
Look-Ahead Acquisition Functions for Bernoulli Level Set Estimation

Benjamin Letham
Meta
bletham@fb.com

Phillip Guan
Reality Labs Research, Meta
philguan@fb.com

Chase Tymms
Reality Labs Research, Meta
tymms@fb.com

Eytan Bakshy
Meta
ebakshy@fb.com

Michael Shvartsman
Reality Labs Research, Meta
michael.shvartsman@fb.com

Abstract

Level set estimation (LSE) is the problem of identifying regions where an unknown function takes values above or below a specified threshold. Active sampling strategies for efficient LSE have primarily been studied in continuous-valued functions. Motivated by applications in human psychophysics where common experimental designs produce binary responses, we study LSE active sampling with Bernoulli outcomes. With Gaussian process classification surrogate models, the look-ahead model posteriors used by state-of-the-art continuous-output methods are intractable. However, we derive analytic expressions for look-ahead posteriors of sublevel set membership, and show how these lead to analytic expressions for a class of look-ahead LSE acquisition functions, including information-based methods. Benchmark experiments show the importance of considering the *global* look-ahead impact on the entire posterior. We demonstrate a clear benefit to using this new class of acquisition functions on benchmark problems, and on a challenging real-world task of estimating a high-dimensional contrast sensitivity function.

1 INTRODUCTION

The *level set estimation* (LSE) problem is to identify the regions where a black-box function $f(\mathbf{x})$ is

above or below a particular threshold γ . Applications include environmental monitoring (*e.g.* identifying areas where a contaminant is at a hazardous level, Gotovos et al., 2013), communications (*e.g.* finding wireless network configurations with acceptable signal quality, Ramakrishnan et al., 2005), and finance (*e.g.* for derivative pricing, Lyu et al., 2021). Evaluating $f(\mathbf{x})$ in these settings entails a time-consuming physical measurement or computer simulation, so the goal is to identify the level set with as few samples as possible.

Sample-efficient LSE is an active learning problem (Kapoor et al., 2007; Hoang et al., 2014), and is related to techniques like Bayesian active learning by disagreement (BALD, Houlsby et al., 2011) that use a surrogate model to perform active sampling. The function $f(\mathbf{x})$ is modeled with a Gaussian process (GP) surrogate, and then active sampling is driven by an acquisition function that selects the most valuable point to sample for identifying the level set. Several acquisition functions have been developed for LSE, as described in Section 2, primarily for continuous-output functions with Gaussian noise.

The motivation for this work comes from psychophysics, a field of science that seeks to understand perception of physical stimuli (Fechner et al., 1966). Understanding and adjusting for the limitations of human perception is important for a variety of downstream applications including audio/visual compression (Pappas et al., 1996; Nadenau et al., 2000), hearing aid design (Moore, 1996), clinical evaluation of auditory and visual impairments (De Boer and Bouwmeester, 1975; Fitzke, 1988), and the design of virtual and augmented reality systems (Kress, 2020).

A common task in psychophysics is to identify *detection thresholds*, the smallest stimulus intensity (*e.g.*, volume of a sound) at which the stimulus can be per-

ceived, usually as a function of other stimulus properties (*e.g.*, frequency). Finding detection thresholds is an LSE problem for the threshold level, notably with Bernoulli observations $y \in \{0, 1\}^n$ indicating whether or not a stimulus was perceived. GP surrogate models and active sampling methods have been applied to psychophysics threshold estimation problems with Bernoulli responses (Gardner et al., 2015; Song et al., 2015, 2017, 2018; Cox and de Vries, 2016; Schlittenlacher et al., 2018, 2020; Owen et al., 2021). This prior work has been limited to a 1-d or 2-d stimulus space of audiograms and, in contrast to our work here, has not used LSE acquisition functions, rather has used global learning methods such as BALD and then extracted the level set *post hoc* from the surrogate posterior. As a notable exception, Owen et al. (2021) used the straddle acquisition, which is described below.

Look-ahead acquisition functions select points based on the impact their observation will have on the subsequent surrogate model. They are the state-of-the-art approach for LSE (Bogunovic et al., 2016; Lyu et al., 2021; Nguyen et al., 2021), as well as for Bayesian optimization (BO) (Scott et al., 2011; Hernández-Lobato et al., 2014; Wang and Jegelka, 2017; Balandat et al., 2020), which is related to LSE by the use of GP surrogates and acquisition functions. They have additionally been shown to be useful in the psychophysics setting, albeit with a parametric model and only in one dimension (Kim et al., 2017).

Look-ahead acquisition functions rely on being able to compute a posterior update given the proposed observation, which can be done analytically with Gaussian observations. Unfortunately, with Bernoulli observations the surrogate model posterior updates are no longer analytic. This is vital for acquisition optimization, which requires computing many thousands of look-ahead posteriors throughout the course of active sampling.

Our work here enables tractable look-ahead acquisition for the Bernoulli LSE problem. Specifically, the contributions of this paper are:

1. Despite the look-ahead surrogate model posterior being intractable with Bernoulli observations, we derive posterior update formulae that, remarkably, enable exact, closed-form computation of several state-of-the-art look-ahead acquisition functions, including information-based approaches.
2. Our look-ahead posterior formulae enable easy construction of novel acquisition functions, which we show by introducing *expected absolute volume change* (EAVC), a new acquisition function inspired by max-value entropy search in BO.
3. We evaluate the acquisition functions with a thorough simulation study, which shows that look-ahead is critical for achieving good level set estimates in high dimensions. We also show that simply being look-ahead is not enough to ensure reliable performance—the acquisition function must also be *global*, a distinction we discuss in Section 2.
4. Our work enables rapid acquisition function computation that is suitable for human experiments, which we show by applying our global look-ahead acquisition functions to a real, high-dimensional psychophysics problem.

2 BACKGROUND

2.1 Models for Level Set Estimation

We consider a black-box function $f(\mathbf{x})$, with $\mathbf{x} \in B$ and $B \subseteq \mathbb{R}^d$ a compact set. Our goal is to identify the set $L_\gamma(f) = \{\mathbf{x} : f(\mathbf{x}) \leq \gamma\}$, known as the *sublevel set*. When f has continuous outputs, it is typical to assume a Gaussian observation model $y = f(\mathbf{x}) + \epsilon$, with ϵ i.i.d. Gaussian noise. We give f a GP prior, then, given a set of observations $D_n = \{(\mathbf{x}_i^{\text{obs}}, y_i^{\text{obs}})\}_{i=1}^n$, the joint posterior for any set of points is a multivariate normal (MVN) with analytic mean and covariance (Rasmussen and Williams, 2006). We denote the marginal posterior at \mathbf{x} as $f(\mathbf{x})|D_n \sim N(\mu(\mathbf{x}|D_n), \sigma^2(\mathbf{x}|D_n))$.

With Bernoulli observations $y \in \{0, 1\}^n$, the standard practice is to use a classification GP based on either a logit or probit model (Kuss and Rasmussen, 2005). Here we focus on the probit case, in which $y \sim \text{Bernoulli}(z(\mathbf{x}))$, with latent probability $z(\mathbf{x}) = \Phi(f(\mathbf{x}))$ and $\Phi(\cdot)$ denoting the Gaussian cumulative distribution function. With Bernoulli observations, the predictive posterior for $f|D_n$ requires approximation, but a variety of efficient approximations have been developed, including Laplace approximation (Williams and Barber, 1998), expectation propagation (Minka, 2001), and variational inference (Hensman et al., 2015). Our experiments use variational inference, but the acquisition functions we develop are agnostic to how inference is done—for our purposes it is sufficient to have an MVN posterior for f .

We define $\theta = \Phi(\gamma)$ to be the desired threshold for the probability function $z(\mathbf{x})$. Then, $L_\gamma(f) = L_\theta(z)$, and so LSE for $z(\mathbf{x})$ can equivalently be done directly on $z(\mathbf{x})$ or in the latent space of $f(\mathbf{x})$. An important quantity for the acquisition functions described below is $\pi(\mathbf{x}|D_n) = P(\mathbf{x} \in L_\gamma(f)|D_n)$, which we call the *level set posterior*. Given the GP posterior, we can compute the level set posterior as

$$\pi(\mathbf{x}|D_n) = \Phi\left(\frac{\gamma - \mu(\mathbf{x}|D_n)}{\sigma(\mathbf{x}|D_n)}\right). \quad (1)$$

2.2 Acquisition for Level Set Estimation

2.2.1 Non-look-ahead Acquisition

During active sampling, at each iteration we select a maximizer of the acquisition function to be the next point sampled. General purpose active sampling strategies such as BALD seek to reduce global uncertainty in the posterior of f or z , which can waste samples by reducing variance in regions that are far from the threshold, the main area of interest for LSE.

Bryan et al. (2005) developed the first acquisition functions tailored for LSE, the most successful of which was the *straddle*, which is of similar flavor to the well-known upper confidence bound (UCB) acquisition function (Srinivas et al., 2010):

$$\alpha_{\text{straddle}}(\mathbf{x}) = j\mu(\mathbf{x} | D_n) - \gamma j + \beta\sigma(\mathbf{x} | D_n).$$

As in UCB, the parameter β drives exploration by selecting higher variance points; Bryan et al. (2005) used $\beta = 1.96$. They also considered as acquisition functions the misclassification probability and the entropy of the level set posterior:

$$\begin{aligned} \alpha_{\text{misclass}}(\mathbf{x}) &= \min(\pi(\mathbf{x} | D_n), 1 - \pi(\mathbf{x} | D_n)), \\ \alpha_{\text{entropy}}(\mathbf{x}) &= H_b(\pi(\mathbf{x} | D_n)), \end{aligned}$$

where $H_b(p) = -p \log_2 p - (1-p) \log_2 (1-p)$ is the binary entropy function. They found that the straddle acquisition performed best, though noted that all of these acquisition functions are “subject to oversampling edge positions.” Gotovos et al. (2013) provided theoretical grounding for LSE and proved sample complexity bounds for extensions of the straddle. Ranjan et al. (2008) adapted the Expected Improvement criterion from BO (Jones et al., 1998) by defining an improvement function with respect to the threshold.

2.3 Look-ahead Acquisition

Subsequent development of acquisition functions for LSE have focused on look-ahead approaches, which consider not just the posterior at the candidate point \mathbf{x} , but how the posterior at a different point \mathbf{x}_q will change as a result of an observation at \mathbf{x} . We denote the look-ahead dataset as $D_{n+1}(\mathbf{x}, y) = D_n \cup \{f(\mathbf{x}, y)\}$, and note that $D_{n+1}(\mathbf{x}, y)$ is a random variable via y . Much of the past work in look-ahead acquisition has relied on the useful GP property that, *with Gaussian observations*, the look-ahead variance $\sigma^2(\mathbf{x}_q | D_{n+1}(\mathbf{x}, y))$ does not depend on y , and can be computed analytically.

Picheny et al. (2010) introduced the first look-ahead acquisition function for LSE, *targeted integrated mean squared error* (tIMSE), which minimizes look-ahead

posterior variance, weighted according to distance to the threshold by some function $w_\gamma(\mathbf{x})$:

$$\begin{aligned} \alpha_{\text{tIMSE}}(\mathbf{x}) &= \int_B \sigma^2(\mathbf{x}_q | D_{n+1}(\mathbf{x}, y)) w_\gamma(\mathbf{x}) d\mathbf{x} \\ &= C \sum_{\mathbf{x}_q \in G} \sigma^2(\mathbf{x}_q | D_{n+1}(\mathbf{x}, y)) w_\gamma(\mathbf{x}_q). \end{aligned} \quad (2)$$

This is an example of a *global* look-ahead acquisition function that evaluates the impact of observing \mathbf{x} on the entire design space, using quasi-Monte Carlo (qMC) integration (Caffisch, 1998). Here G is a quasi-random sequence and $C = \frac{\text{Vol}(B)}{j|G|}$ is a constant that can be ignored for the purpose of acquisition optimization. The tIMSE formula in (2) is analytic due to the analytic form of $\sigma^2(\mathbf{x}_q | D_{n+1}(\mathbf{x}, y))$, and so can be cheaply evaluated in batch across a large set of global reference points G . Bogunovic et al. (2016) used a similar approach by minimizing the total look-ahead posterior variance of the region that was not yet decidedly classified as above or below threshold. Zanette et al. (2018) utilized the analytic look-ahead posterior to construct an expected improvement-based criterion.

