Joint Entropy Search For Maximally-Informed Bayesian Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Information-theoretic Bayesian optimization techniques have become popular for optimizing expensive-to-evaluate black-box functions due to their non-myopic qualities. Entropy Search and Predictive Entropy Search both consider the entropy over the optimum in the input space, while the recent Max-value Entropy Search considers the entropy over the optimal value in the output space. We propose Joint Entropy Search (JES), a novel information-theoretic acquisition function that considers an entirely new quantity, namely the entropy over the joint optimal probability density over both input and output space. To incorporate this information, we consider the reduction in entropy from conditioning on fantasized optimal input/output pairs. The resulting approach primarily relies on standard GP machinery and removes complex approximations typically associated with information-theoretic methods. With minimal computational overhead, JES shows superior decision-making, and yields state-of-the-art performance for information-theoretic approaches across a wide suite of tasks. As a light-weight approach with superior results, JES provides a new go-to acquisition function for Bayesian optimization.

1 Introduction

The optimization of expensive black-box functions is a prominent task, arising across a wide range of applications. *Bayesian optimization* (BO) [24, 33] is a sample-efficient approach, and has been successfully applied to various problems, including machine learning hyperparameter optimization [2, 19, 31, 35], robotics [3, 6, 23], hardware design [10, 25], and tuning reinforcement learning agents like AlphaGo [7]. In BO, a probabilistic surrogate model is used for modeling the (unknown) objective. The selection policy employed by the BO algorithm is dictated by an acquisition function, which draws on the uncertainty of the surrogate to guide the selection of the next query.

The choice of acquisition function is significant for the success of the BO algorithm. A popular line of acquisition functions takes an information-theoretic angle, and considers the *expected information gain* regarding the location of the optimum that is obtained from an upcoming query. *Entropy Search* (ES) [14], *Predictive Entropy Search* (PES) [15] and the earlier work of IAGO [42] select queries by maximizing this quantity. While ES andPESare efficient in the number of queries to optimize the objective, they both require significant computational effort and complex approximations of the expected information gain, which impacts their performance and practical use [15, 43]. To address this computational expense, efficient parameterizations of the objective have been proposed [29]. However, this solution imposes constraints in the ability to accurately model the objective.

A related information-theoretic family of approaches considers the information gain on the optimal objective value [17, 43]. *Max-value Entropy Search* (MES) [43] was the first information-theoretic approach to have a proven convergence rate, albeit only in a noiseless problem setting. Moreover, its consideration of a one-dimensional density over the output space as opposed to a D-dimensional

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

input space and a reduction in intricate approximations yielded a computationally efficient alternative to the ES/PES line of approaches. Despite its empirical success, some crucial shortcomings of MES have been highlighted in recent works [26, 39]. Most importantly, it does not differentiate between the (unobserved) maximal objective value f^* and the observed noisy maximum y_{max}. As such, its assumption on the posterior distribution of the output $p(y|\mathcal{D}, \boldsymbol{x})$ does not hold in any setting where noise is present [26].

We propose an approach which merges the ES/PES and MES lines of work, and provides an all-encompassing perspective on information gain regarding optimality. We introduce Joint Entropy Search (JES), a novel acquisition function which has the following advantages over existing infomation-theoretic approaches:

1. It utilizes two sources of information, by considering the entropy over both the optimum and optimal value;
2. It retains low computational and conceptual complexity by utilizing the full optimal observation, allowing it to rely on standard GP machinery instead of complex approximations; and
3. It intrinsically supports noisy objective functions, making it particularly efficient for these types of tasks.

The code for reproducing the experiments is available at
`https://github.com/jointentropysearch/JointEntropySearch`.

