H-Entropy Search: Generalizing Bayesian Optimization with a Decision-theoretic Uncertainty Measure

Anonymous authors
Paper under double-blind review

Abstract

Bayesian optimization (BO) is a popular method for efficiently inferring optima of an expensive black-box function via a sequence of queries. Existing information-theoretic BO procedures aim to make queries that most reduce the uncertainty about optima, where the uncertainty is captured by Shannon entropy. However, an optimal measure of uncertainty would, ideally, factor in how we intend to use the inferred quantity in some downstream procedure. In this paper, we instead consider the H-entropy, a generalization of Shannon entropy from work in statistical decision theory (DeGroot, 1962; Rao, 1984), which contains a broad class of uncertainty measures parameterized by a problem-specific loss function corresponding to a downstream task. We first show that special cases of the H-entropy lead to popular acquisition functions used in BO procedures such as knowledge gradient, expected improvement, and entropy search. We then show how alternative choices for the loss yield a flexible family of acquisition functions for a variety of specialized optimization tasks, including variants of top-k estimation, level set estimation, and multi-valued search. For special cases of the loss and design space, we develop gradient-based methods to efficiently optimize our proposed family of acquisition functions, and demonstrate that the resulting BO procedure shows strong empirical performance on a diverse set of optimization tasks.

1 Introduction

A popular class of methods for global optimization of a black-box function $\mathbf{f}$ over a design space is information-based Bayesian optimization (BO), which includes the family of *entropy search* methods (Hennig & Schuler, 2012; Hernández-Lobato et al., 2014; Wang & Jegelka, 2017). As in other BO procedures, entropy search methods leverage a probabilistic model of the function, $p(f)$, to select pointwise queries of $\mathbf{f}$ for sample-efficient optimization. Specifically, at each iteration t, these methods query $\mathbf{f}(x_t)$, where $x_t \in \mathcal{X} \subset \mathbb{R}^d$ is the design that is expected to yield the largest reduction in the Shannon entropy of the posterior distribution over the optimal design $x^* = \arg\max_{x\in\mathcal{X}} f(x)$, yielding the query selection criterion

$$x_t = \underset{x\in\mathcal{X}}{\arg\max}\ \mathcal{H}\left[x^* \mid \mathcal{D}_t\right] - \mathbb{E}_{p(y_x|\mathcal{D}_t)}\left[\mathcal{H}\left[x^*|\mathcal{D}_t \cup \{(x, y_x)\}\right]\right], \tag{1}$$

where $\mathcal{H}\left[x^* \mid \mathcal{D}_t\right]$ is the differential Shannon entropy of the posterior distribution $p(x^* \mid \mathcal{D}_t)$ induced by the probabilistic model, $p(y_x \mid \mathcal{D}_t)$ is the posterior predictive distribution at a point $x \in \mathcal{X}$, and $\mathcal{D}_t = \{(x_i, y_{x_i})\}_{i=1}^{t-1}$ is a dataset of observations. Intuitively, this criterion queries an input x_t that is expected to most reduce the uncertainty of $p(x^* \mid \mathcal{D}_t)$.

Shannon entropy is one measure of uncertainty that we could aim to reduce at each iteration of Bayesian optimization—however, it is not the only measure, and it is not necessarily the most-ideal measure for every optimization task. An optimal uncertainty function would, ideally, factor in how we intend to use an inferred quantity (and our uncertainty about it) in some downstream procedure. For example, instead of reducing the Shannon entropy of $p(x^* \mid \mathcal{D}_t)$, we could aim to shrink the posterior uncertainty with repect to how x^* is used downstream, such that the posterior expected downstream performance is maximized. Furthermore, there are many optimization variants where we are interested in reducing the uncertainty about more-complex quantities beyond x^*, such as in the tasks of level set estimation, multi-objective optimization, and top-k estimation.

In this paper, we instead consider the *H-entropy*, a generalization of Shannon entropy from work in statistical decision theory (DeGroot, 1962; Rao, 1984; Grünwald & Dawid, 2004), which contains a broad class of uncertainty measures for a posterior distribution over f, parameterized by a problem-specific loss function and set of actions. For a given optimization task, we can define a loss and action set tailored to the downstream use of the posterior $p(f \mid \mathcal{D}_t)$ after queries are complete. Specifically, we will assume that, after completion of the BO procedure, we will take the *Bayes action* with respect to $p(f \mid \mathcal{D}_t)$—i.e. the member of the action set that minimizes the posterior expected loss. Then our H-entropy-based acquisition function can be viewed as choosing an x_t to query on $\mathbf{f}$ which most reduces the posterior expected loss of the Bayes action.

We will show that special cases of our acquisition function are equivalent to the entropy search acquisition functions in Eq. (1), as well as other popular acquisition functions used in BO, such as the knowledge gradient (Frazier et al., 2009) and expected improvement (Močkus, 1975; Jones et al., 1998) functions. Moreover, we show that alternative choices for the loss and action set provide a flexible family of acquisition functions for a variety of specialized optimization tasks. For example, we define special cases of our acquisition function for the tasks of *top-k estimation with diversity* (estimating a set of k optimal designs under a penalty that encourages diversity), *generalized level set estimation* (partitioning the design space $\mathcal{X}$ based on binned function value), and *multi-value sequence search* (estimating a sequence of points with multiple prescribed values under f).

Finally, we present a full BO procedure, H-ENTROPY SEARCH, and show that it is computationally feasible for a large class of tasks (including each of the examples above). We design a general recipe for gradient-based optimization of the acquisition function, and show how to implement the computation techniques using automatic differentiation (Balandat et al., 2020; Paszke et al., 2019) to accelerate acquisition function optimization. In summary, we provide the following contributions:

- We introduce a family of acquisition functions based on H-entropy, parameterized by a loss function ℓ and action set $\mathcal{A}$. We show that choices of ℓ and $\mathcal{A}$ yield acquisition functions used in popular BO methods such as expected improvement, knowledge gradient, and entropy search.
- By customizing ℓ and $\mathcal{A}$ for a given optimization task, we derive new acquisition functions for specialized optimization settings, including top-k estimation with diversity, generalized level set estimation, and multi-value sequence search.
- Under certain conditions on ℓ, $\mathcal{A}$, and the design space $\mathcal{X}$, we provide gradient-based methods to optimize our proposed acquisition function.
- We demonstrate that H-ENTROPY SEARCH, using gradient-based acquisition optimization, shows strong empirical performance on a diverse set of optimization tasks.

2 SETUP

Let $\mathbf{f} : \mathcal{X} \to \mathcal{Y}$ denote an expensive black-box function that maps from an input search space $\mathcal{X}$ to an output space $\mathcal{Y}$, and where $\mathbf{f} \in \mathcal{F}$. We assume that we can evaluate $\mathbf{f}$ at an input $x \in \mathcal{X}$, and will observe a noisy function value $y_x = \mathbf{f}(x) + \epsilon$, where ϵ is drawn from some noise distribution.

We will assume that, at some point after the optimization procedure, we intend to take an action a from some set of actions $\mathcal{A}$, and then incur some loss based on both this action a and the function $\mathbf{f}$. We denote this loss as $\ell : \mathcal{F} \times \mathcal{A} \to \mathbb{R}$. As one example, after the BO procedure, we may be allowed to make a single guess x^* of the function maximizer, and then incur a loss based on the value of the function at x^*. In this case, the action set is $\mathcal{A} = \mathcal{X}$ and the loss is $\ell(\mathbf{f}, a) = \ell(\mathbf{f}, x^*) = -\mathbf{f}(x^*)$.