An important class of methods are of the form

$$\alpha(\mathbf{x}) = Q(D_n) - E_y [Q(D_{n+1}(\mathbf{x}, y))] \quad (3)$$

where $Q(D_n)$ is a cost function applied to the surrogate model posterior, and $Q(D_{n+1}(\mathbf{x}, y))$ is that same cost function applied to the look-ahead posterior. *Stepwise uncertainty reduction* (SUR, Bect et al., 2012; Chevalier et al., 2014) uses expected classification error as the cost function, and thus maximizes the expected look-ahead misclassification error reduction:

$$\begin{aligned} Q_{\text{GlobalSUR}}(D_n) &= \sum_{\mathbf{x}_q \in G} \min(\pi(\mathbf{x}_q | D_n), 1 - \pi(\mathbf{x}_q | D_n)) \\ &= \sum_{\mathbf{x}_q \in G} \Phi \left(\frac{j\mu(\mathbf{x}_q | D_n) - \gamma j}{\sigma(\mathbf{x}_q | D_n)} \right), \end{aligned} \quad (4)$$

where the global impact is again being estimated via a sum over a quasi-random sequence, as in (2). $Q_{\text{GlobalSUR}}(D_{n+1}(\mathbf{x}, y))$ is computed in the same manner using the look-ahead posterior, so when (5) is plugged into (3), the acquisition function is analytic.

Acquisitions of the form (3) can be formed either as a global look-ahead, as with GlobalSUR, or as a *localized* version that considers the look-ahead impact of observing \mathbf{x} just on \mathbf{x} :

$$Q_{\text{LocalSUR}}(D_n) = \min(\pi(\mathbf{x} | D_n), 1 - \pi(\mathbf{x} | D_n)). \quad (6)$$

Lyu et al. (2021) call this method “gradient SUR.” It avoids the summation over the global reference set G

required to compute GlobalSUR, but provides a less direct measurement of the total value of observing \mathbf{x} .

Azzimonti et al. (2021) applied the SUR strategy to the volume of misclassified points, via the Vorob'ev expectation, with particular focus on active learning of conservative level set estimates that seek to control the type I error rate. In Section 4.1 we also use measures of level set volume to construct an acquisition function, though with a different approach that does not control error rates and is not a SUR strategy.

Global information gain has long been a target for level set acquisition. Bryan et al. (2005) wrote of their acquisition functions, “we believe the good performance of the evaluation metrics proposed below stems from their being heuristic proxies for global information gain.” Nguyen et al. (2021) constructed a localized mutual information (MI) acquisition function by taking

$$Q_{\text{LocalMI}}(D_n) = H_b(\pi(\mathbf{x} | D_n)) \quad (7)$$

in the acquisition form (3), and called this strategy binary entropy search (BES). As with SUR, computation of the look-ahead term $Q_{\text{LocalMI}}(D_{n+1}(\mathbf{x}, y))$ has hitherto relied on the analytic look-ahead posteriors that exist only for Gaussian observations. A criterion for global information gain naturally parallels GlobalSUR:

$$Q_{\text{GlobalMI}}(D_n) = \sum_{\mathbf{x}_q \in 2^G} H_b(\pi(\mathbf{x}_q | D_n)). \quad (8)$$

The only work that has applied LSE acquisition functions to GP classification surrogates is that of Lyu et al. (2021), who studied LSE under high levels of heavy-tailed noise. For the Bernoulli likelihood, they derived an approximation for the intractable look-ahead variance $\sigma^2(\mathbf{x} | D_{n+1}(\mathbf{x}, y))$ based on plug-in estimates, and further approximated $\mu(\mathbf{x} | D_{n+1}(\mathbf{x}, y)) \approx \mu(\mathbf{x} | D_n)$, which enabled approximate SUR computation via (5).

3 BERNOULLI LOOK-AHEAD POSTERIOR UPDATES

With Bernoulli observations, the latent look-ahead posterior $p(f | D_{n+1}(\mathbf{x}, y))$ is not analytic. However, other quantities important for computing acquisition functions are. The moments of the posterior $p(z | D_n)$ are analytic:

Proposition 1. Let $a = \frac{\rho \mu(\mathbf{x} | D_n)}{1 + \sigma^2(\mathbf{x} | D_n)}$ and $c = \frac{\rho}{1 + 2\sigma^2(\mathbf{x} | D_n)}$. Then,

$$\begin{aligned} E[z(\mathbf{x} | D_n)] &= \Phi(a), \\ \text{Var}[z(\mathbf{x} | D_n)] &= \Phi(a) - \Phi(a)^2 - 2T(a, c). \end{aligned}$$

Here $T(\cdot, \cdot)$ is Owen’s T function, which can be computed efficiently (Patefield and Tandy, 2000) and is available in SciPy (Jones et al., 2001). The formula for the mean is well-known and used in several applications of classification GPs (e.g. Houlby et al., 2011); the variance is derived in the supplement. This result enables computing the straddle acquisition on the posterior of z instead of that of f , which accounts for the variance squashing of the $\Phi(\cdot)$ transformation. However, our primary interest lies in look-ahead posteriors, and $p(z | D_{n+1}(\mathbf{x}, y))$ is intractable.

Our main result is that despite the look-ahead posteriors for f and z being intractable, the look-ahead level set posterior $\pi(\mathbf{x}_q | D_{n+1}(\mathbf{x}, y))$ is analytic, in terms of $\Phi(\cdot)$ and $\text{BvN}(\cdot, \cdot; \rho)$, which denotes the standard (zero-mean, unit-variance) bivariate normal distribution function with correlation ρ .

Theorem 1. Let $b_q = \frac{\gamma \mu(\mathbf{x}_q | D_n)}{\sigma(\mathbf{x}_q | D_n)}$, $\sigma(\mathbf{x}_q, \mathbf{x} | D_n) = \text{Cov}[f(\mathbf{x}_q), f(\mathbf{x} | D_n)]$, and

$$Z_q = \text{BvN}\left(a, b_q; \frac{\sigma(\mathbf{x}_q, \mathbf{x} | D_n)}{\sigma(\mathbf{x}_q | D_n) \sqrt{1 + \sigma^2(\mathbf{x} | D_n)}}\right).$$

The level set posterior at \mathbf{x}_q given observation $y = 1$ at \mathbf{x} is

$$\pi(\mathbf{x}_q | D_{n+1}(\mathbf{x}, y = 1)) = \frac{Z_q}{\Phi(a)}.$$

Given observation $y = 0$, the level set posterior at \mathbf{x}_q is

$$\pi(\mathbf{x}_q | D_{n+1}(\mathbf{x}, y = 0)) = \frac{\Phi(b_q) - Z_q}{\Phi(-a)}.$$

If $\mathbf{x}_q = \mathbf{x}$, these results hold with $\sigma(\mathbf{x}_q, \mathbf{x} | D_n) = \sigma^2(\mathbf{x}_q | D_n) = \sigma^2(\mathbf{x} | D_n)$ and $\mu(\mathbf{x}_q | D_n) = \mu(\mathbf{x} | D_n)$.

The proof is in the supplement. There are several routines for efficiently computing $\text{BvN}(\cdot, \cdot; \rho)$ —we use the method of Genz (2004), which produces a differentiable estimate that enables us to compute gradients of the look-ahead level set posterior for acquisition optimization.

This result shows that the look-ahead level set posterior at \mathbf{x}_q given an observation at \mathbf{x} can be computed analytically using only the GP posterior at those points. Fig. 1 shows an example of the global look-ahead posterior for the 2-d discrimination test problem from Section 5. The posteriors are conditional on the binary outcome y , whose distribution is:

Proposition 2. $P(y = 1 | D_n, \mathbf{x}) = \Phi(a)$.

We now show how these results can be used to construct Bernoulli LSE versions of look-ahead acquisition functions from Section 2.3, as well as novel acquisition functions.

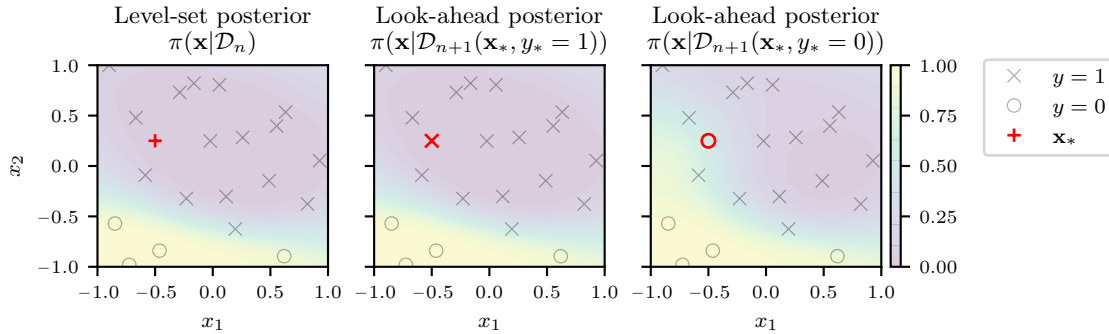


Figure 1: **Level set posteriors.** *Left:* The level set posterior for a model fit to 20 observations (gray markers) of the 2-d discrimination test function. *Middle, Right:* Look-ahead posteriors given an observation at \mathbf{x}^* (red marker) of $y = 1$ (Middle) and $y = 0$ (Right). The look-ahead posteriors are computed analytically using the formulae of Theorem 1, and form the basis of acquisition function computation.

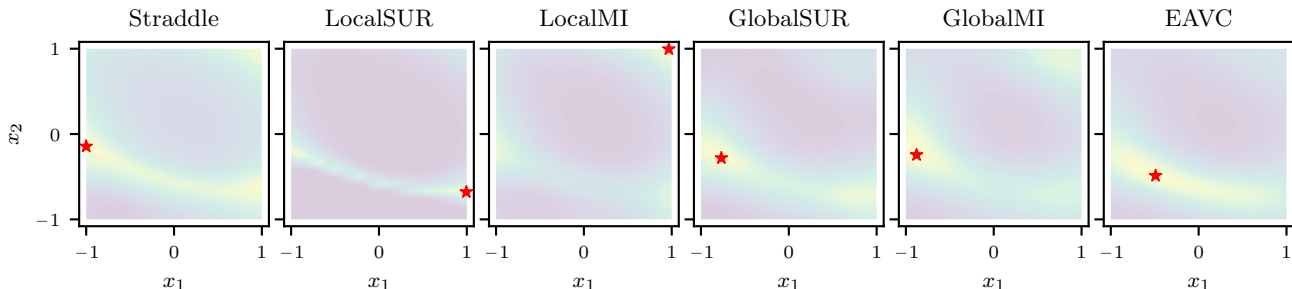


Figure 2: **Acquisition functions.** Acquisition surfaces for the same posterior in Fig. 1. The red star indicates the point that maximizes the acquisition function. Localized look-ahead methods (LocalSUR, LocalMI) show the same edge-seeking as the non-look-ahead Straddle. Global look-ahead methods (GlobalSUR, GlobalMI, EAVC) select interior points near the threshold.