2 Background and related work

Bayesian optimization. We consider the problem of optimizing a black-box function f across a set of feasible inputs $\mathcal{X} \subset \mathbb{R}^d$:

$$\boldsymbol{x}^* \in \arg\max_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}). \tag{1}$$

We assume that $f(\boldsymbol{x})$ is expensive to evaluate, and can potentially only be observed through a noise-corrupted estimate, y, where $y = f(\boldsymbol{x}) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ for some noise level σ_ϵ^2. In this setting, we wish to maximize f in an efficient manner, typically while adhering to a budget which sets a cap on the number of points that can be evaluated. BO aims to globally maximize f by an initial design and thereafter sequentially choosing new points $\boldsymbol{x}_n$ for some iteration n, creating the data $\mathcal{D}_n = \mathcal{D}_{n-1} \cup \{(\boldsymbol{x}_n, y_n)\}$. After each new observation, BO constructs a probabilistic surrogate model $p(f|\mathcal{D}_n)$ and uses that surrogate to build an acquisition function $\alpha(\boldsymbol{x}, \mathcal{D}_n)$. The combination of surrogate model and acquisition function encodes the strategy for selecting the next point $\boldsymbol{x}_{n+1}$. After the full budget of N iterations is exhausted, a best configuration $\boldsymbol{x}_N^*$ is returned as either the $\arg\max$ of the observed values, or the optimum as predicted by the surrogate model.

Gaussian processes. When constructing the surrogate, the most common choice is the *Gaussian process* (GP) [28]. Formally, a GP is an infinite collection of random variables, such that every finite subset of those variables follows a multivariate Gaussian distribution. The GP utilizes a covariance (or kernel) function k, which encodes a prior belief for the smoothness of f, and determines how previous observations influence prediction. Typical choices for the kernel are the squared exponential and Matérn-5/2 [22] kernel. Given the observations $\mathcal{D}_n$ at iteration n, the density $p(f|\mathcal{D}_n)$ over the objective is characterized by the posterior mean m_n and variance s_n of the GP:

$$m_n(\boldsymbol{x}) = \mathbf{k}_n(\boldsymbol{x})^\top (\mathbf{K}_n + \sigma_\epsilon^2 \mathbf{I})^{-1} \mathbf{y}, \quad s_n(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{x}) - \mathbf{k}_n(\boldsymbol{x})^\top (\mathbf{K}_n + \sigma_\epsilon^2 \mathbf{I})^{-1} \mathbf{k}_n(\boldsymbol{x}), \tag{2}$$

where $(\mathbf{K}_n)_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\mathbf{k}_n(\boldsymbol{x}) = [k(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, k(\boldsymbol{x}, \boldsymbol{x}_n)]^\top$ and σ_ϵ^2 is the noise variance. Alternative surrogate models include random forests [18] and Bayesian neural networks [36, 37].

Acquisition functions. The *acquisition function* acts on the surrogate model to quantify the attractiveness of a point in the search space. Acquisition functions employ a trade-off between exploration and exploitation, typically using a greedy heuristic to do so. Simple, computationally cheap heuristics are Expected Improvement(EI) [5, 20]. For a noiseless function, EI selects the next point $\boldsymbol{x}_{n+1}$ as

$$\boldsymbol{x}_{n+1} \in \arg\max_{\boldsymbol{x} \in \mathcal{X}} \mathbb{E}\left[(y_n^* - y_{n+1}^*)^+\right] = \arg\max_{\boldsymbol{x} \in \mathcal{X}} Z s_n(\boldsymbol{x}) \Phi(Z) + s_n(\boldsymbol{x}) \phi(Z), \tag{3}$$

where $Z = (y_n^* - m_n(\boldsymbol{x}))/s_n(\boldsymbol{x})$. Other acquisition functions which use similar heuristics are the Upper Confidence Bound (UCB) [38], and Probability of Improvement(PI) [21]. A more sophisticated approach related to EI is the Knowledge Gradient [11].