We will also assume that our uncertainty about $\mathbf{f}$ is captured by a probabilistic model with prior distribution $p(f)$, which reflects our prior beliefs about $\mathbf{f}$. Given a dataset set of observed function evaluations $\mathcal{D}_t = \{(x_i, y_{x_i})\}_{i=1}^{t-1}$, our model gives a posterior distribution over $\mathcal{F}$, denoted $p(f|\mathcal{D}_t)$.

3 H-ENTROPY SEARCH

We first define the H-entropy, a decision-theoretic notion of uncertainty, which is parameterized by a problem-specific action set $\mathcal{A}$ and loss function ℓ. This uncertainty measure has been introduced previously as a generalization of Shannon entropy (DeGroot, 1962; Rao, 1984; Grünwald & Dawid, 2004). We adopt the phrase *H-entropy* and symbol H, which were both used by Rao (1982; 1984; 1987) to refer to this family of entropy functionals.

Definition 1. *(H-Entropy of f). Given a prior distribution $p(f)$ on functions, and a dataset $\mathcal{D}_t$ of observed function evaluations, the posterior H-entropy with loss ℓ and action set $\mathcal{A}$ is defined to be*

$$H_{\ell,\mathcal{A}}\left[f \mid \mathcal{D}_t\right] = \inf_{a \in \mathcal{A}} \mathbb{E}_{p(f|\mathcal{D}_t)}\left[\ell(f, a)\right]. \tag{2}$$

Intuitively, suppose that we must make a decision by choosing an action $a \in \mathcal{A}$, where this action incurs a loss $\ell(a, f)$ defined by the loss function ℓ and true function f. Given a posterior distribution $p(f \mid \mathcal{D}_t)$ that describes our belief about f after observing $\mathcal{D}_t$, the *Bayes action* $a^* \in \mathcal{A}$ is the action that minimizes the posterior expected loss, i.e. $a^* = \arg\inf_{a\in\mathcal{A}} \mathbb{E}_{p(f|\mathcal{D}_t)}\left[\ell(f, a)\right]$. The H-entropy can then be viewed as the posterior expected loss of the Bayes action.

We propose a family of acquisition functions for BO using this H-entropy, which are similar in structure to information-theoretic acquisition functions such as entropy search (ES) (Hennig & Schuler, 2012), predictive entropy search (PES) (Hernández-Lobato et al., 2014), and max-value entropy search (MES) (Wang & Jegelka, 2017). Our family of acquisition functions are designed to select the query $x_t \in \mathcal{X}$ that maximally reduces the uncertainty, as characterized by the H-entropy, in expectation. We refer to this quantity as the *expected H-information gain* (EHIG).

Definition 2. *(Expected H-Information Gain). Given a prior distribution $p(f)$ on functions and a dataset of observed function evaluations $\mathcal{D}_t$, the expected H-information gain (EHIG), with loss ℓ and action set $\mathcal{A}$, is defined to be*

$$\textit{EHIG}_t(x; \ell, \mathcal{A}) = H_{\ell,\mathcal{A}}\left[f \mid \mathcal{D}_t\right] - \mathbb{E}_{p(y_x|\mathcal{D}_t)}\left[H_{\ell,\mathcal{A}}\left[f \mid \mathcal{D}_t \cup \{(x, y_x)\}\right]\right]. \tag{3}$$

We note that EHIG is similar in structure to the entropy search acquisition function given in Eq. (1): the Shannon entropy $\mathcal{H}[x^* \mid \mathcal{D}_t]$ in (1) is simply replaced by the H-entropy $H_{\ell,\mathcal{A}}[f \mid \mathcal{D}_t]$. However, there are a couple of notable differences arising from this substitution. First, H-entropy characterizes the uncertainty of $p(f \mid \mathcal{D}_t)$, rather than $p(x^* \mid \mathcal{D}_t)$. We will show how this can both generalize the entropy search acquisition function (as well as a few other popular acquisition functions used in BO), and also allow us to tailor our acquisition function to infer quantities beyond x^*, which is desirable in optimization variants (e.g. levelset estimation, multi-objective optimization, and top-k estimation). Second, while the entropy search acquisition function uses the Shannon entropy to describe uncertainty about the inferred quantity, the H-entropy describes uncertainty using a problem-specific loss, which can be tailored to the downstream use of the inferred quantity.

We present H-ENTROPY SEARCH, our full Bayesian optimization procedure using the EHIG acquisition function, in Algorithm 1. This procedure takes as input a loss ℓ, action set $\mathcal{A}$, and prior model $p(f)$. At each iteration, the procedure optimizes $\text{EHIG}_t(x; \ell, \mathcal{A})$ to select a design $x_t \in \mathcal{X}$ to query, and then evaluates the black-box function on this design to observe an outcome $y_{x_t} \sim \mathbf{f}(x_t) + \epsilon$. In Section 6 we will describe methods for optimizing the EHIG acquisition function, including gradient-based procedures for certain cases of $\mathcal{X}$, $\mathcal{A}$, and ℓ.

Algorithm 1 H-ENTROPY SEARCH

```
Input: dataset $\mathcal{D}_1$, prior distribution $p(f)$, action set $\mathcal{A}$, loss $\ell$.
for $t = 1, \ldots, T$ do
    $x_t \leftarrow \arg\max_{x\in\mathcal{X}} \text{EHIG}_t(x; \ell, \mathcal{A})$        ▷ Optimize expected $H$-information gain
    $y_{x_t} \sim \mathbf{f}(x_t) + \epsilon$        ▷ Evaluate $\mathbf{f}$ at $x_t$
    $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{(x_t, y_{x_t})\}$        ▷ Update dataset
Output: distribution $p(f \mid \mathcal{D}_{T+1})$
```

4 EXISTING ACQUISITION FUNCTIONS FOR BO AS SPECIAL CASES

We next show how popular acquisition functions developed for BO are special cases of the proposed EHIG acquisition function family given in Eq. (3), for particular choices of ℓ and $\mathcal{A}$. In particular, we will show this for the knowledge gradient (Frazier et al., 2009), entropy search (Hennig & Schuler, 2012; Hernández-Lobato et al., 2014; Wang & Jegelka, 2017), and expected improvement (Močkus, 1975; Jones et al., 1998) acquisition functions. To do so, we will view each acquisition function from a decision-theoretic perspective: after the BO procedure is complete, we must make some decision and then incur a loss, and we want to make a sequence of queries that reduce the uncertainty of the posterior distribution over f in order to help best make this decision with low loss.

Knowledge gradient (KG). The knowledge gradient (KG) acquisition function can be written

$$\mathrm{KG}_t(x) = \mathbb{E}_{p(y_x|\mathcal{D}_t)}\left[\mu^*_{t+1}(x, y_x)\right] - \mu^*_t \tag{4}$$

where $\mu^*_t = \sup_{x' \in \mathcal{X}} \mathbb{E}_{p(f|\mathcal{D}_t)}\left[f(x')\right]$ is the maximum value of the posterior mean of f given data $\mathcal{D}_t$, and $\mu^*_{t+1}(x, y_x) = \sup_{x' \in \mathcal{X}} \mathbb{E}_{p(f|\mathcal{D}_t \cup \{(x,y_x)\})}\left[f(x')\right]$ is the maximum value of the posterior mean of f, given both data $\mathcal{D}_t$ and observation (x, y_x).