4 BERNOULLI LOOK-AHEAD ACQUISITION FUNCTIONS

The formulae in Theorem 1 enable efficient computation of any look-ahead acquisition function that depends on the model only via the level set posterior. In particular, acquisition functions of the form (3) can be computed for any $Q(\cdot)$ that is a function of $\pi(\cdot)$, which includes both SUR and MI. The key quantity is the expectation over the look-ahead posteriors, $\mathbb{E}_y [Q(D_{n+1}(\mathbf{x}, y))]$, which can be computed via Theorem 1 and Proposition 2. For instance,

$$\mathbb{E}_y [Q_{\text{LocalMI}}(D_{n+1}(\mathbf{x}, y))] = \Phi(a) H_b \left(\frac{Z_q}{\Phi(a)} \right) + \Phi(-a) H_b \left(\frac{\Phi(b_q)}{\Phi(-a)} Z_q \right).$$

Expressions for GlobalSUR, LocalSUR, LocalMI, and GlobalMI in the Bernoulli case are obtained by plugging the posterior formulae into (4), (6), (7), and (8), respectively. These expressions are fully analytic; complete expressions for each acquisition function are given in the supplement.

4.1 A Novel Volume Acquisition Function

Besides SUR and MI, other quantities of interest for acquisition functions can be computed with the formulae in Theorem 1. In BO, the successful max-value entropy search acquisition function (Wang and Jegelka, 2017) finds the point that is most informative about the best *function value* as opposed to being informative about the *best point*. The same concept can be applied to LSE by seeking points that are informative about the *volume* of the sublevel set $L_\gamma(f)$. The qMC expected sublevel-set volume is

$$\tilde{V}(D_n) = C \sum_{\mathbf{x}_q \in \mathcal{G}} \pi(\mathbf{x}_q | D_n). \quad (9)$$

There are many approaches one might take to assess how informative a candidate point \mathbf{x} is about $\text{Vol}(L_\gamma(f))$, and here we consider the expected absolute volume change (EAVC) produced by observing \mathbf{x} , which is a direct measure of how sensitive $\text{Vol}(L_\gamma(f))$ is to the outcome at \mathbf{x} :

$$\alpha_{\text{EAVC}}(\mathbf{x}) = \mathbb{E}_y \left[\left| \tilde{V}(D_n) - \tilde{V}(D_{n+1}(\mathbf{x}, y)) \right| \right].$$

The look-ahead volumes under $y = 0$ and $y = 1$ can be computed by plugging the look-ahead posteriors into (9), producing an analytic acquisition function; the complete expression is given in the supplement. Along with SUR and MI, EAVC shows the breadth of the acquisition functions that can be computed using Theorem 1. The acquisition functions cover a broad set of target criteria (misclassification error, entropy, and volume), and also have variety in their functional form: SUR and MI are both of the form (3), while EAVC is not, showing that acquisition functions do not have to be of the form (3) in order to be computed via Theorem 1, or to be useful.

Fig. 2 shows each acquisition function when computed on the posterior of Fig. 1. The acquisition functions are broadly similar, with elevated values along the threshold and in the high-uncertainty region of the top-right corner. However, they have substantially different maxima, and thus propose different candidates for the next iteration. Straddle selects the point on the threshold at the edge of the design space, consistent with the observation of Bryan et al. (2005) that it oversamples the edges. The localized look-ahead methods (LocalMI, LocalSUR) also select points on the edges. Edge points have high uncertainty in GP posteriors, and edge samples are highly informative about the edge point itself, the criterion for a localized look-ahead method. However, edge points are less informative about the global surface as a whole, and so we see the global look-ahead methods (EAVC, GlobalMI, GlobalSUR) select interior points along the highest-uncertainty portion of the threshold.

5 BENCHMARK EXPERIMENTS

We use three benchmark problems to evaluate and understand the performance of look-ahead acquisition functions for Bernoulli LSE. The first is a binarized version of the classic Hartmann 6-d function, using the same modified version of Lyu et al. (2021), plus an affine transformation and an inverse probit transform to produce Bernoulli responses; see the supplement for the full functional form.

Inspired by our primary application area of psychophysics, the other benchmark problems are a low-dimensional ($d = 2$) and a high-dimensional ($d = 8$) synthetic function modeled after psychophysical discrimination tasks. The 2-d discrimination function is from Owen et al. (2021), and is linear in an intensity dimension (x_2) with slope given by a polynomial function of the other dimension (x_1). It is modeled after psychometric functions in domains such as haptics and multisensory perception. The 8-d discrimination function is similarly linear in an intensity dimension, with

a slope given by a sum of shifted and scaled sinusoids, whose parameters form the other seven dimensions. Functional forms of both are given in the supplement.

Both discrimination test functions mimic psychophysics tasks in which the participant must identify which of two images/sounds/etc. has the stimulus, and we record if the identification was correct ($y = 1$) or incorrect ($y = 0$). When the stimulus intensity is very low, the participant must guess and the probability of a correct response is lower-bounded at $p = 0.5$, and reaches this minimum along many of the edges of the search space. The goal in the experiment is to identify the detection threshold, where $p = 0.75$.

We applied eight active sampling strategies to each of the three problems: the non-look-ahead straddle, applied to the posterior of the response probability z as in Proposition 1; localized look-ahead methods LocalSUR and LocalMI; global look-ahead methods GlobalSUR, GlobalMI, and EAVC; the approximate global SUR method of Lyu et al. (2021), ApproxGlobalSUR; and quasi-random search with a scrambled Sobol sequence (Owen, 1998). To ensure differences are due solely to the acquisition function, all methods used the same GP classification surrogate model and the same gradient optimization of the acquisition function—see the supplement for details¹. We evaluated performance using the Brier score (Brier, 1950), a strictly proper scoring rule (Gneiting and Raftery, 2007) that assesses the quality and calibration of the level set posterior. See the supplement for extended results, including additional evaluation metrics, additional baseline methods such as BALD, and a sensitivity study.

Fig. 3 shows the results of the benchmark experiments. In the 2-d discrimination problem, all LSE acquisition functions performed significantly better than the quasi-random baseline, and straddle and LocalMI were among the best-performing methods. However, on the high-dimensional problems, LocalMI performed worse than quasi-random search by a substantial margin, as did straddle and LocalSUR on the 8-d problem. Global methods generally outperformed localized methods and the quasi-random baseline, consistent with the conclusion that global look-ahead is a key ingredient needed to achieve consistently strong performance. Among global methods, GlobalSUR showed variable performance, consistent with the findings of Lyu et al. (2021) who noted that SUR underperformed with classification metamodels. Interestingly, ApproxGlobalSUR generally outperformed GlobalSUR, which seemed to underexplore on the 8-d discrimination

¹Software for reproducing all of the methods and experiments in this paper, including the real-world task, is available at https://github.com/facebookresearch/bernoulli_lse/

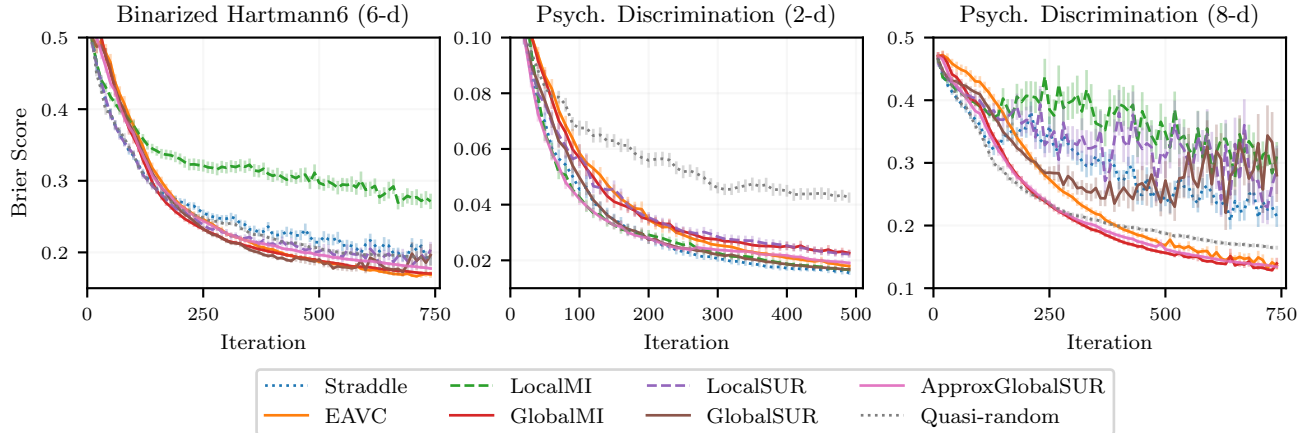


Figure 3: **Benchmark results.** Brier score (lower is better) for the level set posterior as a function of active sampling iteration, averaged over 280 repeated runs with error bars showing two standard errors. On the high-dimensional problems, straddle and the localized look-ahead methods (LocalMI, LocalSUR) did not perform better than the quasi-random baseline. Global methods GlobalMI and EAVC were best in high dimensions.

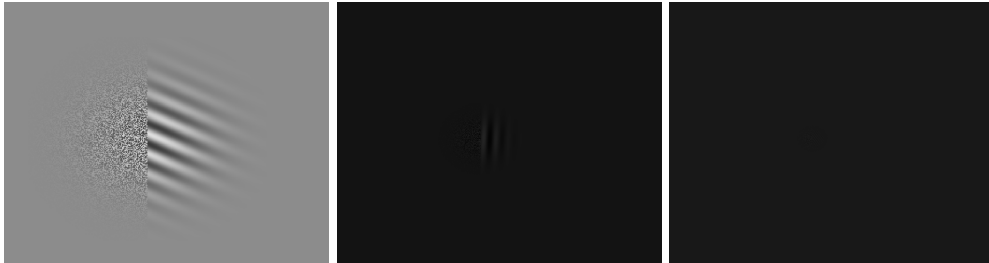


Figure 4: **Example stimuli from the real psychophysical discrimination task.** The human’s task is to determine whether the white noise is on the right or the left of the image. In the three examples in this figure, the correct response is “left.” *Left:* A stimulus whose discrimination probability is approximately 1. The stimulus is both large and clearly visible against the background. *Middle:* A stimulus near the detection threshold of $p = 0.75$. The contrast is very low and the stimulus is small. *Right:* A stimulus whose discrimination probability is $p = 0.5$. The stimulus is essentially invisible against the background and the participant must resort to random guessing. The real stimulus was additionally animated with some temporal frequency, and appeared at some distance and angle from the center of the screen.

problem, suggesting that the posterior approximations encouraged better exploration. GlobalMI and EAVC both performed consistently well across problems.

Consistent with Fig. 2, we found that localized look-ahead methods sampled significantly more near the edges. On the Binarized Hartmann6 problem, 99% of samples with LocalMI were near an edge, compared to 80% with GlobalMI, 55% with EAVC, and 47% with quasi-random search—see the supplement for a full analysis of edge sampling behavior. On the low-dimensional problem, the tendency to oversample the edges was not as detrimental and the localized methods showed some advantage. However, they failed badly in high dimensions where a higher degree of exploration was critical.

Importantly, the wall time to select the next point with

the global acquisition functions was generally under a second in a standard multi-core setting, making these methods suitable for real human experiments—see the supplement for details on running times.