Information-theoretic acquisition functions. Information-theoretic acquisition functions [14, 15, 30, 43] and their various adaptations [1, 16, 32] seek to maximize the expected information gain I from observing a subsequent query $(\boldsymbol{x}, y)$ regarding the optimum, $\boldsymbol{x}^*$. This equates to reducing the uncertainty of the density over the optimum, $p(\boldsymbol{x}^*|\mathcal{D}) = \mathbb{P}(\boldsymbol{x} = \arg\max_{\boldsymbol{x}' \in \mathcal{X}} f(\boldsymbol{x}')|\mathcal{D})$, using the information obtained through $(\boldsymbol{x}, y)$. By quantifying uncertainty through differential entropy H, design points are selected based on the expected reduction in this quantity over $p(\boldsymbol{x}^*|\mathcal{D})$. Formally, this is expressed as the difference between the current entropy over $p(\boldsymbol{x}^*|\mathcal{D})$, and the expected entropy of that density after observing the next query:

$$\alpha_{\texttt{ES}}(\boldsymbol{x}) = I((\boldsymbol{x}, y); \boldsymbol{x}^*|\mathcal{D}) = \mathrm{H}[p(\boldsymbol{x}^*|\mathcal{D})] - \mathbb{E}_y \left[\mathrm{H}[p(\boldsymbol{x}^*|\mathcal{D} \cup (\boldsymbol{x}, y)]\right]. \tag{4}$$

By utilizing the symmetric property of the mutual information, one can arrive at an equivalent expression, where the entropy is computed with regard to the density over the output y,

$$\alpha_{\texttt{PES}}(\boldsymbol{x}) = I(y; (\boldsymbol{x}, \boldsymbol{x}^*)|\mathcal{D}) = \mathrm{H}[p(y|\mathcal{D}, \boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}^*} \left[\mathrm{H}[p(y|\mathcal{D}, \boldsymbol{x}, \boldsymbol{x}^*)]\right]. \tag{5}$$

Eq. 4 is the original formulation used in ES [14] and Eq. 5 is the formulation introduced with PES [15]. Both formulations require a series of approximations and expensive computational steps to compute the entropy in the second term. For PES specifically, with n data points of dimension d, the second term is estimated through Monte Carlo (MC) methods by computing Cholesky decompositions of size $\mathcal{O}(n + d^2/2)^3$, and approximating the Hessian at the optimum for each MC sample.

MES [43] avoids this computational hurdle by considering the information gain $I((\boldsymbol{x}, y); y^*)$ regarding the optimal value y^*. As such, it computes the entropy reduction for a one-dimensional density:

$$\alpha_{\texttt{MES}}(\boldsymbol{x}) = I(y; (\boldsymbol{x}, y^*)|\mathcal{D}) = \mathrm{H}[p(y|\mathcal{D}, \boldsymbol{x})] - \mathbb{E}_{y^*} \left[\mathrm{H}[p(y|\mathcal{D}, \boldsymbol{x}, y^*)]\right]. \tag{6}$$

Here, it is assumed that the posterior predictive distribution $p(y|\mathcal{D}, \boldsymbol{x}, y^*)$ is a truncated Gaussian distribution, for which the entropy can be computed in closed form. However, $p(y|\mathcal{D}, \boldsymbol{x}, y^*)$ takes this form only in a strictly noiseless setting [26, 39], where it holds true that $f^* = y^*$, i.e., when the maximal observation and the optimal value of the objective function are the same. For noisy applications, this assumption leads to an overestimation of the entropy reduction [26].

3 Joint Entropy Search

We now present Joint Entropy Search (JES), a novel information-theoretic approach for Bayesian optimization. As for other information-theoretic acquisition functions, JES considers a mutual information quantity. However, unlike its predecessors, JES adds an additional piece of information: compared to ES/PES, it adds the density over the noiseless optimal value f^*, and compared to MES, it adds the density over $\boldsymbol{x}^*$. It utilizes a novel two-step reduction in the predictive entropy from conditioning on sampled optima and their associated values. Throughout the section, we will refer to a sampled optimum and its associated value, $(\boldsymbol{x}^*, f^*)$, as an *optimal pair*.

