumptions about the parametric model of the

function relating threshold (e.g., contrast) to some independent variable of interest (e.g., spatial frequency). Specifically, so long as the model error is smaller than the uncertainty inherent in the data, FAST will remain a more efficient alternative than running multiple independent staircases. However, when sufficient data are gathered such that they provide threshold estimates with precision exceeding the model error, then FAST will lead to less efficient estimation, and independent staircases should be used instead.

Acknowledgement

This work was supported by a Russell Sage, NDSEG, an NSF Dissertation grant (EV), and National Institutes of Health Grant EY01711 (DM).

Notes

1. The method of adjustment tends to be quite efficient, but is often disfavored in practice because making a given setting takes a substantial amount of time, so it is inapplicable to quickly decaying effects, and may be susceptible to 'overthinking' on the part of the participants.
2. We do not discuss here current developments in non-parametric adaptive testing procedures, which improve upon classic staircase methods without invoking an explicit psychometric function (Faes *et al*., 2007).
3. Each point is not equally informative, and thus the testing procedure should be adaptive, to choose the most informative points.
4. To see this, recall that the mean squared value is the variance plus the squared mean, for any distribution; if the mode is adopted as an estimate of a random variable, the distribution of error in the estimate is the distribution of deviations from the mode, and the mean squared error is increased by the squared difference between mean and mode.

References

Cobo-Lewis, A. B. (1997). An adaptive psychophysical method for subject classification, *Percept. Psychophys.* **59**, 989–1003.

Cornsweet, T. and Teller, D. (1965). Relation of increment thresholds to brightness and luminance, *J. Optic. Soc. Amer.* **55**, 1303–1308.

Dixon, W. and Mood, A. (1948). A method for obtaining and analyzing sensitivity data, *J. Amer. Statist. Assoc.* **43**, 109–126.

Faes, L., Nollo, G., Ravelli, F., Ricci, L., Vescoci, M., Turatto, M., Pavani, F. and Antolini, R. (2007). Small-sample characterization of stochastic approximation staircases in forced-choice adaptive threshold estimation, *Percept. Psychophys.* **29**, 254–262.

Fechner, G. (1860). *Elemente der psychophysik.* Breitkopf and Hartel, Leipzig, Germany.

Kaernbach, C. (2001). Adaptive threshold estimation with unforced-choice tasks, *Percept. Psychophys.* **63**, 1377–1388.

King-Smith, P. E. and Rose, D. (1997). Principles of an adaptive method for measuring the slope of the psychometric function, *Vision Research* **37**, 1595–1604.

Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: a commentary, *Percept. Psychophys.* **63**, 1421–1455.

Kontsevich, L. L. and Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold, *Vision Research* **39**, 2729–2737.

Kujala, J. and Lukka, T. (2006). Bayesian adaptive estimation: the next dimension, *J. Mathemat. Psychol.* **50**, 369–389.

Kuss, M., Jakel, F. and Wichmann, F. (2005). Approximate Bayesian inference for psychometric functions using mcmc sampling, *Max-Planck-Institut fur biologische Kybernetik Max Planck Institute for Biological Cybernetics Technical Report No. 135*.

Leek, M. R. (2001). Adaptive procedures in psychophysical research, *Percept. Psychophys.* **63**, 1279–1292.

Lesmes, L., Jeon, S., Lu, Z. and Dosher, B. (2006). Bayesian adaptive estimation of threshold *versus* contrast external noise functions: the quick tvc method, *Vision Research* **46**, 3160–3176.

Lesmes, L., Lu, Z.-L., Baek, J. and Albright, T. D. (2010). Bayesian adaptive estimation of the contrast sensitivity function: the quick csf method, *J. Vision* **10**, 1–21.

McKee, S., Klein, S. and Teller, D. (1985). Statistical properties of forced-choice psychometric functions: implications of probit analysis, *Percept. Psychophys.* **37**, 286–298.

Mollon, J. D., Stockman, A. and Polden, P. G. (1987). Transient tritanopia of a second kind, *Vision Research* **27**, 637–650.

Myung, J. and Pitt, M. (2009). Optimal experimental design for model discrimination, *Psycholog. Rev.* **116**, 499–518.