6 REAL PSYCHOPHYSICS TASK

The contrast sensitivity function (CSF) describes how human visual sensitivity depends on stimulus properties such as spatial frequency and contrast. It is a crucial model of human vision used for clinical assessment (Owsley, 2003) and in applied settings to estimate visual appearance (Campbell and Robson, 1968; Mantuik et al., 2011, 2021). Contrast sensitivity is affected by a number of variables including eccentricity, size, color, orientation, mean luminance, spatial frequency, and temporal frequency (Robson, 1966; Wright and

Johnston, 1983; Mullen, 1985; Foley et al., 2007; Kim et al., 2020). Contrast sensitivity thresholds across these dimensions have been previously measured piecemeal with traditional psychophysical methods which cannot scale beyond three or four dimensions, and therefore a definitive CSF simultaneously accounting for all of these variables does not exist.

To evaluate our methods on CSF threshold identification, we ran a real CSF psychophysical discrimination study on one of the authors using Psychopy (Peirce et al., 2019). As is standard for CSF measurement, stimuli were animated, Gaussian-windowed sinusoidal gratings, conventionally known as Gabor patches (Gabor, 1946), generated by convolving a sinusoid with a Gaussian, which then had one half of the image scrambled, and were animated by advancing the phase of the sinusoid. Fig. 4 shows three examples of stimuli used in this experiment, where the task for the participant was to identify whether the scrambled half of the image was on the left or the right, and the y response is whether they correctly identified the side with the stimulus ($y = 1$) or not ($y = 0$). This was thus a discrimination task like those in Section 5 where the success rate was lower-bounded by $p = 0.5$ (guessing). The goal was to identify the detection threshold where $p = 0.75$. For each stimulus, we varied eight properties that are known to affect the CSF: background luminance, stimulus contrast, stimulus orientation, temporal frequency of the animation, spatial frequency, stimulus size, and location on the screen (angle and distance from center). We collected responses to 1000 quasi-random stimuli which were used to fit a 6-d surrogate model for benchmarking purposes—see the supplement for details. We evaluated the same LSE methods from Section 5, using the surrogate model as ground truth from which Bernoulli responses were simulated.

The results on the real-world CSF task are shown in Fig. 5. As in the high-dimensional synthetic problems, straddle and the local methods (LocalMI and LocalSUR) failed to outperform the non-active quasi-random baseline. Global look-ahead methods continued to consistently perform best.

7 DISCUSSION

We have derived analytic formulae for the look-ahead level-set posteriors with Bernoulli observations, and used them to construct acquisition functions for Bernoulli LSE. The formulae enabled applying state-of-the-art approaches for Gaussian observations, SUR and MI, to the Bernoulli setting, while also making it easy to construct EAVC, a novel acquisition function. Prior to this work, none of the look-ahead acquisition

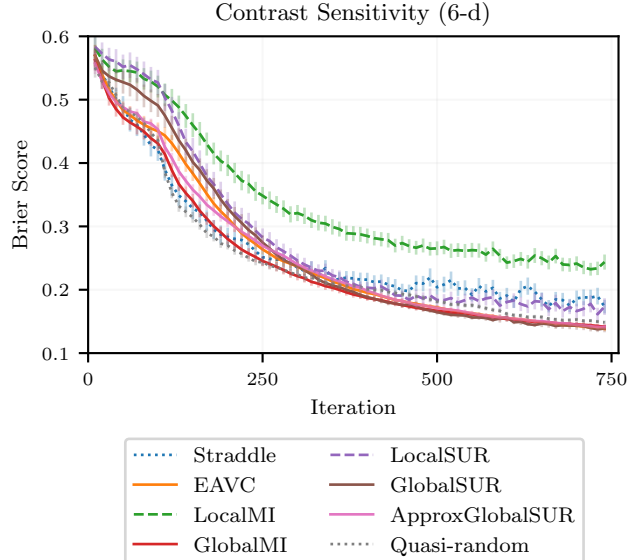


Figure 5: **Real psychophysics task results.** Performance averaged across 280 repeated runs on the real contrast sensitivity function task. As in the synthetic benchmarks, non-global acquisition functions generally performed poorly, and the quasi-random baseline was remarkably strong. Global look-ahead methods performed best, and significantly better than the other methods.

functions could be applied directly to Bernoulli LSE, leaving the straddle and quasi-random search as the primary available strategies. The results of Theorem 1 have thus greatly expanded the Bernoulli LSE acquisition toolbox.

Our empirical results showed that the *global* look-ahead acquisition functions developed in this paper are essential for consistently achieving good estimates of the level set. The localized look-ahead methods LocalSUR and LocalMI have previously been studied with Gaussian observations, and shown to perform well on those problems (Picheny et al., 2010; Nguyen et al., 2021); we found that in high-dimensional problems with Bernoulli observations, they often performed worse than quasi-random search. This highlights the importance of evaluating acquisition functions particularly for this setting, as well as the key differences between this setting and the more typical Gaussian observations.

The setting we consider here is particularly challenging for active learning because Bernoulli observations in essence have very high noise levels in the region of interest. The standard deviation of a Bernoulli random variable is $\sqrt{p(1-p)}$, which at the target threshold here of $p = 0.75$ is approximately 0.43; this nearly

equals the total variation of the entire function (p from 0.5 to 1). Noise levels on par with the total function variation make it difficult to learn a decent global surrogate, and mean that many observations are required to significantly reduce posterior variance. This exacerbates existing boundary over-exploration pathologies of GPs in general (Siivola et al., 2018) and classification GPs in particular (e.g. Song et al., 2017). Localized look-ahead methods target areas of high variance on the edges, and, unlike in the Gaussian case, get stuck on the edges because in high dimensions the variance is never sufficiently reduced. Thus having a look-ahead acquisition function is not by itself sufficient to achieve good performance, we must look ahead to the global impact of a point.

The strong performance of quasi-random search as a baseline on the high-dimensional problems highlights an interesting difference between LSE and BO, where quasi-random search does not typically provide as strong a comparator. In BO, the target is often a single point, the global optimum. In LSE, we are trying to learn the boundary of the set $L_\gamma(f)$, which in general could be a $(d - 1)$ -dimensional manifold. LSE thus inherently requires more global evaluation than BO. Quasi-random search performs maximally global evaluation of the function, which is much less detrimental for LSE than for BO. Much of the literature on LSE has not included a critical evaluation of random or quasi-random search in benchmark experiments—our results show that one should always be included.

Our real-world application focused on a visual psychophysics task, but there is a broad set of other useful applications for Bernoulli LSE. In robotics, one may wish to find the set of controller parameters under which a robot can successfully traverse an obstacle with high probability (Tesch et al., 2013). This problem can be cast as Bernoulli LSE. Several other important classes of problems have discrete outputs, such as ordinal regression (Chu et al., 2005) and preference learning (Chu and Ghahramani, 2005; Fürnkranz and Hüllermeier, 2010). Finding all configurations that are preferred to a current baseline via preference learning can be cast as Bernoulli LSE. Finally, psychophysics itself comprises many application areas: it is a foundational component of AR/VR research, a rapidly developing area of computing, while also having several important applications in disease diagnostics and management, as described in the Introduction.

While our results show strong estimation performance with the acquisition functions developed here, there remain several important areas of future work. First, in acquisition function development: There is no single best acquisition function for all problems, and LSE as a field will benefit from expanding the acquisition tool-

box. Our results make it easy to compute any look-ahead acquisition function that is a function of the level set posterior, which can accelerate development for Bernoulli LSE. Second, in the scope of the look-ahead: Recent work in BO has targeted multi-step acquisition functions, in which we look ahead multiple steps of acquisition rather than just one as is done here (González et al., 2016; Jiang et al., 2020a,b). Our results here could form a basis for non-myopic Bernoulli LSE. Finally, in the model classes: Similar results may exist for other types of classification GPs, such as the logit model or skew GPs (Benavoli et al., 2020), which have different advantages relative to the probit model used here.

Acknowledgements

Thanks to James Wilson for developing a differentiable BVN routine that enabled gradient optimization of the look-ahead acquisition functions; Lucy Owen for assistance with implementing the human experiments; the AEPsych team for development of the AEPsych platform in which the experiments were run; and the anonymous reviewers for their constructive feedback.

References

- Azzimonti, D., Ginsbourger, D., Chevalier, C., Bect, J., and Richet, Y. (2021). Adaptive design of experiments for conservative estimation of excursion sets. *Technometrics*, 63(1):13–26.
- Balandat, M., Karrer, B., Jiang, D. R., Daulton, S., Letham, B., Wilson, A. G., and Bakshy, E. (2020). BoTorch: A framework for efficient Monte-Carlo Bayesian optimization. In *Advances in Neural Information Processing Systems 33*, NeurIPS, pages 21524–21538.
- Bect, J., Ginsbourger, D., Li, L., Picheny, V., and Vazquez, E. (2012). Sequential design of computer experiments for the estimation of a probability of failure. *Statistics and Computing*, 22:773–793.
- Benavoli, A., Azzimonti, D., and Piga, D. (2020). Skew Gaussian process for classification. *Machine Learning*, 109:1877–1902.
- Bogunovic, I., Scarlett, J., Krause, A., and Cevher, V. (2016). Truncated variance reduction: A unified approach to Bayesian optimization and level-set estimation. In *Advances in Neural Information Processing Systems 29*, NIPS, pages 1507–1515.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.
- Bryan, B., Nichol, R. C., Genovese, C. R., Schneider, J., Miller, C. J., and Wasserman, L. (2005). Active

- learning for identifying function threshold boundaries. In *Advances in Neural Information Processing Systems 18*, NIPS, pages 163–170.
- Caffisch, R. E. (1998). Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica*, 7:1–49.
- Campbell, F. and Robson, J. (1968). Application of Fourier analysis to the visibility of gratings. *The Journal of Physiology*, 197(3):551–66.
- Chevalier, C., Bect, J., Ginsbourger, D., Vazquez, E., Picheny, V., and Richet, Y. (2014). Fast parallel Kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. *Technometrics*, 56(4):455–465.
- Chu, W. and Ghahramani, Z. (2005). Preference learning with Gaussian processes. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML, pages 137–144.
- Chu, W., Ghahramani, Z., and Williams, C. K. (2005). Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1019–1041.
- Cox, M. and de Vries, B. (2016). A Bayesian binary classification approach to pure tone audiometry. *arXiv preprint arXiv:1511.08670*.
- De Boer, E. and Bouwmeester, J. (1975). Clinical psychophysics: Illustrated by the problem of auditory overload. *Audiology*, 14(4):274–299.
- Fechner, G. T., Howes, D. H., and Boring, E. G. (1966). *Elements of psychophysics*, volume 1. Holt, Rinehart and Winston, New York.
- Fitzke, F. (1988). Clinical psychophysics. *Eye*, 2(1):S233–S241.
- Foley, J. M., Varadharajan, S., Koh, C. C., and Farias, M. C. (2007). Detection of Gabor patterns of different sizes, shapes, phases and eccentricities. *Vision Research*, 47(1):85–107.
- Fürnkranz, J. and Hüllermeier, E. (2010). Preference learning and ranking by pairwise comparison. In Fürnkranz, J. and Hüllermeier, E., editors, *Preference learning*, pages 65–82. Springer, Berlin, Heidelberg.
- Gabor, D. (1946). Theory of communication. Part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, 93(26):429–441.
- Gardner, J. R., Song, X. D., Weinberger, K. Q., Barbour, D., and Cunningham, J. P. (2015). Psychophysical detection testing with Bayesian active learning. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI, pages 286–297.
- Genz, A. (2004). Numerical computation of rectangular bivariate and trivariate normal and t probabilities. *Statistics and Computing*, 14:251–260.
- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.
- González, J., Osborne, M., and Lawrence, N. (2016). GLASSES: Relieving the myopia of Bayesian optimisation. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, AISTATS, pages 790–799.
- Gotovos, A., Casati, N., Hitz, G., and Krause, A. (2013). Active learning for level set estimation. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI, pages 1344–1350.
- Hensman, J., Matthews, A., and Ghahramani, Z. (2015). Scalable variational Gaussian process classification. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, AISTATS, pages 351–360.
- Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. (2014). Predictive entropy search for efficient global optimization of black-box functions. In *Advances in Neural Information Processing Systems 27*, NIPS, pages 918–926.
- Hoang, T. N., Low, K. H., Jaillet, P., and Kankanhalli, M. (2014). Nonmyopic ϵ -bayes-optimal active learning of Gaussian processes. In *Proceedings of the 31th International Conference on Machine Learning*, ICML, pages 739–747.
- Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. (2011). Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.
- Jiang, S., Chai, H., Gonzalez, J., and Garnett, R. (2020a). BINOCULARS for efficient, nonmyopic sequential experimental design. In *Proceedings of the 37th International Conference on Machine Learning*, ICML, pages 4794–4803.
- Jiang, S., Jiang, D., Balandat, M., Karrer, B., Gardner, J., and Garnett, R. (2020b). Efficient nonmyopic Bayesian optimization via one-shot multi-step trees. In *Advances in Neural Information Processing Systems 33*, NeurIPS, pages 18039–18049.
- Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492.
- Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open source scientific tools for Python.