3.1 Joint density over the optimum and optimal value: pictorial

JES considers the joint probability density $p(\boldsymbol{x}^*, f^*)$ over both the optimum $\boldsymbol{x}^*$ and the true, noiseless optimal value f^*. Fig. 1 visualizes the densities $p(\boldsymbol{x}^*)$ and $p(f^*)$, considered by ES/PES and MES, respectively, and the joint density $p(\boldsymbol{x}^*, f^*)$, considered by JES. As highlighted by the vertical dashed lines for the point selection of each strategy (bottom), PES chooses strictly to reduce the uncertainty over $\boldsymbol{x}^*$, and as such, considers a region where the uncertainty over the optimal value is low. However, it can effectively determine that the right side of the local optimum is more promising to query next. MES seeks to reduce the tail of the probability density over f^* (right), which in this case leads to an exploratory query. JES' joint probability density over optimum and optimal value captures uncertainties over both "where" and "how large" the optimum will be. As such, it selects a point which is uncertain under both measures. For the selected query in Fig. 1, JES will learn substantially about both $\boldsymbol{x}^*$ and f^* by querying it, whereas PES and MES learn only about one of them.

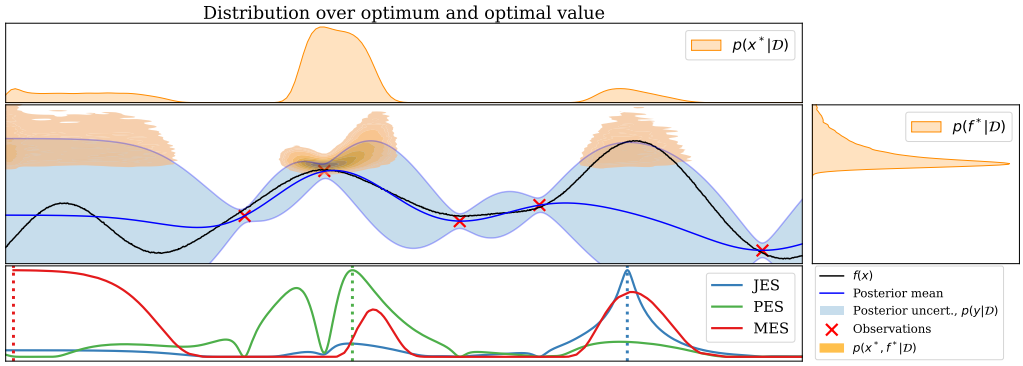

Figure 1: The densities considered by ES/PES (top), MES (right) and JES (center) on a one-dimensional toy example. The multimodal density $p(\boldsymbol{x}^*, f^*)$ is reduced to a heavy-tailed density over f^* for the density used by MES (right), which does not capture the multi-modality of the density over the optimum. The density over $\boldsymbol{x}^*$ used by PES (top) does not capture the apparent exploration/exploitation trade-off that exists between the modes. The acquisition functions and their next point selections are highlighted with dashed lines (bottom).

3.2 The Joint Entropy Search acquisition function

We consider the mutual information between the random variables $(\boldsymbol{x}^*, f^*)$ and a future query $(\boldsymbol{x}, y)$:

$$\alpha_{\texttt{JES}}(\boldsymbol{x}) = I((\boldsymbol{x}, y); (\boldsymbol{x}^*, f^*)|\mathcal{D}_n) \tag{7}$$
$$= \mathrm{H}[p(y|\mathcal{D}, \boldsymbol{x})] - \mathbb{E}_{(\boldsymbol{x}^*, f^*)}\left[\mathrm{H}[p(y|\mathcal{D}, \boldsymbol{x}, \boldsymbol{x}^*, f^*)]\right] \tag{8}$$
$$= \mathrm{H}[p(y|\mathcal{D}, \boldsymbol{x})] - \mathbb{E}_{(\boldsymbol{x}^*, f^*)}\left[\mathrm{H}[p(y|\mathcal{D} \cup (\boldsymbol{x}^*, f^*), \boldsymbol{x}, f^*)]\right]. \tag{9}$$