It is straightforward to view this acquisition function as a special case of EHIG_t. Suppose, after the BO procedure is complete, we must make a guess x^* for the maximizer of f, and then incur a loss equal to the value of the function at x^*. In this case, we can view the action set as $\mathcal{A} = \mathcal{X}$, and the loss function as $\ell(f, a) = \ell(f, x) = -f(x)$. Note that the Bayes action will then be equal to the maximizer of the posterior mean, the H-entropy will be equal to $-\mu^*_t$, and thus the EHIG_t will equal KG_t. We formalize this in the following proposition.

Proposition 1. *If we choose $\mathcal{A} = \mathcal{X}$ and $\ell(f, x) = -f(x)$, then the* EHIG *is equivalent to the knowledge gradient acquisition function, i.e.* $\mathrm{EHIG}_t(x; \ell, \mathcal{A}) = \mathrm{KG}_t(x)$.

Proof of Proposition 1. The proof is given in the appendix. □

Entropy search (ES, PES, MES). We will restate the entropy search acquisition function given in Eq. (1) to include a broader family of information-based BO objectives. Let $\theta_f \in \Theta$ denote a property of f that we would like to infer. For example, we could set $\theta_f = \arg\max_{x \in \mathcal{X}} f(x) = x^* \in \mathcal{X}$, i.e. the location of the global optimizer of f, or $\theta_f = \max_{x \in \mathcal{X}} f(x) \in \mathbb{R}$, i.e. the maximum value achieved by f in $\mathcal{X}$. This generalized entropy search acquisition function can then be written

$$\mathrm{ES}_t(x) = \mathcal{H}\left[\theta_f \mid \mathcal{D}_t\right] - \mathbb{E}_{p(y_x|\mathcal{D}_t)}\left[\mathcal{H}\left[\theta_f \mid \mathcal{D}_t \cup \{(x, y_x)\}\right]\right], \tag{5}$$

where $\mathcal{H}[\theta_f \mid \mathcal{D}_t] = -\int p(\theta_f \mid \mathcal{D}_t) \log p(\theta_f \mid \mathcal{D}_t)\, df$ denotes the differential Shannon entropy of the induced posterior distribution over θ_f.

We can then view this acquisition function as a special case of EHIG_t in the following way. Suppose, after the BO procedure is complete, we must choose a distribution q from a set of distributions $\mathcal{P}(\Theta)$, and then we will incur a loss equal to the negative log-likelihood of q for the true value of θ_f. In this case, we view the action set as $\mathcal{A} = \mathcal{P}(\Theta)$ and the loss function as $\ell(f, a) = \ell(f, q) = -\log q(\theta_f)$. The H-entropy will then be equal to the Shannon entropy of θ, and thus the EHIG_t will be equal to ES_t. We formalize this in the following proposition.

Proposition 2. *If we choose $\mathcal{A} = \mathcal{P}(\Theta)$ and $\ell(f, q) = -\log q(\theta_f)$, then the* EHIG *is equivalent to the entropy search acquisition function, i.e.* $\mathrm{EHIG}_t(x; \ell, \mathcal{A}) = \mathrm{ES}_t(x)$.

Proof of Proposition 2. The proof is given in the appendix. □

Expected improvement (EI). The expected improvement (EI) acquisition function can be written

$$\mathrm{EI}_t(x) = \mathbb{E}_{p(f|\mathcal{D}_t)}\left[\max(0, f^*_t - f(x))\right]. \tag{6}$$

where we define $f^*_t = \max\{\hat{f}(x_i)\}_{i=1}^{t-1}$, for $x_i \in \mathcal{D}_t$, where $\hat{f}(x_i)$ is the posterior expected value of f at x_i. Note that this definition is equal to the standard formulation of EI in the noiseless setting (i.e. when $y_x = \mathbf{f}(x)$ for queried x) and is equal to the *plug-in* formulation of EI in the noisy setting, when $y_x = \mathbf{f}(x) + \epsilon$ (Picheny et al., 2013; Brochu et al., 2010).

We can then view this acquisition function as a special case of EHIG_t in the following way. Suppose, after the BO procedure is complete, we incur a loss based on the value of f at the best queried point $x^*_t \in \mathcal{D}_t$, where $x^*_t = \arg\max\{\hat{f}(x_i)\}_{i=1}^{t-1}$. In this case, we can use a time-dependent action set that depends on the previous queries, i.e. $\mathcal{A}_t = \{x_i\}_{i=1}^{t-1}$ and can define the loss function as $\ell(f, a) = \ell(f, x_i) = -f(x_i)$. The Bayes action will then be equal to x^*_t, the H-entropy will be equal to $-\hat{f}(x^*_t) = -f^*_t$, and the EHIG_t will be equal to EI_t. We formalize this as follows.

Proposition 3. *If we choose $\mathcal{A}_t = \{x_i\}_{i=1}^{t-1}$, where $x_i \in \mathcal{D}_t$, and $\ell(f, x_i) = -f(x_i)$, then the* EHIG *is equal to the expected improvement acquisition function, i.e.* $\mathrm{EHIG}_t(x; \ell, \mathcal{A}) = \mathrm{EI}_t(x)$.

Proof of Proposition 3. The proof is given in the appendix. □

Practical guidance and summary In summary, via our EHIG framework, we can view multiple acquisition functions (KG, ES, PES, MES, EI) from a decision-theoretic perspective, and provide guidance on which to choose based on the preferred loss in a given optimiztion scenario, namely:

- *Knowledge gradient (KG):* one should choose this if, after the BO procedure ends, one wants to know the point $x^* \in \mathcal{X}$, at which the value $f(x^*)$ is expected to be highest.
- *Entropy search on θ (ES, PES, MES):* one should choose this if, after the BO procedure ends, one wants a posterior distribution q over a function property θ_f, in which the log-likelihood $\log q(\theta_f)$ is expected to be highest.
- *Expected improvement (EI):* one should choose this if, after the BO procedure ends, one wants a dataset $\mathcal{D}_t$ in which the maximal queried value $\max_{x_i \in \mathcal{D}_t} f(x_i)$ is expected to be highest.

While the above methods are for standard BO, which focus on estimating a single optimal point under the black-box function, in the following sections we will show how our EHIG framework extends to a broader class of specialized optimization tasks.

5 Novel Acquisition Functions for Specialized Optimization

We use EHIG to derive novel acquisition functions for specialized optimization settings, including top-k estimation with diversity, (generalized) level set estimation, and multi-value sequence search.

Top-k estimation with diversity. For a discrete design space $\mathcal{X}$, the task of top-k estimation is to estimate the subset of $\mathcal{X}$ with size k that has the highest values under the black-box function $\mathbf{f}$; when the domain $\mathcal{X}$ is continuous, the task of *top-k estimation with diversity* aims to solve the constrained optimization problem: $\max_{\{x_i\}_{i=1}^k \in \mathcal{X}^k} \sum_{i=1}^k \mathbf{f}(x_i)$ such that $\forall i, j \in \{1, \dots, k\}$, $d(x_i, x_j) \geq c$, where $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is some predefined distance function on $\mathcal{X}$ (e.g. Euclidean distance) and c is a distance threshold to encourage diversity. This type of task arises, for example, in materials discovery (Liu et al., 2017), sensor networks (Abbasi et al., 2008), and medicine (Xie, 2018).