- Kapoor, A., Grauman, K., Urtasun, R., and Darrell, T. (2007). Active learning with Gaussian processes for object categorization. In *Proceedings of the 11th IEEE International Conference on Computer Vision, ICCV*.
- Kim, M., Ashraf, M., Pérez-Ortiz, M., Martinovic, J., Wuerger, S., and Mantiuk, R. (2020). Contrast sensitivity functions for HDR displays. *London Imaging Meeting*, 2020:44–48.
- Kim, W., Pitt, M. A., Lu, Z. L., and Myung, J. I. (2017). Planning beyond the next trial in adaptive experiments: A dynamic programming approach. *Cognitive Science*, 41(8):2234–2252.
- Kress, B. C. (2020). *Optical architectures for augmented-, virtual-, and mixed-reality headsets*. SPIE, Bellingham, Washington.
- Kuss, M. and Rasmussen, C. E. (2005). Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research*, 6:1679–1704.
- Lyu, X., Binois, M., and Ludkovski, M. (2021). Evaluating Gaussian process metamodels and sequential designs for noisy level set estimation. *Statistics and Computing*, 31(43):1–21.
- Mantiuk, R., Kim, K. J., Rempel, A. G., and Heidrich, W. (2011). HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on Graphics*, 30(4):1–14.
- Mantiuk, R. K., Denes, G., Chapiro, A., Kaplanyan, A., Rufo, G., Bachy, R., Lian, T., and Patney, A. (2021). FovVideoVDP: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics*, 40(4):1–19.
- Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, UAI*, pages 362–369.
- Moore, B. C. (1996). Perceptual consequences of cochlear hearing loss and their implications for the design of hearing aids. *Ear and Hearing*, 17(2):133–161.
- Mullen, K. (1985). The contrast sensitivity of human color vision to red-green and blue-yellow chromatic gratings. *The Journal of Physiology*, 359:381–400.
- Nadenau, M. J., Winkler, S., Alleysson, D., and Kunt, M. (2000). Human vision models for perceptually optimized image processing – a review. *Proceedings of the IEEE*, 32.
- Nguyen, Q. P., Low, B. K. H., and Jaillet, P. (2021). An information-theoretic framework for unifying active learning problems. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence, AAAI*, pages 9126–9134.
- Owen, A. B. (1998). Scrambling Sobol’ and Niederreiter-Xing points. *Journal of Complexity*, 14:466–489.
- Owen, D. B. (1980). A table of normal integrals. *Communications in Statistics - Simulation and Computation*, 9(4):389–419.
- Owen, L., Browder, J., Letham, B., Stocck, G., Tymms, C., and Shvartsman, M. (2021). Adaptive nonparametric psychophysics. *arXiv preprint arXiv:2104.09549*.
- Owsley, C. (2003). Contrast sensitivity. *Ophthalmology Clinics of North America*, 16(2):171–177.
- Pappas, T. N., Michel, T. A., and Hinds, R. O. (1996). Supra-threshold perceptual image coding. In *Proceedings of 3rd IEEE International Conference on Image Processing, ICIP*, pages 237–240.
- Patefield, M. and Tandy, D. (2000). Fast and accurate calculation of Owen’s T function. *Journal of Statistical Software*, 5(5):1–25.
- Peirce, J., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E. K., and Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51:195–203.
- Picheny, V., Ginsbourger, D., Roustant, O., Haftka, R. T., and Kim, N.-H. (2010). Adaptive designs of experiments for accurate approximation of a target region. *Journal of Mechanical Design*, 132(7):071008.
- Ramakrishnan, N., Bailey-Kellogg, C., Tadepalli, S., and Pandey, V. N. (2005). Gaussian processes for active data mining of spatial aggregates. In *Proceedings of the 2005 SIAM International Conference on Data Mining, SDM*, pages 427–438.
- Ranjan, P., Bingham, D., and Michailidis, G. (2008). Sequential experiment design for contour estimation from complex computer codes. *Technometrics*, 50(4):527–541.
- Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, Massachusetts.
- Robson, J. G. (1966). Spatial and temporal contrast-sensitivity functions of the visual system. *Journal of the Optical Society of America*, 56(8):1141–1142.
- Schlittenlacher, J., Turner, R. E., and Moore, B. C. (2020). Application of Bayesian active learning to the estimation of auditory filter shapes using the notched-noise method. *Trends in Hearing*, 24.

- Schlittenlacher, J., Turner, R. E., and Moore, B. C. J. (2018). Audiogram estimation using Bayesian active learning. *The Journal of the Acoustical Society of America*, 144(1):421–430.
- Scott, W., Frazier, P., and Powell, W. (2011). The correlated knowledge gradient for simulation optimization of continuous parameters using Gaussian process regression. *SIAM Journal of Optimization*, 21(3):996–1026.
- Siivola, E., Vehtari, A., Vanhatalo, J., Gonzalez, J., and Andersen, M. R. (2018). Correcting boundary over-exploration deficiencies in Bayesian optimization with virtual derivative sign observations. In *Proceedings of the 28th IEEE International Workshop on Machine Learning for Signal Processing, MLSP*.
- Song, X. D., Garnett, R., and Barbour, D. L. (2017). Psychometric function estimation by probabilistic classification. *The Journal of the Acoustical Society of America*, 141(4):2513–2525.
- Song, X. D., Sukesan, K. A., and Barbour, D. L. (2018). Bayesian active probabilistic classification for psychometric field estimation. *Attention, Perception, & Psychophysics*, 80(3):798–812.
- Song, X. D., Wallace, B. M., Gardner, J. R., Ledbetter, N. M., Weinberger, K. Q., and Barbour, D. L. (2015). Fast, continuous audiogram estimation using machine learning. *Ear and Hearing*, 36(6):e326–e335.
- Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning, ICML*.
- Tesch, M., Schneider, J., and Choset, H. (2013). Expensive function optimization with stochastic binary outcomes. In *Proceedings of the 30th International Conference on Machine Learning, ICML*, pages 1283–1291.
- Wang, Z. and Jegelka, S. (2017). Max-value entropy search for efficient Bayesian optimization. In *Proceedings of the 34th International Conference on Machine Learning, ICML*, pages 3627–3635.
- Williams, C. K. I. and Barber, D. (1998). Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351.
- Wright, M. and Johnston, A. (1983). Spatiotemporal contrast sensitivity and visual field locus. *Vision Research*, 23:983–989.
- Zanette, A., Zhang, J., and Kochenderfer, M. J. (2018). Robust super-level set estimation using Gaussian processes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases, ECML PKDD*, pages 276–291.

Look-Ahead Acquisition Functions for Bernoulli Level Set Estimation: Supplementary Materials

S1 PROOFS

Here we provide proofs of the results in Propositions 1 and 2, and Theorem 1. Owen (1980) undertook the Herculean effort of producing a comprehensive collection of solutions to Gaussian integrals. We use several of his results, given in the following Lemma. We use the following notation for special functions:

- $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function.
- $\phi(\cdot)$ is the standard Gaussian density function.
- $\phi(\cdot; \mu, \sigma^2)$ is the Gaussian density with mean μ and variance σ^2 , so that $\phi(x; \mu, \sigma^2) = \frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right)$.
- $T(\cdot, \cdot)$ is Owen's T function.
- $\text{BvN}(\cdot, \cdot; \rho)$ is the standard bivariate normal cumulative distribution function.

Lemma S1.

$$\int_{\gamma}^{+\gamma} \Phi(a + bx)\phi(x)dx = \Phi\left(\rho\frac{a}{1+b^2}\right), \quad (\text{S1})$$

$$\int_{\gamma}^{+\gamma} \Phi(a + bx)^2\phi(x)dx = \Phi\left(\rho\frac{a}{1+b^2}\right) - 2T\left(\rho\frac{a}{1+b^2}, \rho\frac{1}{1+2b^2}\right), \quad (\text{S2})$$

$$\int_{\gamma}^{+\gamma} \Phi(a + bx)\Phi(h + kx)\phi(x)dx = \text{BvN}\left(\rho\frac{a}{1+b^2}, \rho\frac{h}{1+k^2}; \rho\frac{bk}{1+b^2}\rho\frac{1}{1+k^2}\right). \quad (\text{S3})$$

These results are 10,010.8, 20,010.4, and 20,010.3, respectively, from Owen (1980).

Throughout the proofs in this section, for notational convenience and clarity we will use the shorthand $f = f(\mathbf{x})$ to represent the latent function value at \mathbf{x} , and will let $\mu = \mu(\mathbf{x} | D_n)$ and $\sigma^2 = \sigma^2(\mathbf{x} | D_n)$ indicate the posterior mean and variance of f . Thus, $f | D_n \sim N(\mu, \sigma^2)$. Table S1 provides a complete list of the abbreviated notation used throughout this supplement.

Short-hand notation	Definition
f, f_q	$f(\mathbf{x}), f(\mathbf{x}_q)$
μ, μ_q	$\mu(\mathbf{x} D_n), \mu(\mathbf{x}_q D_n)$
σ, σ_q	$\sigma(\mathbf{x} D_n), \sigma(\mathbf{x}_q D_n)$
σ_q	$\text{Cov}[f(\mathbf{x}_q), f(\mathbf{x}) D_n]$
a	$\rho\frac{\mu}{1+\sigma^2}$
c	$\rho\frac{1}{1+2\sigma^2}$
b, b_q	$\frac{\gamma - \mu}{\sigma}, \frac{\gamma - \mu_q}{\sigma_q}$
Z_q	$\text{BvN}\left(a, b_q; \frac{\rho\sigma_q}{\sigma_q(1+\sigma^2)}\right)$
Z	$\text{BvN}\left(a, b; \rho\frac{\sigma}{1+\sigma^2}\right)$

Table S1: Abbreviated notation used throughout the proofs and other results in this supplement.