Pelli, D. G., Rubin, G. S. and Legge, G. E. (1986). Predicting the contrast sensitivity of low-vision observers, *J. Optic. Soc. Amer. A*, **13**, 56.

Taylor, M. and Creelman, C. (1967). Pest: efficient estimates on probability functions, *J. Acoustic. Soc. Amer.* **41**, 782–787.

Treisman, M. (1965). Signal detection theory and Crozier's law: derivation of a new sensory scaling procedure, *J. Mathemat. Psychol.* **2**, 205–218.

Treutwein, B. (1995). Adaptive psychophysical procedures, *Vision Research* **35**, 2503–2522.

Twer, T. von der and MacLeod, D. (2001). Optimal nonlinear codes for the perception of natural colours, *Network* **12**, 395–407.

Vul, E., Krizay, E. and MacLeod, D. I. A. (2008). The McCollough effect reflects permanent and transient adaptation in early visual cortex, *J. Vision* **8**, 1–12.

Watson, A. B. and Ahumada, A. J. (2005). A standard model for foveal detection of spatial contrast, *J. Vision* **5**, 717–740.

Watson, A. B. and Pelli, D. G. (1983). Quest: a Bayesian adaptive psychometric method, *Percept. Psychophys.* **33**, 113–120.

Wetherill, G. and Levitt, H. (1965). Sequential estimation of points on a psychometric function, *Brit. J. Mathemat. Statist. Psychol.* **18**, 1–10.

Appendix A: Functional Adaptive Sequential Testing (FAST) Matlab Toolbox

Our implementation of FAST as a Maltab toolbox can be downloaded from the Google Code repository: http://code.google.com/p/fast-toolbox/

The implementation includes functions for:

1. Initializing a FAST structure (*fastFull* and *fastStart*).
2. Updating the structure with new data (*fastUpdate*).
3. Dynamic resampling — implementing the roving range (*fastResample*).
4. Obtaining estimates and confidence intervals on parameters (*fastEstimate*).
5. Plotting the current data and function fits (*fastPlot*).
6. Selecting the next stimulus (*fastChooseY*).

Details on the use of these functions may be found in the associated help files.

The FAST toolbox contains several common threshold functions:

1. A single value — (*funcVal*), a function that does not vary with x — used for simple, unidimensional, threshold estimation: $y_c = \Theta_1$.
2. Linear — (*funcLine*), a simple linear change: $y_c = \Theta_1 + \Theta_2 x$; or re-parameterized in (*funcMscale*) to have more intuitive parameters as Magnification scaling with eccentricity: $y_c = \Theta_1(1 + \Theta_2 x)$.
3. Hyperbolic/Simplified exponential decay — as often used to describe delay discounting, or simple decay of aftereffects (*funcHyperbolic*: $y_c = 1/(1 + \Theta_1 x)$ and *funcExpS*: $y_c = \exp(-\Theta_1 x)$).
4. Exponential decay — (*funcExp*) as commonly used to describe time-courses of aftereffects and adaptation with more variable parameters:
$$y_c = \Theta_1 + (\Theta_2 - \Theta_1)\left(1 - \exp\left(-\frac{x}{\Theta_3}\right)\right).$$
5. Polynomial — (*funcPolynomial*), a polynomial of any degree (determined by the number of parameters provided):
$$y_c = \sum_i \Theta_i x^{i-1}.$$
6. CSF — the contrast threshold function as described by Pelli *et al.* (1986), which is simply a log-parabola relating log(threshold contrast) to log(spatial frequency) (*funcCSF* — more complex parameterizations are included as options):
$$y_c = \Theta_1 + 10^{\Theta_3}(\log_{10}(x) - \Theta_2)^2.$$
7. Threshold *vs* Contrast — including both a simple, linear, threshold *vs* contrast function (*funcTVC*):
$$y_c = \begin{cases} \Theta_2 \dfrac{x}{\Theta_1}, & \text{for } x > \Theta_1, \\ \Theta_2, & \text{for } x \leqslant \Theta_1 \end{cases}$$

and a non-linear version (*funcTVCnl*):

$$y_c = \begin{cases} \Theta_2 \left(\dfrac{x}{\Theta_1} \right)^{\Theta_3}, & \text{for } x > \Theta_1, \\ \Theta_2, & \text{for } x \leqslant \Theta_1. \end{cases}$$