To approach this task in our EHIG framework, we design a loss using soft constraints, which are suitable for our continuous domain. In particular, we define the action set as $\mathcal{A} = \mathcal{X}^k$ (where an action $\mathbf{a} = (a_1, \dots, a_k) \in \mathcal{A}$ denotes a set of top-k points) and define the loss function as

$$\ell(f, \mathbf{a}) = -\sum_i f(a_i) - \sum_{1 \leq i < j \leq k} d(a_i, a_j). \tag{7}$$

Generalized level set estimation. The goal of level set estimation (LSE) is to estimate a subset of the design space $\mathcal{X}$, where function values are larger than a given threshold c, $\mathcal{S}_c = \{x \in \mathcal{X} : f(x) > c\}$. This task appears in a number of applications, including catalyst design (Zhong et al., 2020), interactive learning (Boecking et al., 2020), and environmental monitoring (Singh, 2008).

As an initial approach to LSE under our EHIG framework, we focus on carrying out LSE for a discrete subset of design points $\mathcal{X}_0 \subset \mathcal{X}$ of size J, i.e. $|\mathcal{X}_0| = J$. We then define the loss and action set using $\mathcal{X}_0$. The action set, defined as $\mathcal{A} = [0, 1]^J$, represents a set of weights associated with each element in $\mathcal{X}_0$, which can be interpreted as the confidence of an element belonging to the super-level set. We define the loss as

$$\ell(f, a(x)) = -\sum_{x \in \mathcal{X}_0} a(x) \left(f(x) - c\right). \tag{8}$$

such that at optimality $a(x) = 1$ for each $x \in \mathcal{X}_0$ with $f(x) > c$, and $a(x) = 0$ otherwise. Similarly, we can consider a generalized level set estimation problem, where we are given m thresholds satisfying $c_1 < \ldots < c_m$ and we are interested at estimating $m + 1$ level sets: $\mathcal{S}_i = \{x \in \mathcal{X} : c_i < f(x) < c_{i+1}\}$ for $i = \{0, \dots, m\}$ (where $c_0 := -\infty$ and $c_{m+1} := +\infty$). In this case, we define the action set to be $\mathcal{A} = [0, 1]^{m \times J}$ and the loss to be

$$\ell(f, a(x)) = -\sum_{i=1}^m \sum_{x \in \mathcal{X}_0} a_i(x) \left(f(x) - c_i\right). \tag{9}$$

When the loss achieves its maximum value, for each $i \in \{1, \dots, m\}$, $a_i(x)$ determines the c_i-super level set, i.e. $a_i(x) = 1$ for each $x \in \mathcal{X}_0$ with $f(x) > c_i$ and $a_i(x) = 0$ otherwise.

Multi-value sequence search. Given a black-box function $\mathbf{f}$, the goal of multi-value sequence search is to estimate a sequence of inputs $(x_1, \dots, x_m) \in \mathcal{X}^m$, each with a different pre-specified function value $(y_1^{\circledast}, \dots, y_m^{\circledast})$. This task arises when we want to estimate the inverse function $h : \mathcal{Y} \to \mathcal{X}$ (where h returns a point $x \in \mathcal{X}$ for some desired function value y) for a sequence of values $(y_1, \dots, y_m)$. In the context of public health applications, for example, we may be interested in a set of locations where vaccination rates approximate some pre-specified values (e.g. $(20\%, \dots, 80\%)$) when designing the next round of vaccine allocations, as we describe in Section 7.

To solve this problem with our EHIG framework, we let $\mathcal{A} = \mathcal{X}^m$, and define the loss to be

$$\ell(f, \mathbf{a}) = \sum_{i=1}^{m} (f(\mathbf{a}_m) - y_m^{\circledast})^2. \tag{10}$$

6 Acquisition Optimization

At each iteration of H-Entropy Search (Algorithm 1), we optimize the acquisition function to select the next query $x_t = \arg\max_{x \in \mathcal{X}} \text{EHIG}_t(x; \ell, \mathcal{A})$. Historically, zeroth order optimization routinues have often been used for acquisition optimization in BO. However, recent work has developed gradient-based methods for optimizing certain acquisition functions (Wilson et al., 2018; Balandat et al., 2020), which can allow for efficient acquisition optimization over $\mathcal{X}$. We work on similar methodology here—namely, we develop a gradient-based acquisition optimization procedure for appropriate settings (i.e. assuming continuous $\mathcal{X}$ and $\mathcal{A}$, and certain conditions on ℓ). We have implemented this gradient based optimization for each of the acquisition functions described in Section 5, for which we show experimental results in Section 7.

6.1 Gradient-based Acquisition Optimization

Similar to previous related work (Wilson et al., 2018; Balandat et al., 2020), we will provide the following derivation with a focus on Gaussian process (GP) models of the black-box function $\mathbf{f}$, though the methodology can be extended to other models in which we can apply the reparameterization procedure described below to differentiate through posterior model parameters.

Differentiable loss function We first describe a few assumptions that must be satisfied to carry out the gradient-based optimization procedure.

Denote the posterior expected loss given $\mathcal{D}$ by $L(\mathcal{D}, a) := \mathbb{E}_{p(f|\mathcal{D})}[\ell(f, a)]$. We assume that this loss function depends only on the function value of f at a finite number of points, i.e. there exists functions $\mathfrak{x}_1(a), \cdots, \mathfrak{x}_K(a)$, and a function $\ell' : \mathbb{R}^K \times \mathcal{A} \to \mathbb{R}$, for $K \in \mathbb{N}$, such that

$$\ell(f, a) = \ell'(f(\mathfrak{x}_1(a)), f(\mathfrak{x}_2(a)), \cdots, f(\mathfrak{x}_K(a)), a). \tag{11}$$

This requirement is satisfied by the loss functions in Section 5. For brevity, denote the sequence $\mathfrak{x}_1(a), \cdots, \mathfrak{x}_K(a)$ by $\mathfrak{x}_{1:K}(a)$ and $f(\mathfrak{x}_1(a)), \cdots, f(\mathfrak{x}_K(a))$ by $f(\mathfrak{x}_{1:K}(a))$. We assume that the functions $\mathfrak{x}_k$ and ℓ' are differentiable with respect to all arguments. Given a dataset $\mathcal{D}$ and GP prior, the posterior distribution of $f(\mathfrak{x}_K(a))$ is also Gaussian. In particular, there exist functions

$$\mu : \mathfrak{x}_{1:K}(a) \times \mathcal{D} \mapsto \mathbb{R}^K \quad \text{and} \quad U : \mathfrak{x}_{1:K(a)} \times \mathcal{D} \mapsto \mathbb{R}^{K \times K}. \tag{12}$$

such that $f(\mathfrak{x}_{1:K}(a)) = \mu(\mathfrak{x}_{1:K}(a); \mathcal{D}) + U(\mathfrak{x}_{1:K}(a); \mathcal{D})\epsilon$ where ϵ is drawn from a K-dimensional standard normal distribution. We can combine the above results to get

$$L(\mathcal{D}, a) = \mathbb{E}_\epsilon \left[\ell'(\mu(\mathfrak{x}_{1:K}(a); \mathcal{D}) + U(\mathfrak{x}_{1:K}(a); \mathcal{D})\epsilon, a)\right]. \tag{13}$$

A key property is that we can compute unbiased gradients of this with respect to both $\mathcal{D}$ and a, as

$$\nabla L(\mathcal{D}, a) = \mathbb{E}_\epsilon \left[\nabla \ell'(\mu(\mathfrak{x}_{1:K}(a); \mathcal{D}) + U(\mathfrak{x}_{1:K}(a); \mathcal{D})\epsilon, a)\right]. \tag{14}$$