Eq. 8 is similar to Eq. 6 but with the addition of $\boldsymbol{x}^*$ and the replacement of y^* with f^* in the conditioning of the second term. The expectation is computed with respect to a $D+1$-dimensional joint probability density over $\boldsymbol{x}^*$ and f^*. In Eq. 9, we make it explicit that the conditional density inside the expectation is obtained after (1) conditioning the GP on the previous data $\mathcal{D}$, plus one additional noiseless optimal pair $(\boldsymbol{x}^*, f^*)$, and (2) knowing that the noiseless optimal value is in fact f^*. By utilizing the complete observation $(\boldsymbol{x}^*, f^*)$, and of just $\boldsymbol{x}^*$ *or* f^*, we can treat it like any other (noiseless) observation. As such, we quantify much of the entropy reduction by utilizing standard GP functionality. For (2), we gain knowledge of the maximum of the noiseless objective, globally. Following [26, 43], the resulting effect is to truncate the GP's posterior over f, upper bounding it to f^*. The expectation in Eq. 9 is approximated through MC by sampling L optimal pairs $\{(\boldsymbol{x}^*_\ell, f^*_\ell)\}_{\ell=1}^L$ from $p(\boldsymbol{x}^*, f^*)$ using an approximate version of *Thompson Sampling* (TS) [40]. as explained in Sec. 3.3. In Fig. 2, the resulting posterior distribution of the two-step conditioning is shown in greater detail. As pointed out in [26, 39], after conditioning on f^*, the posterior predictive density over y is a sum of a truncated Gaussian distribution over f and the Gaussian noise ϵ. As such, we can loosely relate the entropy reduction to two separate variance reduction steps over $p(y|\mathcal{D}, \boldsymbol{x})$: a conditioning term and a truncation term. The former primarily reduces the entropy over $\boldsymbol{x}^*$, while the latter mainly reduces the entropy over f^*. Moreover, the former is easily computed exactly, and the latter, which is generally a smaller term, requires approximation, as shown in Sec. 3.4.

Fig. 3 shows the difference in log variance over $p(y|\mathcal{D}, \boldsymbol{x})$ resulting from conditioning (in blue) and truncation (in orange) for the scenario in Fig. 2. The overall reduction is largest close to the point of conditioning, and the truncation term mainly contributes at uncertain regions far away from the conditioned point. Moreover, the magnitude of the conditioning term will rely on the prior variance at the conditioned point, as a larger prior variance will lead to a larger reduction in entropy from conditioning. As we average over optimal pairs, many such entropy reduction terms accumulate.

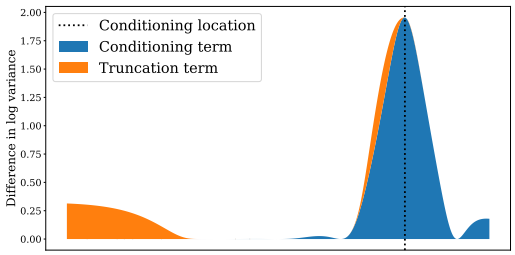

Figure 3: Reduction in log variance from the conditioning step and the truncation step as visualized in Fig. 2. The local conditioning term (blue), and the smaller, globally variance-reducing truncation term (orange).