There is also a set of psychometric functions built into FAST; each of these is the full ogival cumulative density function (0–1) parameterized by a scale parameter S. These full ogival functions are then scaled depending on task parameters (number of alternatives, and lapse parameters) by the FAST function *fastPsyScale*:

1. Gumbel: $p_r(y) = 1 - \exp(-\exp(S(y - y_c))$.
2. Normal: $p_r(y) = \frac{1}{2} + \frac{1}{2}(erf((y - y_c)/(\sqrt{2}S)))$.
3. Half-Normal: $p_r(y) = erf((\frac{y}{y_c})^S/\sqrt{2})$, $y > 0$.
4. Logistic: $p_r(y) = (1 + \exp(-\frac{y - y_c}{S}))^{-1}$.
5. Weibull: $p_r(y) = 1 - \exp(-(\frac{y}{y_c})^S)$, $y > 0$.

The Weibull function differs from the others listed in that the critical value: y_c scales the stimulus magnitude instead of subtracting from it. But it is equivalent to case 1, the Gumbel function, if y and y_c are replaced by their logarithms there.

A more thorough description of the virtues and uses of each of these different functions may be found in the manual, or in the help for the individual files in the FAST toolbox.

Appendix B: Minimizing Expected Posterior Entropy

While others have described the calculations necessary to compute the points that minimizes expected posterior entropy (Kontsevich and Tyler, 1999), for the sake of completeness, we will provide the description here as well: First, we define the probability of obtaining a particular response (r as 0 or 1) to a particular stimulus (x, y), by integrating over all possible parameter values (Θ, S):

$$\begin{aligned} P(r = 1|x, y) &= \sum_{\Theta, S} P(r = 1|x, y, \Theta, S) P(\Theta, S), \\ P(r = 0|x, y) &= \sum_{\Theta, S} P(r = 0|x, y, \Theta, S) P(\Theta, S). \end{aligned} \tag{B.1}$$

Using these probabilities as normalizing constants, we can compute the posterior

probability over parameters given a particular response:

$$\begin{aligned} P_1 &= P(\Theta, S|r=1, x, y) = \frac{P(r=1|x, y, \Theta, S)P(\Theta, S)}{P(r=1|x, y)} \\ P_0 &= P(\Theta, S|r=0, x, y) = \frac{P(r=0|x, y, \Theta, S)P(\Theta, S)}{P(r=0|x, y)}. \end{aligned} \tag{B.2}$$

For each of these posterior probability distributions we can define the entropy over parameters as:

$$H(P) = -\sum_{\Theta,S} P \log_2 P, \tag{B.3}$$

and we can compute the expected entropy:

$$\mathbb{E}[H(P)] = H(P_1)P(r=1|x, y) + H(P_0)P(r=0|x, y). \tag{B.4}$$

The goal of global entropy minimization is the solution to:

$$\arg\min_{x,y}\{\mathbb{E}[H(P)]\}, \tag{B.5}$$

and solving for the local solution amounts to solving this with x fixed (to c):

$$\arg\min_{y}\{\mathbb{E}[H(P)]|x=c\}. \tag{B.6}$$

Appendix C: Minimizing Expected Average Posterior Predictive Variance

Here we introduce an optimality criterion that provides a principled, and general, method for trading off efficiency of estimating different parameters. The goal of minimizing posterior variance is to choose a stimulus $(x_\mathrm{p}, y_\mathrm{p})$ that is expected to alter the posterior distribution in such a way that it decreases the variance of the predicted critical values (y^*) for a number of relevant x values (x^{eval}). In other words, this optimality criterion aims to increase the confidence of our predictions about the threshold (y^* value) for a number of points along the threshold function (specified by x^{eval}). In contrast to entropy minimization, the minimization of posterior predictive variance is supported by a rational cost function, that associates errors with a cost proportional to the square of their magnitudes, whereas entropy minimization is tantamount to minimizing error frequency in parameter space without regard to the magnitude of the error either in parameter space, or its consequences for our predictions (Twer and MacLeod, 2001).