Differentiable acquisition function For a given input $x \in \mathcal{X}$, let $y(x, \mathcal{D})$ denote the posterior predictive distribution of our model. Note that there exists a deterministic function $\bar{y}(x, \mathcal{D}, \lambda)$ such that $y(x, \mathcal{D}) = \bar{y}(x, \mathcal{D}, \lambda)$ where λ is drawn from a standard normal distribution. Hence, if ℓ satisfies Eq. (11), then we can optimize EHIG_t with gradient descent. In particular, we can write

$$\begin{aligned} \inf_{x \in \mathcal{X}} -\text{EHIG}_t(x; \ell, \mathcal{A}) &= \inf_{x \in \mathcal{X}} \mathbb{E}_\lambda [\inf_{a \in \mathcal{A}} \mathbb{E}_{p(f|\mathcal{D} \cup \bar{y}(x, \mathcal{D}, \lambda))}[\ell(f, a)]] \\ &= \inf_{x \in \mathcal{X}} \inf_{\mathfrak{a} : \lambda \mapsto \mathcal{A}} \mathbb{E}_{\lambda, \epsilon}[\ell'(\hat{\mu}(x, \mathfrak{a}(\lambda)) + \hat{U}(x, \mathfrak{a}(\lambda))\epsilon, \mathfrak{a}(\lambda))] \end{aligned} \tag{15}$$

where in Eq. (15), to avoid clutter, we define the notation shorthand $\hat{\mu}(x, \mathfrak{a}(\lambda)) := \mu(\mathfrak{x}_{1:K}(\mathfrak{a}(\lambda)); \mathcal{D} \cup \bar{y}(x, \mathcal{D}, \lambda))$, and $\hat{U}(x, \mathfrak{a}(\lambda)) := U(\mathfrak{x}_{1:K}(\mathfrak{a}(\lambda)); \mathcal{D} \cup \bar{y}(x, \mathcal{D}, \lambda))$. The important property of Eq. (15) is that we can compute the unbiased gradient of the quantity $\mathbb{E}_{\lambda,\epsilon}[\ell'(\hat{\mu}(x, \mathfrak{a}(\lambda)) + \hat{U}(x, \mathfrak{a}(\lambda))\epsilon, \mathfrak{a}(\lambda))]$. In practice, we can also take gradients of a Monte Carlo estimate of Eq. (15) (Balandat et al., 2020), by fixing samples of λ, ϵ throughout the optimization. Specifically, we can sample $\lambda_1, \cdots, \lambda_M$ and $\epsilon_1, \cdots, \epsilon_N$ and approximate Eq. (15) via

$$\inf_{x \in \mathcal{X}} -\text{EHIG}_t(x; \ell, \mathcal{A}) \approx \inf_{x \in \mathcal{X}} \inf_{a_1, \ldots, a_M} \frac{1}{NM} \sum_{m,n} \ell'(\hat{\mu}(x, a_m) + \hat{U}(x, a_m)\epsilon_k, a_m), \tag{16}$$

where we use $a_m = a(\lambda_m)$ for brevity. Under the assumptions above, we can compute the unbiased gradient of this quantity. Using systems such as GPyTorch (Gardner et al., 2018) and BoTorch (Balandat et al., 2020) we can compute these gradients efficiently via automatic differentiation.

7 Experiments

We evaluate our proposed methods on the three optimization applications described in Section 5: top-k estimation with diversity, level set estimation, and multi-value sequence search. In each, we evaluate our method against a set of baselines on real and synthetic black-box functions.

Comparison methods. In each experiment, we compare the following set of acquisition strategies:

- H-Entropy Search (HES). We follow Algorithm 1, using the loss and action set for each task as described in Section 5, and the Monte Carlo gradient-based acquisition optimization procedure outlined in Section 6.1.
- Random Search (RS). At each iteration, we draw a sample x_t uniformly at random from $\mathcal{X}$.
- Uncertainty Sampling (US). At each iteration, we select $x_t = \arg\max_{x \in \mathcal{X}} p(y_x \mid \mathcal{D}_t)$.
- Knowledge Gradient (KG). We also wish to compare against a representative existing BO procedure. KG allows us to carry out a similar Monte Carlo gradient-based acquisition optimization procedure (as it is a special case of HES) on a sensible loss, as detailed in Section 4.

7.1 Top-k Estimation with Diversity

In our first task, the goal is to find a set of k diverse elements in $\mathcal{X}$, each with a high value of $\mathbf{f}$. To assess each method, at each iteration we record $-\ell(\mathbf{f}, a^*)$ using Eq. (7)—i.e. the *negative top-k with diversity loss* of the Bayes action $a^* = \arg\inf_{a \in \mathcal{A}} \mathbb{E}_{p(f|\mathcal{D}_t)}[\ell(f, a)]$ on the true function $\mathbf{f}$—using the set of queries $\mathcal{D}_t$ produced by the given method. Intuitively, if a method makes a set of queries that yield a good estimate of diverse top-k elements, it will score a high value of $-\ell(\mathbf{f}, a^*)$.

In Figure 1 (*top row*) we show qualitative results on the multimodal *Alpine-d* function, defined as *Alpine-d*$(x) = \sum_{i=1}^{d} |x_i \sin(x_i) + 0.1x_i|$, for $x \in \mathbb{R}^d$. Here, HES concentrates queries over five local optima of this function, while KG allocates a majority of samples on only the highest peak, and both US and RS distribute their queries over the full domain $\mathcal{X}$. We compare performance of methods quantitatively in Figure 1 (*bottom row*), where we plot $-\ell(\mathbf{f}, a^*)$ versus iteration on two higher dimensional examples, and can see the advantage of the HES selection strategy.

We also compare performance of each method for this task the *Vaccination* function (provided by Yuan et al. (2021)), which returns the vaccination rate for locations in the continential United States, given an input *(latitude, longitude)*. Here, we restrict the design domain $\mathcal{X}$ to the state of Pennsylvania, due to its rectangular shape. The goal of this task is then to efficiently find a set of five diverse locations over the state that achieve a high vaccination rate. We show results in Figure 1 (*bottom row, right*), and see a similar advantage of HES over comparison methods.

7.2 Level Set Estimation

In our second task, the goal is to carry out binary level set estimation. Here, it is easier to assess each method using a more conventional metric: we produce an estimate of the level $a(x)$ for every $x \in \mathcal{X}_0$, using the model's posterior mean (given the queries selected by a particular comparison method), and then can record the accuracy of this estimate. Intuitvely, a method will achieve a higher accuracy if it chooses queries that yield a fine-grained estimate of the function near the boundaries of the level set.