Proof of Proposition 1.

$$\begin{aligned} \mathbb{E}[z(\mathbf{x}) | D_n] &= \mathbb{E}[\Phi(f) | D_n] \\ &= \int_{\gamma}^{+\infty} \Phi(f) \phi(f; \mu, \sigma^2) df \end{aligned} \quad (\text{S4})$$

$$\begin{aligned} &= \frac{1}{\sigma} \int_{\gamma}^{+\infty} \Phi(f) \phi\left(\frac{f - \mu}{\sigma}\right) df \\ &= \int_{\gamma}^{+\infty} \Phi(\mu + \sigma \tilde{f}) \phi(\tilde{f}) d\tilde{f} \end{aligned} \quad (\text{S5})$$

$$= \Phi\left(\frac{\mu}{\sqrt{1 + \sigma^2}}\right), \quad (\text{S6})$$

where (S5) used the change of variables $\tilde{f} = \frac{f - \mu}{\sigma}$, and (S6) used (S1). For the variance,

$$\text{Var}[z(\mathbf{x}) | D_n] = \mathbb{E}[z(\mathbf{x})^2 | D_n] - (\mathbb{E}[z(\mathbf{x}) | D_n])^2,$$

where, similarly as before,

$$\begin{aligned} \mathbb{E}[z(\mathbf{x})^2 | D_n] &= \int_{\gamma}^{+\infty} \Phi(f)^2 p(f | D_n) df \\ &= \int_{\gamma}^{+\infty} \Phi(\mu + \sigma \tilde{f})^2 \phi(\tilde{f}) d\tilde{f} \\ &= \Phi\left(\frac{\mu}{\sqrt{1 + \sigma^2}}\right)^2 \quad 2T\left(\frac{\mu}{\sqrt{1 + \sigma^2}}, \frac{1}{\sqrt{1 + 2\sigma^2}}\right) \end{aligned}$$

using (S2). Letting $a = \frac{\mu}{\sqrt{1 + \sigma^2}}$ and $c = \frac{1}{\sqrt{1 + 2\sigma^2}}$ as in Proposition 1, we have that

$$\text{Var}[z(\mathbf{x}) | D_n] = \Phi(a)^2 \quad \Phi(a)^2 \quad 2T(a, c).$$

□

Proof of Proposition 2.

$$\mathbb{P}(y = 1 | D_n, \mathbf{x}) = \int_{\gamma}^{+\infty} \mathbb{P}(y = 1 | f) p(f | D_n) df = \int_{\gamma}^{+\infty} \Phi(f) \phi(f; \mu, \sigma^2) df,$$

which we have already seen in (S4) equals $\Phi(a)$. □

For the proof of Theorem 1, we will introduce additional shorthand notation $f_q = f(\mathbf{x}_q)$, and as before will let $f_q | D_n \sim \mathcal{N}(\mu_q, \sigma_q^2)$. We let σ_q denote the covariance between f and f_q . We use the following result on the conditional distribution between f and f_q .

Lemma S2. *Let*

$$\begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}\right).$$

Then, the conditional density for f_1 given $f_2 = \gamma$ is

$$p(f_1 | f_2 = \gamma) = \frac{\phi\left(\frac{f_1 - \mu_1}{\sigma_1}\right)}{\sigma_1 \Phi\left(\frac{\gamma - \mu_2}{\sigma_2}\right)} \Phi\left(\frac{\gamma - \mu_2 - \frac{\sigma_{12}}{\sigma_1^2}(f_1 - \mu_1)}{\sqrt{\sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2}}}\right)$$

Proof. By Bayes' theorem,

$$p(f_1 | f_2 = \gamma) = \frac{p(f_2 = \gamma | f_1) p(f_1)}{p(f_2 = \gamma)} = \frac{p(f_2 = \gamma | f_1) \phi\left(\frac{f_1 - \mu_1}{\sigma_1}\right)}{\sigma_1 \Phi\left(\frac{\gamma - \mu_2}{\sigma_2}\right)}. \quad (\text{S7})$$

It is well-known that

$$p(f_2 | f_1 = x) = N(\mu_{2|1}, \sigma_{2|1})$$

where $\mu_{2|1} = \mu_2 + \frac{\sigma_{12}^2}{\sigma_1^2}(x - \mu_1)$ and $\sigma_{2|1}^2 = \sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2}$. Thus,

$$p(f_2 = \gamma | f_1) = \Phi\left(\frac{\gamma - \mu_{2|1}}{\sigma_{2|1}}\right) = \Phi\left(\frac{\gamma - \mu_2 - \frac{\sigma_{12}^2}{\sigma_1^2}(f_1 - \mu_1)}{\sqrt{\sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2}}}\right),$$

which when plugged into (S7) produces the result. \square

Proof of Theorem 1. By Bayes' theorem, we have that

$$\begin{aligned} \pi(\mathbf{x}_q | D_{n+1}(\mathbf{x}, y = 1)) &= P(f_q = \gamma | D_n, \mathbf{x}, y = 1) \\ &= \frac{P(y = 1 | D_n, \mathbf{x}, f_q = \gamma) P(f_q = \gamma | D_n)}{P(y = 1 | D_n, \mathbf{x})}. \end{aligned}$$

From Proposition 2 we know the denominator equals $\Phi(a)$ and can easily compute $P(f_q = \gamma | D_n) = \Phi\left(\frac{\gamma - \mu_q}{\sigma_q}\right) = \Phi(b_q)$, so the only term remaining is $P(y = 1 | D_n, \mathbf{x}, f_q = \gamma)$.

$$\begin{aligned} P(y = 1 | D_n, \mathbf{x}, f_q = \gamma) &= \int_1^{+1} P(y = 1 | f) p(f | D_n, f_q = \gamma) df \\ &= \int_1^{+1} \Phi(f) p(f | D_n, f_q = \gamma) df \\ &= \frac{1}{\sigma} \Phi\left(\frac{\gamma - \mu_q}{\sigma_q}\right) \int_1^{+1} \Phi(f) \Phi\left(\frac{\gamma - \mu_q - \frac{\sigma_q}{\sigma^2}(f - \mu)}{\sqrt{\sigma_q^2 - \frac{\sigma_q^2}{\sigma^2}}}\right) \phi\left(\frac{f - \mu}{\sigma}\right) df \end{aligned} \quad (\text{S8})$$

$$\begin{aligned} &= \frac{1}{\Phi\left(\frac{\gamma - \mu_q}{\sigma_q}\right)} \int_1^{+1} \Phi(\mu + \sigma \tilde{f}) \Phi\left(\frac{\gamma - \mu_q - \frac{\sigma_q}{\sigma} \tilde{f}}{\sqrt{\sigma_q^2 - \frac{\sigma_q^2}{\sigma^2}}}\right) \phi(\tilde{f}) d\tilde{f} \\ &= \frac{1}{\Phi\left(\frac{\gamma - \mu_q}{\sigma_q}\right)} \text{BvN}\left(\frac{\mu}{\sqrt{1 + \sigma^2}}, \frac{\gamma - \mu_q}{\sqrt{\sigma_q^2 - \frac{\sigma_q^2}{\sigma^2}} \sqrt{1 + \frac{\sigma_q^2}{\sigma^2 \sigma_q^2}}}; \frac{\rho \frac{\sigma}{\sigma^2 \sigma_q^2} \frac{\sigma_q}{\sigma_q}}{\sqrt{1 + \sigma^2} \sqrt{1 + \frac{\sigma_q^2}{\sigma^2 \sigma_q^2}}}\right) \end{aligned} \quad (\text{S9})$$

$$= \frac{1}{\Phi\left(\frac{\gamma - \mu_q}{\sigma_q}\right)} \text{BvN}\left(\frac{\mu}{\sqrt{1 + \sigma^2}}, \frac{\gamma - \mu_q}{\sigma_q}; \frac{\sigma_q}{\sigma_q \sqrt{1 + \sigma^2}}\right).$$

Here (S8) used Lemma S2, and (S9) used (S3) with $a = \mu$, $b = \sigma$, $h = \frac{\gamma - \mu_q}{\sqrt{\sigma_q^2 - \frac{\sigma_q^2}{\sigma^2}}}$, and $k = \rho \frac{\sigma}{\sigma^2 \sigma_q^2} \frac{\sigma_q}{\sigma_q}$.

Combining this term with the other terms, and using the convenient definitions of a and b_q , we have that

$$\pi(\mathbf{x}_q | D_{n+1}(\mathbf{x}, y = 1)) = \frac{1}{\Phi(a)} \text{BvN}\left(a, b_q; \frac{\sigma_q}{\sigma_q \sqrt{1 + \sigma^2}}\right).$$

For the $y = 0$, case,

$$\pi(\mathbf{x}_q | D_{n+1}(\mathbf{x}, y = 0)) = \frac{P(y = 0 | D_n, \mathbf{x}, f_q = \gamma) P(f_q = \gamma | D_n)}{P(y = 0 | D_n, \mathbf{x})}$$

The terms are easily computed from what we have already found: $P(y = 0 | D_n, \mathbf{x}) = 1 - \Phi(a)$, and $P(y = 0 | D_n, \mathbf{x}, f_q = \gamma) = 1 - \frac{1}{\Phi(b_q)} \text{BvN}\left(a, b_q; \frac{\rho \sigma_q}{\sigma_q \sqrt{1 + \sigma^2}}\right)$. Plugging these in yields the result in the Theorem. \square

S2 ACQUISITION EXPRESSIONS

Here we provide the full expression used to compute each look-ahead acquisition function, using the same posterior short-hand notation as in the previous section.

Global SUR

$$\alpha_{\text{GlobalSUR}}(\mathbf{x}) = \sum_{\mathbf{x}_q \in \mathcal{Z}_G} (\min(\Phi(b_q), 1 - \Phi(b_q)) \min(Z_q, \Phi(a) - Z_q) \min(\Phi(b_q) - Z_q, \Phi(a) - \Phi(b_q) - Z_q)).$$

Localized SUR

$$\alpha_{\text{LocalSUR}}(\mathbf{x}) = \min(\Phi(b), 1 - \Phi(b)) \min(Z, \Phi(a) - Z) \min(\Phi(b) - Z, \Phi(a) - \Phi(b) - Z).$$

Localized MI

$$\alpha_{\text{LocalMI}}(\mathbf{x}) = H_b(\Phi(b)) - \Phi(a) H_b\left(\frac{Z}{\Phi(a)}\right) + \Phi(a) H_b\left(\frac{\Phi(b) - Z}{\Phi(a)}\right).$$

Global MI

$$\alpha_{\text{GlobalMI}}(\mathbf{x}) = \sum_{\mathbf{x}_q \in \mathcal{Z}_G} \left(H_b(\Phi(b_q)) - \Phi(a) H_b\left(\frac{Z_q}{\Phi(a)}\right) + \Phi(a) H_b\left(\frac{\Phi(b_q) - Z_q}{\Phi(a)}\right) \right).$$

EAVC

$$\alpha_{\text{EAVC}}(\mathbf{x}) = \Phi(a) \left| \sum_{\mathbf{x}_q \in \mathcal{Z}_G} \left(\Phi(b_q) - \frac{Z_q}{\Phi(a)} \right) \right| + \Phi(a) \left| \sum_{\mathbf{x}_q \in \mathcal{Z}_G} \left(\Phi(b_q) - \frac{\Phi(b_q) - Z_q}{\Phi(a)} \right) \right|.$$

S3 ADDITIONAL BENCHMARK EXPERIMENT RESULTS

S3.1 Synthetic Functions

The synthetic functions were designed to explore a variety of input and output patterns that are present in real LSE problems, and in psychophysics problems in particular. A common experimental paradigm in psychophysics is the *two-alternative forced choice* (2AFC) method in which the participant is given two options and forced to select one. The CSF study in Section 6, and illustrated in Fig. 4, is an example of a 2AFC task. For 2AFC tasks, the minimum probability of being correct is 0.5, because participants are forced to make a choice and in the absence of a detectable stimulus will guess randomly. Thus the probability output space is $[0.5, 1]$, and the goal in these experiments is typically to find the $\theta = 0.75$ threshold, as is done in our experiment. However, there are other experimental designs, and other Bernoulli LSE tasks, in which the probability of success will vary from 0 to 1, and so to show that the methods are not limited to the 2AFC setting, we designed the Binarized Hartmann6 function to have probabilities from 0 to 1, and there set the target threshold to $\theta = 0.5$. We now give the functional form for each synthetic function.