The calculation starts just as in Appendix B: First, we define the probability of obtaining a particular response (r as 0 or 1) to a particular stimulus $(x_\mathrm{p}, y_\mathrm{p})$, by integrating over all possible parameter values (Θ, S):

$$\begin{aligned} P(r=1|x_p, y_p) &= \sum_{\Theta,S} P(r=1|x_\mathrm{p}, y_\mathrm{p}, \Theta, S)P(\Theta, S), \\ P(r=0|x_\mathrm{p}, y_\mathrm{p}) &= \sum_{\Theta,S} P(r=0|x_\mathrm{p}, y_\mathrm{p}, \Theta, S)P(\Theta, S). \end{aligned} \tag{C.1}$$

Using these probabilities as normalizing constants, we can compute the posterior probability over parameters given a particular response:

$$\begin{aligned} P(\Theta, S|r=1, x_\mathrm{p}, y_\mathrm{p}) &= \frac{P(r=1|x_\mathrm{p}, y_\mathrm{p}, \Theta, S)P(\Theta, S)}{P(r=1|x_\mathrm{p}, y_\mathrm{p})}, \\ P(\Theta, S|r=0, x_\mathrm{p}, y_\mathrm{p}) &= \frac{P(r=0|x_\mathrm{p}, y_\mathrm{p}, \Theta, S)P(\Theta, S)}{P(r=0|x_\mathrm{p}, y_\mathrm{p})}. \end{aligned} \tag{C.2}$$

Using these (possible) posterior probability distributions, we then compute the variance of the predicted critical values (y^*) over k potential x^{eval} values (x_1^{eval} to x_k^{eval}). These x^{eval} values correspond to the relevant x values that it is our goal to learn about.

Recalling that $y^* = F(x, \Theta)$, the mean and variance for y^* for each x^{eval}, and each response r, can be written as follows:

$$\begin{aligned} \mathbb{E}\big[\mu_i^{y^*}|r=1\big] &= \sum_\Theta F\big(x_i^{\mathrm{eval}}, \Theta\big)P(\Theta, S|r=1, x_\mathrm{p}, y_\mathrm{p}), \\ \mathbb{E}\big[\mu_i^{y^*}|r=0\big] &= \sum_\Theta F\big(x_i^{\mathrm{eval}}, \Theta\big)P(\Theta, S|r=0, x_\mathrm{p}, y_\mathrm{p}), \\ \mathbb{E}\big[\sigma_i^{y^*}|r=1\big] &= \sum_\Theta F\big(x_i^{\mathrm{eval}}, \Theta\big)^2 P_1(\Theta, S) - \mathbb{E}[\mu_i^{y^*}|r=1]^2, \\ \mathbb{E}\big[\sigma_i^{y^*}|r=0\big] &= \sum_\Theta F\big(x_i^{\mathrm{eval}}, \Theta\big)^2 P_0(\Theta, S) - \mathbb{E}[\mu_i^{y^*}|r=0]^2. \end{aligned} \tag{C.3}$$

From these expected posterior predictive variances for each x^{eval} conditioned on a response, we compute the expected average posterior predictive variances by averaging over x^{eval}, and integrating out the two possible responses, weighted by their prior probability:

$$\begin{aligned} \mathbb{E}[\sigma^{y^*}] = {} & \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\big[\sigma_i^{y^*}|r=1\big]P(r=1|x_\mathrm{p}, y_\mathrm{p}) \\ & + \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\big[\sigma_i^{y^*}|r=0\big]P(r=0|x_\mathrm{p}, y_\mathrm{p}). \end{aligned} \tag{C.4}$$

The goal of global predictive variance minimization is the solution to:

$$\arg\min_{x_\mathrm{p}, y_\mathrm{p}} \{\mathbb{E}[\sigma^{y^*}]\} \tag{C.5}$$

and solving for the local solution amounts to solving this with x fixed (to c):

$$\arg\min_{y_\mathrm{p}} \{\mathbb{E}[\sigma^{y^*}]|x_\mathrm{p} = c\}. \tag{C.6}$$

Finally, if some x^{eval} values are more important than others — for instance, if we aim to estimate the contrast sensitivity function to assess reading function, where specific spatial frequencies may be expected to play a greater role, the variances associated with different x^{eval} values can be differentially weighted to reflect this.