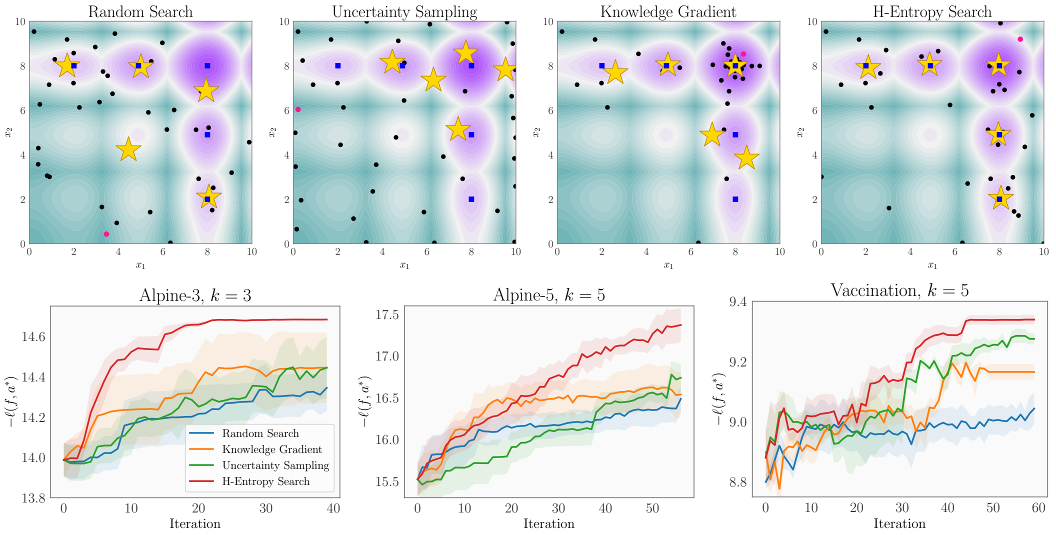

Figure 1: **Top-k estimation with diversity.** *Top row:* A comparison of methods on the *Alpine-2* function, showing the set of ground-truth top-k diverse design points (blue squares), queries $\mathcal{D}_t$ taken (black dots), acquisition function optimizer (pink dot), and the estimated set of top-k diverse design points (gold stars). *Bottom row:* Plots of $-\ell(\mathbf{f}, a^*)$ versus iteration for the set of comparison methods, on the *Alpine-3*, *Alpine-5* and *Vaccination* functions, averaged over 3 trials, where error bars represent one standard error.

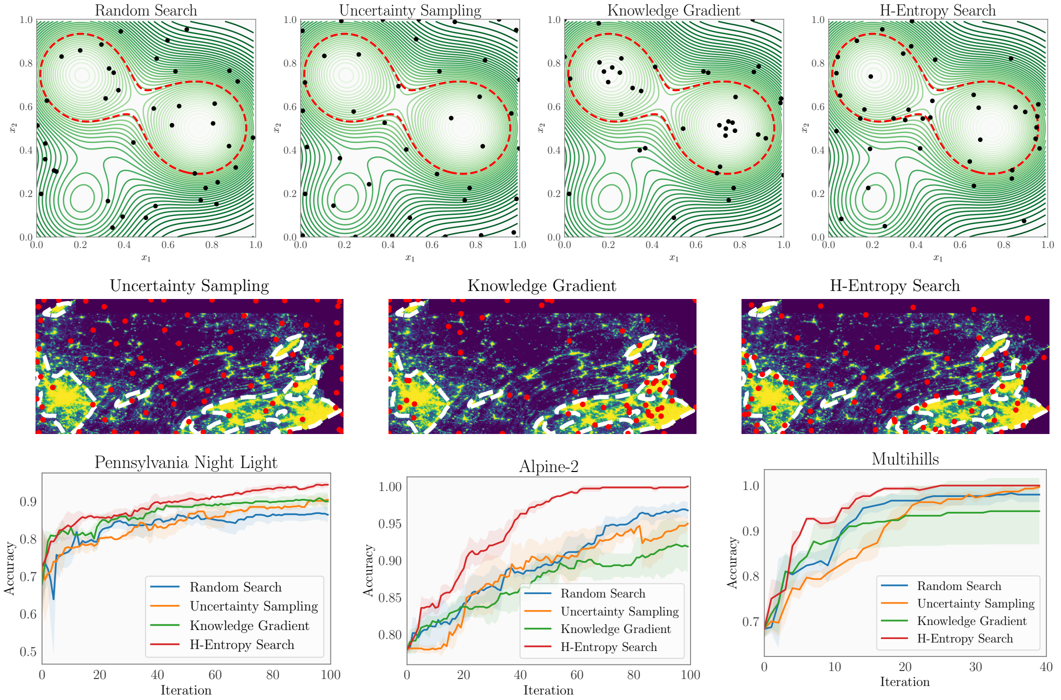

Figure 2: **Level set estimation.** *Top and middle rows:* A comparison of methods on the *Multihills* (*top*) and *Pennsylvania Night Light* (*middle*) functions, showing the ground-truth level set boundary (dashed line) and queries $\mathcal{D}_t$ taken (black or red dots). *Bottom row*: Plots of accuracy versus iteration for the set of comparison methods, on three functions, each averaged over 3 trials, where error bars represent one standard error.

In the *top row* of Figure 2, we show qualitative results on the *Multihills* function, defined as a mixture density (details given in appendix), and in the *middle row* of Figure 2, we show qualitative results on the *Pennsylvania Night Light* function[1], released by NASA (additional details in the appendix), which

[1] https://earthobservatory.nasa.gov/features/NightLights

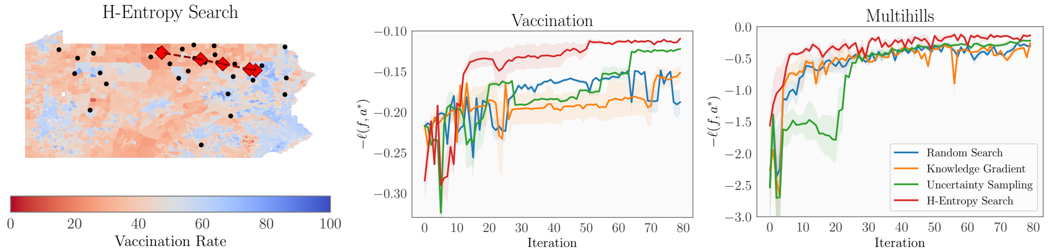

Figure 3: **Multi-value sequence search.** *Left:* Visualization of the *Vaccination* function, along with the queries $\mathcal{D}_t$ taken by HES (black dots), and the estimated sequence $(x_1^{\circledast}, \ldots, x_5^{\circledast})$ (red diamonds), such that $(f(x_1^{\circledast}), \ldots, f(x_5^{\circledast})) = (30\%, 40\%, 50\%, 60\%, 70\%)$. *Middle and right:* Plots of $-\ell(\mathbf{f}, a^*)$ versus iteration for the set of comparison methods, on the *Vaccination* and *Multihills* functions, averaged over 3 trials, where error bars represent one standard error.

returns the relative level of light at a location in Pennsylvania, as queried by a satellite image. The goal of this experiment is to determine the portion of land at which night light is above a specified threshold value. In both cases, HES concentrates queries along the boundary of the level set (denoted as a dashed line). In the *bottom row*, we plot the accuracy vs. iteration of each method, and see that this allocation strategy leads to a higher accuracy relative to comparison methods, which allocate queries to optima (KG) or distribute them over the full domain (US and RS).

7.3 Multi-value Sequence Search

In our third task, the goal is find a sequence of elements whose value under the black-box function matches some sequence of pre-specified function values $(y_1^{\circledast}, \ldots, y_m^{\circledast})$.