S3.1.1 Binarized Hartmann6 Function

The Binarized Hartmann6 function was a binarization of the modified Hartmann 6-d function used by Lyu et al. (2021). Their modified Hartmann 6-d function is:

$$h(\mathbf{x}) = 1 - \sum_{i=1}^4 \alpha_i \exp\left(-\sum_{j=1}^6 A_{ij}(x_j - P_{ij})^2\right)$$

with $\alpha = [2.0, 2.2, 2.8, 3.0]$,

$$A = \begin{pmatrix} 8 & 3 & 10 & 3.5 & 1.7 & 6 \\ 0.5 & 8 & 10 & 1.0 & 6 & 9 \\ 3 & 3.5 & 1.7 & 8 & 10 & 6 \\ 10 & 6 & 0.5 & 8 & 1.0 & 9 \end{pmatrix}, \text{ and } P = 10^{-4} \begin{pmatrix} 1312 & 1696 & 5569 & 124 & 8283 & 5886 \\ 2329 & 4135 & 8307 & 3736 & 1004 & 9991 \\ 2348 & 1451 & 3522 & 2883 & 3047 & 6650 \\ 4047 & 8828 & 8732 & 5743 & 1091 & 381 \end{pmatrix}.$$

We used $f(\mathbf{x}) = 3h(\mathbf{x}) - 2$ as the ground-truth latent function, so that Bernoulli samples were simulated according to $\Phi(f(\mathbf{x}))$. The input space for this problem is as in the classic Hartmann6 problem, $\mathbf{x} \in [0, 1]^6$. The output probabilities, $\Phi(f(\mathbf{x}))$, span $[0, 1]$, so for this problem the target threshold was set as $\theta = 0.5$.

S3.1.2 Psychophysical Discrimination, 2-d

The latent function for this problem is computed as

$$f(x_1, x_2) = \frac{1 + x_2}{0.05 + 0.4x_1^2(0.2x_1 - 1)^2}.$$

The input domain is $x_1, x_2 \in [-1, 1]$, and the output probabilities span $[0.5, 1]$, with the target threshold $\theta = 0.75$.

S3.1.3 Psychophysical Discrimination, 8-d

For $\mathbf{x} = [x_1, \dots, x_8]$, we define

$$c(\mathbf{x}) = \left(\frac{x_3}{2} \left(1 + \cos \left(\frac{3}{5} \pi x_2 x_8 + x_7 \right) \right) + x_4 \right) \left(2 - x_6 \left(1 + \sin \left(\frac{3}{10} \pi x_2 x_8 + x_7 \right) \right) \right) - 1.$$

Then, the Bernoulli probability for the 8-d Psychophysical Discrimination function is computed as

$$z(\mathbf{x}) = \frac{1}{2} + \frac{1}{2} \Phi \left(\frac{x_1 - c(\mathbf{x})}{x_5(2 + c(\mathbf{x}))} \right).$$

The input space is $\mathbf{x} \in [-1, 1]^8$ and, as in the 2-d discrimination function, the output probabilities span $[0.5, 1]$, so the target threshold was set to 0.75.

S3.2 Surrogate Model and Acquisition Optimization

All methods and all experiments used the same surrogate model: A typical variational classification GP (Hensman et al., 2015) with 100 inducing points and an RBF kernel. Kernel hyperpriors were taken as the defaults from the Botorch package (Balandat et al., 2020). Inducing points were selected by applying k-means to the observations. In each iteration of active learning, the model was updated with the new observation by refitting the variational distribution and kernel hyperparameters. In most iterations, the refitting was warm-started by beginning the fitting at the previous values. To avoid getting stuck in a local optimum, and as is common in Bayesian optimization, every 10th iteration the re-fitting was done from scratch with a refreshed set of inducing points.

To avoid conflating acquisition quality with the ability to optimize the acquisition function, all acquisition functions were optimized in the same manner, using the gradient-based acquisition optimization utilities from the Botorch package. For the global acquisition functions, the reference set G was taken as a quasi-random (scrambled Sobol) set of 500 points, which was changed for each iteration. Straddle in the probability space was evaluated as a Monte Carlo acquisition function (Balandat et al., 2020) due to the lack of a differentiable implementation of Owen’s T function.

Each benchmark run initialized with an initial design of 10 quasi-random points, after which the surrogate model was fit and all subsequent iterations used active sampling with the specified acquisition function. Throughout the active sampling, performance metrics were computed using a quasi-random test-set of 1000 points, which was sampled independently from anything done for the modeling or acquisition optimization.

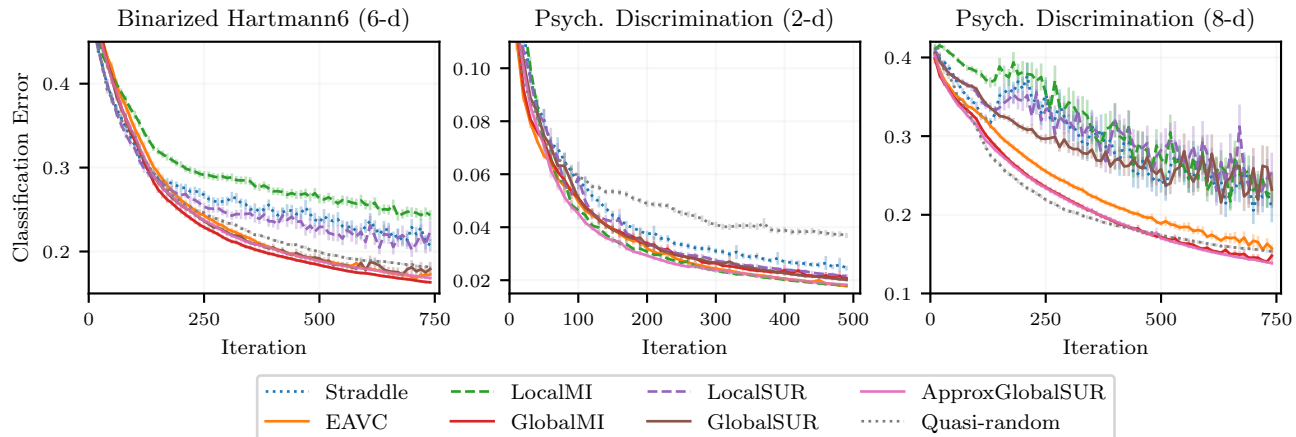


Figure S1: **Benchmark results: alternative metric.** Expected classification error (lower is better) of the GP surrogate model as a function of active sampling iteration, for the same benchmark results as in Fig. 3. As before, shown is the mean and two standard errors over 200 replications. On the high-dimensional problems, the localized look-ahead methods (LocalMI, LocalSUR) performed significantly worse than the quasi-random baseline. Global methods GlobalMI and EAVC were consistently the best.

S3.3 Additional Evaluation Metrics

The results in the main text showed performance evaluated using Brier score, a strictly proper scoring rule. Proper scoring rules are an appropriate evaluation metric for this problem space because they assess not only the quality of the model point prediction, but also the calibration of posterior uncertainty. An alternative metric for evaluating classifiers in particular is the expected classification error, defined as $p(1 - y) + (1 - p)y$ for a classifier that provides p as the probability that $y = 1$ (*i.e.*, that a point is below threshold), and y the actual outcome (*i.e.*, if the point was actually below threshold). Fig. S1 show the results of the benchmark experiments when evaluated using expected classification error. The conclusions of the experiments do not change under this alternative evaluation metric.

Fig. S2 shows the amount of wall time required to optimize the acquisition function for a model fit to 250 observations. Global acquisition functions required more wall-time to compute due to the global reference set G , however with a per-iteration time between half a second and a second for the complete optimization when given multiple threads, they were well within the speed required for experiments with human participants.

S3.4 Comparison to BALD and BALV

The main text discusses how the use of global active sampling methods such as BALD (Houlsby et al., 2011) or Bayesian active learning by variance (BALV) (Song et al., 2015) can be inefficient for level-set estimation because they may focus sampling effort on reducing variance in areas that are not close to the threshold. Fig. S3 shows empirically that this is the case, by evaluating BALD and BALV on the same benchmark problems used in the main text. For LSE in high dimensions, BALD and BALV performed significantly worse than quasi-random search, which further emphasizes the importance of developing acquisition functions specifically for LSE.

S3.5 An Analysis of Edge Sampling Behavior

We highlighted in the main text that a source of poor performance for localized look-ahead methods is their tendency to oversample edge locations. This behavior is shown empirically in Fig. S4, using the benchmark results from Section 5. For each benchmark run, we evaluated the proportion of active learning samples that were within 5% of the search space range of an edge. For instance, on the Binarized Hartmann6 problem where the domain is $[0, 1]^6$, this was the proportion of points with an element less than 0.05 or greater than 0.95. In high dimensions, the localized look-ahead methods LocalMI and LocalSUR, along with the straddle acquisition, sampled significantly more edge locations than the global look-ahead methods, or quasi-random sampling. LocalMI was particularly focused on the edges, with 99% edge samples for Binarized Hartmann6.

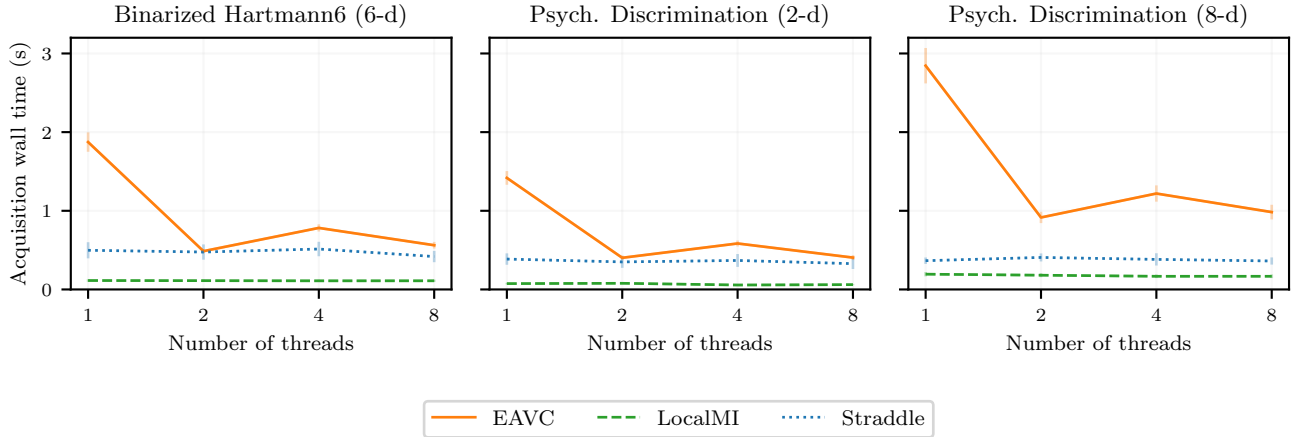


Figure S2: **Benchmark results: wall time.** Wall time required for acquisition optimization, in seconds, based on a surrogate model fit to 250 observations drawn from a Sobol sequence. This evaluation was done on an AWS EC2 c61.18xlarge instance and is the average of 20 replications. Global acquisition methods required less than a second per iteration to identify the next point for evaluation.