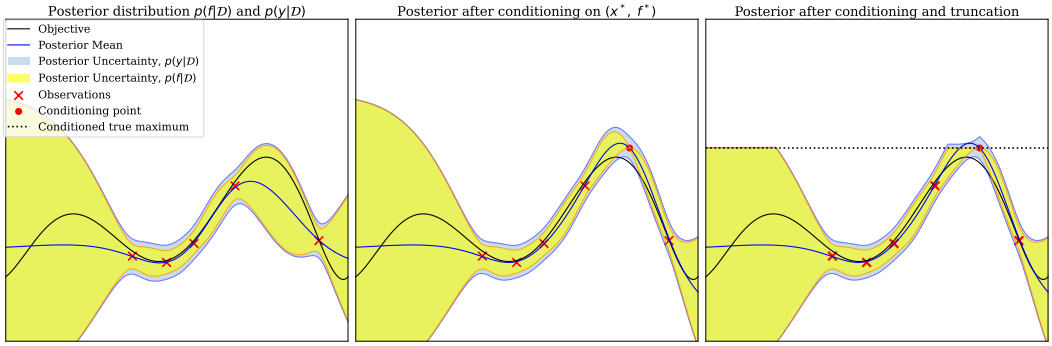

Figure 2: Step-by-step modeling when conditioning on one optimal pair $(\boldsymbol{x}^*, f^*)$. The posterior with noise $p(y|\mathcal{D})$ and without noise $p(f|\mathcal{D})$ are illustrated in blue and yellow, respectively. The GP after 5 (noisy) observations, before conditioning on $(\boldsymbol{x}^*, f^*)$ is shown on the left. In the middle panel, we draw $(\boldsymbol{x}^*, f^*)$ and condition on it, making $p(f|\mathcal{D} \cup (\boldsymbol{x}^*, f^*))$ a delta distribution at the conditioning point as the fantasized observation f^* is noiseless. Since f^* is also the presumed noiseless maximum, we truncate its posterior $p(f|\mathcal{D} \cup (\boldsymbol{x}^*, f^*), f^*)$ globally in the right panel. The observation noise allows for non-zero density on $p(y > f^*|\mathcal{D} \cup (\boldsymbol{x}^*, f^*), f^*)$. We note that, while the noise is homoscedastic, its relative contribution to the total variance differs over the input space. As such, and since we're plotting standard deviations (not variances), the blue region is wider near observed data, where $p(f|\mathcal{D})$ has lower variance.

3.3 Incorporating optimal pairs

To obtain samples $(\boldsymbol{x}^*, f^*)$, we utilize an approximate variant of TS [40], originally proposed in PES [15]. We utilize Bochner's theorem [4], which, for any stationary kernel k, asserts the existence of its Fourier dual $s(\boldsymbol{w})$. By normalizing $s(\boldsymbol{w})$, we obtain the spectral density $p(\boldsymbol{w}) = s(\boldsymbol{w})/\alpha$, where α is a normalization constant. We can then write the kernel as an expectation,

$$k(\boldsymbol{x}, \boldsymbol{x}') = \alpha \mathbb{E}_{p(\boldsymbol{w},\boldsymbol{b})}[e^{i\boldsymbol{w}^\intercal(\boldsymbol{x}-\boldsymbol{x}')}] = 2\alpha \mathbb{E}_{p(\boldsymbol{w},\boldsymbol{b})}[\cos(\boldsymbol{w}^\intercal \boldsymbol{x} + \boldsymbol{b}) \cos(\boldsymbol{w}^\intercal \boldsymbol{x}' + \boldsymbol{b})], \tag{10}$$

where $\boldsymbol{b} \sim \mathcal{U}(\boldsymbol{0}, 2\pi \boldsymbol{I})$. Following Rahimi and Recht [27], we sample $\boldsymbol{b}$ and $\boldsymbol{w}$ to obtain an unbiased estimate of the kernel k. From this approximation, approximate sample paths can be drawn as a weighted sum of basis functions. This form allows for fast and dense querying of the sample paths – the $\arg\max$ and $\max$ of which is an approximate draw from $p(\boldsymbol{x}^*, f^*)$. In PES, each sample $\boldsymbol{x}^*_\ell$ along with its inverted Hessian is required for computing the acquisition function. To obtain the Hessian, each sample needs to be thoroughly optimized through gradient-based optimization. JES on the other hand, only requires $(\boldsymbol{x}^*, f^*)$. As such, it can rely on cheap, approximate optimization of these samples, e.g., by densely querying sample points on a non-uniform grid.