To assess each method, at each iteration we again record $-\ell(\mathbf{f}, a^*)$ from Eq. (10)—i.e. the *negative multi-value sequence loss* of the Bayes action $a^* = \arg\inf_{a \in \mathcal{A}} \mathbb{E}_{p(f|\mathcal{D}_t)} [\ell(f, a)]$ on the true function $\mathbf{f}$—using the set of queries $\mathcal{D}_t$ produced by the given method. Intuitively, if a method makes a set of queries that yield a good estimate of a sequence of points $(x_1^{\circledast}, \ldots, x_m^{\circledast})$ such that $(\mathbf{f}(x_1^{\circledast}), \ldots, \mathbf{f}(x_m^{\circledast})) \approx (y_1^{\circledast}, \ldots, y_m^{\circledast})$ it will score highly on $-\ell(\mathbf{f}, a^*)$.

In Figure 3 (*left*) we show qualitative results on the *Vaccination* function (described in Section 7.1). Here, our goal is to find a sequence of five *(latitude, longitude)* coordinates with vaccination rates equal to $(y_1^{\circledast}, \ldots, y_m^{\circledast}) = (30\%, 40\%, 50\%, 60\%, 70\%)$. Information about sequences of values under an expensive black-box function such as this can be useful when making policy decisions involving a vaccine response or allocation. In this case, we see that HES concentrates queries along a route from the relatively highly vaccinated region in the East to the relatively lowly vaccinated region in the North. The *middle* and *right* plots in Figure 3 provide a quantitive comparison of methods on the *Vaccination* and *Multihills* functions, plotting $-\ell(\mathbf{f}, a^*)$ vs. iteration. These again show the benefits of query selection performed by HES relative to the comparison strategies.

8 Conclusion

In this paper, we take a decision making perspective on acquisition functions in Bayesian optimization: after the BO procedure is complete, we assume that we must make some decision a^* and then incur a loss $\ell(a^*, \mathbf{f})$. Our goal is then to make a sequence of queries that reduce the uncertainty of the posterior distribution $p(f \mid \mathcal{D}_t)$ in a way to help best make this decision with low loss. Using H-entropy (DeGroot, 1962; Rao, 1984), we can define an EHIG acquisition function which carries this out directly: it selects a point that is expected to maximally reduce the posterior expected loss of the Bayes action a^*. We incorporate this acquisition function into a BO procedure called H-Entropy Search, and show, under certain conditions, that we can perform efficient gradient-based optimization of this acquisition function.

There are multiple interesting future directions of study. First, we hope to develop efficient acquisition optimization procedures for a broader array of settings, such as for non-continuous action sets $\mathcal{A}$ or design spaces $\mathcal{X}$. One interesting avenue is hybrid optimization settings, where we can take gradient steps with respect to either design or action variables, but must resort to zeroth order optimization methods for the other. We also wish to further study how the EHIG framework and optimization strategies proposed in this paper could be used to improve existing BO procedures and provide insights on existing acquisition functions.

REFERENCES

Ali Abbasi, Ahmad Khonsari, and Navid Farri. Mote: efficient monitoring of top-k set in sensor networks. In *2008 IEEE Symposium on Computers and Communications*, pp. 957–962. IEEE, 2008.

Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. Botorch: A framework for efficient monte-carlo bayesian optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Benedikt Boecking, Willie Neiswanger, Eric Xing, and Artur Dubrawski. Interactive weak supervision: Learning useful heuristics for data labeling. *arXiv preprint arXiv:2012.06046*, 2020.

Eric Brochu, Vlad M Cora, and Nando De Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

M H DeGroot. Uncertainty, information, and sequential experiments. *Ann. Math. Stat.*, 33(2): 404–419, 1962.

Peter Frazier, Warren Powell, and Savas Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS journal on Computing*, 21(4):599–613, 2009.

Jacob R Gardner, Geoff Pleiss, David Bindel, Kilian Q Weinberger, and Andrew Gordon Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *arXiv preprint arXiv:1809.11165*, 2018.

Peter D Grünwald and A Philip Dawid. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *Ann. Stat.*, 32(4):1367–1433, August 2004.

Philipp Hennig and Christian J Schuler. Entropy search for Information-Efficient global optimization. *J. Mach. Learn. Res.*, 13(57):1809–1837, 2012.

José Miguel Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *NIPS*, 2014.

Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

Yue Liu, Tianlu Zhao, Wangwei Ju, and Siqi Shi. Materials discovery and design using machine learning. *Journal of Materiomics*, 3(3):159–177, 2017.

Jonas Močkus. On bayesian methods for seeking the extremum. In *Optimization techniques IFIP technical conference*, pp. 400–404. Springer, 1975.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32: 8026–8037, 2019.

Victor Picheny, Tobias Wagner, and David Ginsbourger. A benchmark of kriging-based infill criteria for noisy optimization. *Structural and Multidisciplinary Optimization*, 48(3):607–626, 2013.

C Radakrishna Rao. Differential metrics in probability spaces. *Differential geometry in statistical inference*, 10:217–240, 1987.

C Radhakrishna Rao. Diversity and dissimilarity coefficients: A unified approach, 1982.

C Radhakrishna Rao. Convexity properties of entropy functions and analysis of diversity. *Lecture Notes-Monograph Series*, pp. 68–77, 1984.

Aarti Singh. *Nonparametric Set Estimation Problems in Statistical Inference and Learning*. PhD thesis, University of Wisconsin–Madison, 2008.

Zi Wang and Stefanie Jegelka. Max-value entropy search for efficient bayesian optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3627–3635, 2017.

James T Wilson, Frank Hutter, and Marc Peter Deisenroth. Maximizing acquisition functions for bayesian optimization. *arXiv preprint arXiv:1805.10196*, 2018.

Pengtao Xie. *Diversity-promoting and large-scale machine learning for healthcare*. PhD thesis, University of Pittsburgh Medical Center, 2018.

Yuan Yuan, Eaman Jahani, Shengjia Zhao, Yong-Yeo Ahn, and Alex Sandy Pentland. Mobility network reveals the impact of geographic vaccination heterogeneity on covid-19. *medRxiv*, 2021.

Miao Zhong, Kevin Tran, Yimeng Min, Chuanhao Wang, Ziyun Wang, Cao-Thang Dinh, Phil De Luna, Zongqian Yu, Armin Sedighian Rasouli, Peter Brodersen, et al. Accelerated discovery of co 2 electrocatalysts using active machine learning. *Nature*, 581(7807):178–183, 2020.

A PROOFS OF PROPOSITIONS

Here we prove the propositions stated in Section 4.

Proposition 1. If we choose $\mathcal{A} = \mathcal{X}$ and $\ell(f, x) = -f(x)$, then the EHIG is equivalent to the knowledge gradient acquisition function, i.e. $\text{EHIG}_t(x; \ell, \mathcal{A}) = \text{KG}_t(x)$.