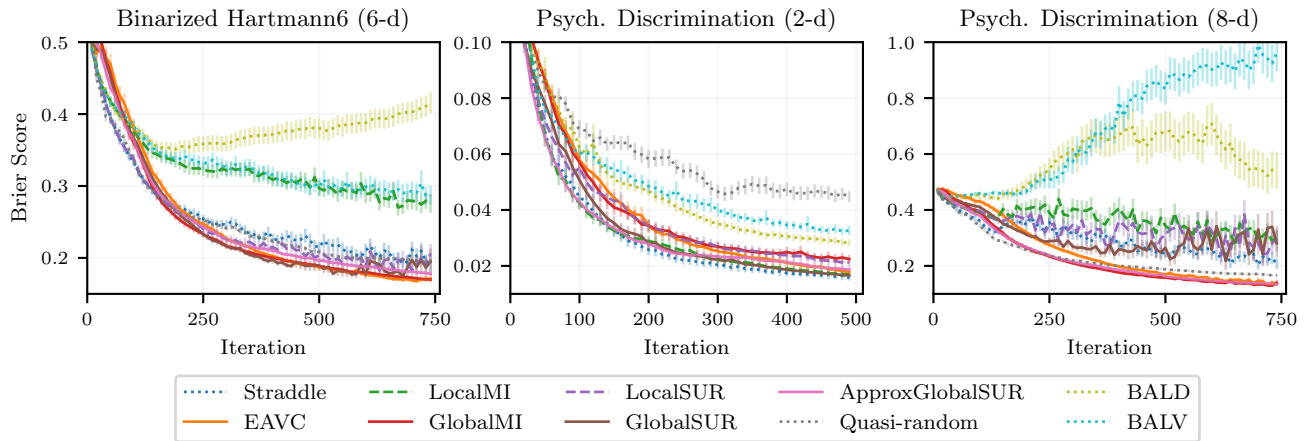


Figure S3: **Benchmark results: BALD.** The same benchmark results as in Fig. 3, with the addition of global active sampling methods BALD and BALV. Global active sampling methods can waste samples reducing uncertainty in areas far from the threshold, and here they performed worse than quasi-random search for LSE.

EAVC had the least tendency to sample edges, and in high dimensions had an edge sampling rate comparable to quasi-random search.

S3.6 Sensitivity Study

The look-ahead acquisition functions described and developed in this paper do not have any hyperparameters that must be tuned. The straddle acquisition function has the β hyperparameter, and Lyu et al. (2021) have done a sensitivity study of various strategies for selecting β and found that none consistently performed well. Here we study sensitivity to two aspects of the experiments: the initial design, and the target threshold.

S3.6.1 Initial Design Sensitivity

Each benchmark run in the results of Section 5 was initialized with 10 quasi-random points. We saw in Fig. S4 that localized look-ahead methods over-sampled edge locations, and hypothesized that a larger initial design would provide a better initial global surrogate model, under which the high degree of exploitation in the localized look-ahead methods could actually be beneficial.

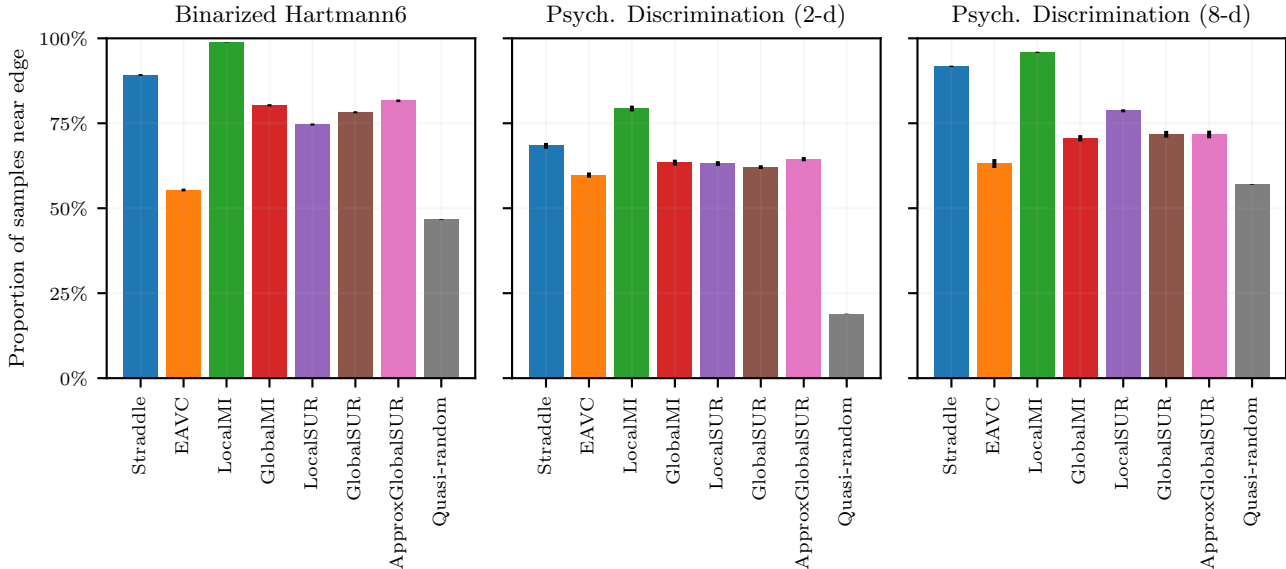


Figure S4: **Benchmark results: edge sampling.** For the benchmark results from Fig. 3, the proportion of active learning iterations in which the evaluated point was within 5% of an edge of the search space. Like the Straddle acquisition, localized look-ahead methods did significantly more edge sampling than global look-ahead methods in high-dimensions.

Fig. S5 shows the final Brier score after 750 total iterations, for increasingly large initial designs. Note that the total number of iterations was fixed at 750, so that the initial design of size 10 had 740 active sampling iterations, while that of size 500 had only 250 active sampling iterations. The sensitivity study focuses on a subset of methods (EAVC, LocalMI, and GlobalMI) that were most characteristic. As hypothesized, LocalMI benefited significantly from having a larger initial design, and the improved global surrogate that a larger initial design entails. Particularly on the Binarized Hartmann6 problem, LocalMI performed significantly worse than quasi-random sampling with the small initial design of 10 points, but with an initial design of 250 (out of 750) points, was able to do slightly better than quasirandom. While LocalMI did perform better with a larger initial design on all three high-dimensional problems, it still did not match the performance of the global methods. The global methods GlobalMI and EAVC, in contrast, performed best for the smallest initial designs, which permitted the most active sampling. Generally, performance with the global methods was robust to the size of the initial design up to 250 points (one third of the total budget).

S3.6.2 Target Threshold Sensitivity

Fig. S6 shows how the final Brier score varies with the target threshold for the problem. Changing the target threshold significantly alters the LSE problem, by focusing the active sampling in a different part of the search space. LSE performance thus changes when the target threshold is changed, however Fig. S6 shows that across this large set of target level sets, the global look-ahead methods continued to consistently be the best.

S4 DETAILS OF REAL-WORLD EXPERIMENT

For the real-world CSF experiment, the stimulus feature space was 8-dimensional and we collected 1000 stimuli generated from a Sobol sequence over that stimulus space. We used a GP classification surrogate model as ground truth. The model was fit using an RBF kernel over six of the stimulus features to create a 6-d problem space. Two stimulus properties, angular dimensions of eccentricity and orientation, were left unmodeled. This effectively added noise to the surrogate function and increased the difficulty of level-set estimation. The model was a variational classification GP with training locations used for inducing points. The GP mean was used for the ground-truth latent f from which Bernoulli responses were simulated.

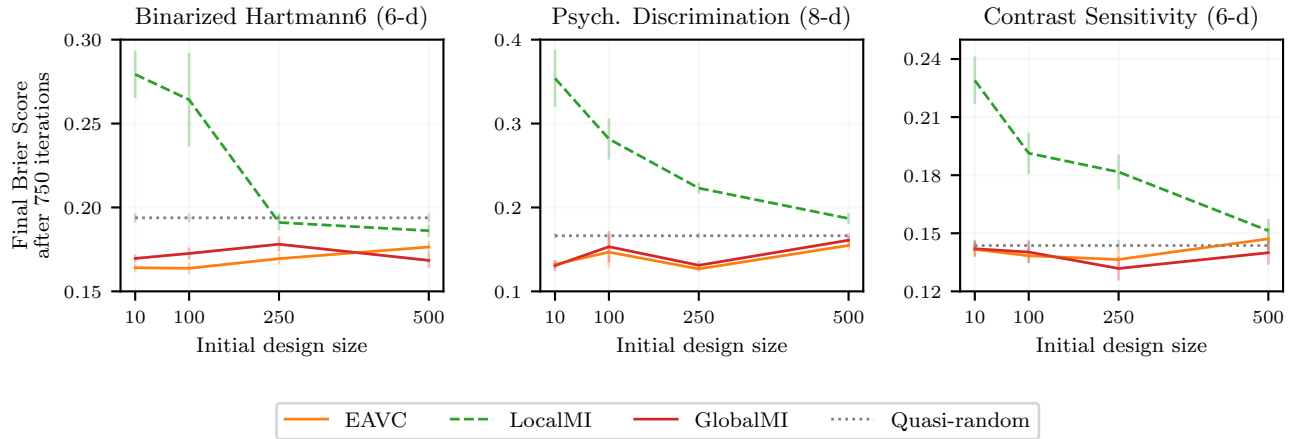


Figure S5: **Sensitivity study: initial design.** The Brier score (mean and two standard errors) after 750 total iterations (initial design plus active sampling) as a function of the size of the initial design. Global methods performed better when given more iterations of active sampling (smaller initialization). LocalMI benefited from a larger initialization, but never achieved the best performance of global methods on the Binarized Hartmann6 and Psychophysical Discrimination (8-d) problems.

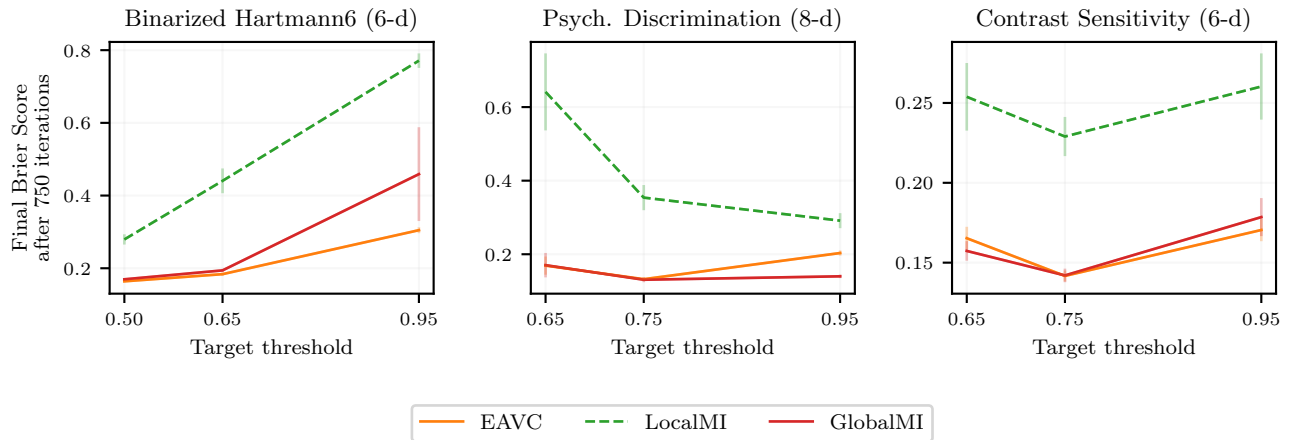


Figure S6: **Sensitivity study: target threshold.** The Brier score (mean and two standard errors) after 750 iterations when changing the target threshold. Changing the target threshold effectively changes the LSE problem. While some target thresholds pose more challenging tasks than others, across this range of settings the global look-ahead methods continued to generally perform best.