After obtaining a set of optimal pairs $\{(\boldsymbol{x}^*_\ell, y^*_\ell)\}_{\ell=1}^L$, JES computes the conditional entropy quantity over the output y. Concretely, we generate L GPs, each modeling a posterior density $\{p(y|\mathcal{D} \cup (\boldsymbol{x}^*_\ell, f^*_\ell), \boldsymbol{x})\}_{\ell=1}^L$ conditioned on an optimal pair and previously observed data $\mathcal{D}$. Since each optimal pair is drawn from the current GP hyperparameter set, we know that the current hyperparameter set is the correct one even after adding the optimal pair to the data. By consequence, JES can compute the updated inverse Gram matrix, $(K + \sigma_\epsilon^2)^{-1}$, through a rank-1 update, instead of solving a linear system of equations. Utilizing the Sherman–Morrison formula [34], we obtain updated Gram matrices in $\mathcal{O}(n^2)$ for each sample, as opposed to $\mathcal{O}(n^3)$ for solving the linear system of equations.

3.4 Approximating the truncated entropy

As highlighted in the right panel of Fig. 2, conditioning on f^* yields a truncated normal distribution $p(f|\mathcal{D}, f^*)$, but the entropy is computed with regard to the density over noisy observations, $y = f + \epsilon$, which leads to an intractable entropy. We approximate this quantity through moment matching of the truncated Gaussian distribution over f. Consequently, we obtain two Gaussian densities $\hat{p}(f|\mathcal{D}, f^*) \sim \mathcal{N}(m_{f|f^*}, \sigma^2_{f|f^*})$ and $p(\epsilon) \sim \mathcal{N}(0, \sigma^2_\epsilon)$, where $m_{f|f^*}$ and $\sigma^2_{f|f^*}$ are the mean and variance of the truncated Gaussian posterior $p(f|\mathcal{D}, f^*)$. Due to independence between f and ϵ and the linearity of Gaussian distributions, we can then compute the entropy of the approximate density $\hat{p}_y$ exactly as $\mathrm{H}[\hat{p}(y|\mathcal{D} \cup (\boldsymbol{x}^*, f^*), \boldsymbol{x}, f^*)] = \log(2\pi(\sigma^2_\epsilon + \sigma^2_{f|f^*}))$. The quality of this approximation is studied in greater detail in Appendix E.

3.5 Greedy selection to guard against model misspecification

As with all information-theoretic approaches, JES aims to reduce the uncertainty over the location of the optimum. With this strategy, the incentive to query the perceived optimum is often lower than for heuristic approaches, such as EI. In cases where the surrogate model is misspecified, information-theoretic approaches risk reducing the entropy based on a faulty belief of the optimum, which can drastically impact their performance. As a remedy, we utilize an inverse-γ-greedy approach: with probability γ, JES will query the $\arg\max$ of the posterior mean to confirm its belief of the location of the optimum. If the model is misspecified, these greedy steps enable the algorithm to reconsider its beliefs, rather than continuing to act based on faulty ones. In Appendix F, we show how this approach can substantially improve performance in cases of surrogate model misspecification, while having negligible impact on performance in the worst case.