Proof of Proposition 1. The proof follows directly from the definition of H-entropy and the EHIG, namely

$$\begin{aligned}
\text{EHIG}_t(x) &= \inf_{a\in\mathcal{A}} \mathbb{E}_{p(f|\mathcal{D}_t)}\left[\ell(f,a)\right] - \mathbb{E}_{p(y_x|\mathcal{D}_t)}\left[\inf_{a\in\mathcal{A}} \mathbb{E}_{p(f|\mathcal{D}_t\cup\{(x,y_x)\})}\left[\ell(f,a)\right]\right] && (17)\\
&= \inf_{x'\in\mathcal{X}} \mathbb{E}_{p(f|\mathcal{D}_t)}\left[-f(x')\right] - \mathbb{E}_{p(y_x|\mathcal{D}_t)}\left[\inf_{x'\in\mathcal{X}} \mathbb{E}_{p(f|\mathcal{D}_t\cup\{(x,y_x)\})}\left[-f(x')\right]\right] && (18)\\
&= -\sup_{x'\in\mathcal{X}} \mathbb{E}_{p(f|\mathcal{D}_t)}\left[f(x')\right] + \mathbb{E}_{p(y_x|\mathcal{D}_t)}\left[\sup_{x'\in\mathcal{X}} \mathbb{E}_{p(f|\mathcal{D}_t\cup\{(x,y_x)\})}\left[f(x')\right]\right] && (19)\\
&= \mathbb{E}_{p(y_x|\mathcal{D}_t)}\left[\mu^*_{t+1}(x,y_x)\right] - \mu^*_t && (20)\\
&= \text{KG}_t(x) && (21)
\end{aligned}$$

□

Proposition 2. If we choose $\mathcal{A} = \mathcal{P}(\Theta)$ and $\ell(f, q) = -\log q(\theta_f)$, then the EHIG is equivalent to the entropy search acquisition function, i.e. $\text{EHIG}_t(x; \ell, \mathcal{A}) = \text{ES}_t(x)$.

Proof of Proposition 2. We first prove that under our definition of loss ℓ, the H-entropy $H[f \mid \mathcal{D}_t]$ is equivalent to the Shannon entropy of the posterior distribution over θ_f (where, as an example, θ_f could be equal to the global maximizer x^* of f).

Note that the H-entropy is the expected loss of the Bayes action

$$q^* = \arg\inf_{q\in\mathcal{P}(\mathcal{X})}\mathbb{E}_{p(f|\mathcal{D}_t)}\left[-\log q(\theta_f)\right].$$

We want to show that q^* defined above is equal to $p(\theta_f \mid \mathcal{D}_t)$. To do so, note that

$$\begin{aligned}
q^* &= \arg\inf_{q\in\mathcal{P}(\mathcal{X})}\mathbb{E}_{p(f|\mathcal{D}_t)}\left[-\log q(\theta_f|\mathcal{D}_t)\right] && (22)\\
&= \arg\inf_{q\in\mathcal{P}(\mathcal{X})}\mathbb{E}_{p(\theta_f|\mathcal{D}_t)}\left[-\log q(\theta_f|\mathcal{D}_t)\right] && (23)\\
&= p(\theta_f|\mathcal{D}_t), && (24)
\end{aligned}$$

where the first equality holds since

$$E_X[f(g(X))] = E_Z[f(Z)], \text{when } Z = g(X). \quad (25)$$

Therefore, under this loss and action set, using the definition of the EHIG we can write

$$\text{EHIG}_t(x;\ell,\mathcal{A}) = \mathcal{H}(p(\theta_f \mid \mathcal{D}_t)) - \mathbb{E}_{p(y_x|\mathcal{D}_t)}[\mathcal{H}(p(\theta_f \mid \mathcal{D}_t \cup \{x, y_x\}))] = \text{ES}_t(x). \quad (26)$$

□

Proposition 3. If we choose $\mathcal{A}_t = \{x_i\}_{i=1}^{t-1}$, where $x_i \in \mathcal{D}_t$, and $\ell(f, x_i) = -f(x_i)$, then the EHIG is equal to the expected improvement acquisition function, i.e. $\text{EHIG}_t(x; \ell, \mathcal{A}) = \text{EI}_t(x)$.

Proof of Proposition 3. The first term in Eq. (3) is equal to:

$$\mathcal{H}_{\ell,\mathcal{A}_t}[f \mid \mathcal{D}_t] = \inf_{a\in\mathcal{A}_t} \mathbb{E}_{p(f|\mathcal{D}_t)}\left[\ell(f,a)\right] = -\max_{i\leq t-1} \hat{f}(x_i) := -f^*_t \quad (27)$$

where $\hat{f}(x_i)$ is the posterior expected value of f at x_i.

The second term in Eq. (3) is:

$$
\begin{aligned}
&\mathbb{E}_{p(y_x|\mathcal{D}_t)}\left[H_{\ell,\mathcal{A}_{t+1}}\left[f \mid \mathcal{D}_t \cup \{(x,y_x)\}\right]\right] && (28)\\
=&\mathbb{E}_{p(y_x|\mathcal{D}_t)}\left[\mathbb{E}_{p(f|\mathcal{D}_t\cup\{(x,y_x)\})}\left[\inf_{a\in A_{t+1}} \ell(f,a)\right]\right] && (29)\\
=&\mathbb{E}_{p(y_x|\mathcal{D}_t)}\left[\mathbb{E}_{p(f|\mathcal{D}_t\cup\{(x,y_x)\})}\left[-\max(f_t^*, f(x))\right]\right] && (30)\\
=&\mathbb{E}_{p(y_x|\mathcal{D}_t)}\left[-\max(f_t^*, y_x)\right] && (31)
\end{aligned}
$$

Putting it together, the EHIG_t acquisition function in Eq. (3) will reduce to:

$$
\begin{aligned}
\mathrm{EHIG}_t(x;\ell,\mathcal{A}) &= -f_t^* - \mathbb{E}_{p(y_x|\mathcal{D}_t)}\left[-\max(f_t^*, y_x)\right] && (32)\\
&= \mathbb{E}_{p(y_x|\mathcal{D}_t)}[\max(0, y_x - f_t^*)] && (33)\\
&= \mathrm{EI}_t(x). && (34)
\end{aligned}
$$

□

B Additional Details on Experiments

Here we show results from additional experiments and datasets.

Details on the *Multihills* function The *Multihills* function is defined as a mixture density as follows. $\textit{Multihills}(x) = \sum_{j=1}^{J} w_j \mathcal{N}(x \mid \mu_j, C_j)$, for $x \in \mathbb{R}^d$, where $\mathcal{N}$ denotes a multivariate normal density, $\{\mu_j\}$ are a set of J means, $\{C_j\}$ are a set of J covarance matrices, and $\{w_j\}$ are a set of J weights.

Details on the *Pennsylvania Night Light* function We consider the 2012 gray scale global night-light raster with resolution 0.1 degree per pixel. The data is downloaded from NASA[2]. We focus on Pennsylvania and normalize the raster data before using.

More visualization results for level set estimation on Alpine-2 function. We provide an additional visualization result for the level set estimation experiment on the Alpine-2 function in Figure 4.

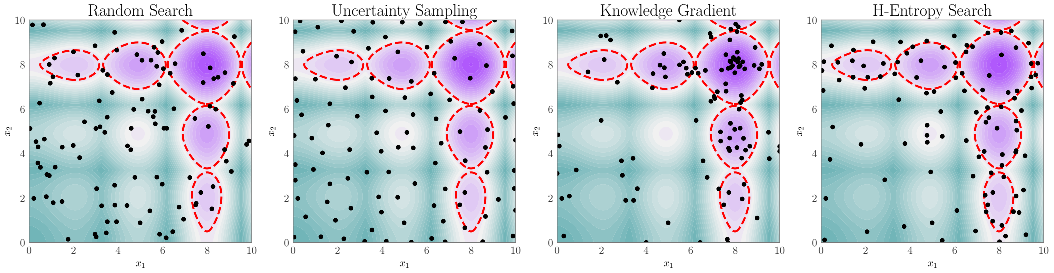

Figure 4: Level set estimation for Alpine-2 function. We show the ground-truth level set boundary with red dashed line and queries $\mathcal{D}_t$ taken with black dots.

[2] https://earthobservatory.nasa.gov/features/NightLights