3.6 Putting it all together: The JES algorithm

For a sampled set of size L, containing optimal pairs $\{(\boldsymbol{x}_\ell^*, y_\ell^*)\}_{\ell=1}^L$ and GPs with mean and covariance functions $\{m_n^\ell(\boldsymbol{x}), s_n^\ell(\boldsymbol{x})\}_{\ell=1}^L$, the expression for the JES acquisition function is

$$\alpha(\boldsymbol{x})_{\texttt{JES}} = \mathrm{H}[p(y|\mathcal{D}, \boldsymbol{x})] - \mathbb{E}_{(\boldsymbol{x}^*, f^*)}\left[\mathrm{H}[p(y|\mathcal{D} \cup (\boldsymbol{x}^*, f^*), \boldsymbol{x}, f^*)]\right] \tag{11}$$

$$\approx \log(2\pi(s_n(\boldsymbol{x}) + \sigma_\epsilon^2)) - \frac{1}{L}\sum_{\ell=1}^{L} \log(2\pi(\sigma_\epsilon^2 + \sigma_{f|f^*}^2(\boldsymbol{x}; \mathcal{D}, \boldsymbol{x}_\ell^*, f_\ell^*))), \tag{12}$$

where $\sigma_{f|f^*}^2(\boldsymbol{x}; \mathcal{D}, \boldsymbol{x}_\ell^*, f_\ell^*) = \sigma_T^2(f^*; m_n^\ell(\boldsymbol{x}), s_n^\ell(\boldsymbol{x}))$ and $\sigma_T^2(\alpha; \mu, \sigma^2)$ is the variance of an upper truncated Gaussian distribution with parameters (μ, σ^2), truncated at α. The first term in 12 is simply the entropy of a Gaussian that can be computed in closed form. The second term contains both the conditioning term, which is exact, and the truncation, which is approximated as described in Sec. 3.4. In Algorithm 1, we outline the pseudocode for JES in its entirety.

Algorithm 1 JES Algorithm

Input: Black-box function f, input space $\mathcal{X}$, size M of the initial design, max number of optimization iterations N, number of posterior MC samples L, fraction of greedy samples γ.
Output: Optimized design $\boldsymbol{x}^*$.
$\mathcal{D}_M \leftarrow \{(\boldsymbol{x}_i, y_i)\}_{i=1}^M$ ▷ initial design
for $\{n = M+1, \ldots, M+N\}$ **do**
  $\mu(\boldsymbol{x}), s^2(\boldsymbol{x}) \leftarrow \textsc{FitGP}(\mathcal{D}_{n-1})$
  if $\textsc{Rand}(0,1) < \gamma$ **then** ▷ as described in Sec. 3.5
    $\boldsymbol{x}_n \leftarrow \arg\max_{\boldsymbol{x}\in\mathcal{X}} m_n(\boldsymbol{x})$
  else
    for $\{\ell = 1, \ldots, L\}$ **do**
      $(\boldsymbol{x}_\ell^*, y_\ell^*) \leftarrow \texttt{TS}(f)$ ▷ as described in Sec. 3.3
      $p(y|\mathcal{D}_{n-1} \cup (\boldsymbol{x}_\ell^*, f_\ell^*), \boldsymbol{x}) \leftarrow \textsc{UpdateGP}(\boldsymbol{x}_\ell^*, f_\ell^*)$ ▷ as described in Sec. 3.3
    end for
    $\boldsymbol{x}_n = \arg\max_{\mathcal{X}} \alpha(\boldsymbol{x})_{JES}$ ▷ defined in Eq. 11
  end if
  $y_n = f(\boldsymbol{x}_n) + \epsilon$
  $\mathcal{D}_n \leftarrow \mathcal{D}_{n-1} \cup \{(\boldsymbol{x}_n, y_n)\}$
end for
return $\boldsymbol{x}^* \leftarrow \arg\max_{\boldsymbol{x}\in\mathcal{X}} m_n(\boldsymbol{x})$

4 Experimental evaluation

Benchmarks. We now evaluate JES on a suite of diverse tasks. We consider three different types of benchmarks: samples drawn from a GP prior, commonly used synthetic test functions [15], and a collection of classification tasks on tabular data using an MLP, provided through HPOBench [9]. For the GP prior tasks, the hyperparameters are known for all methods to evaluate the effect of the acquisition function in isolation. Consequently, we do not use the inverse greedy approach from

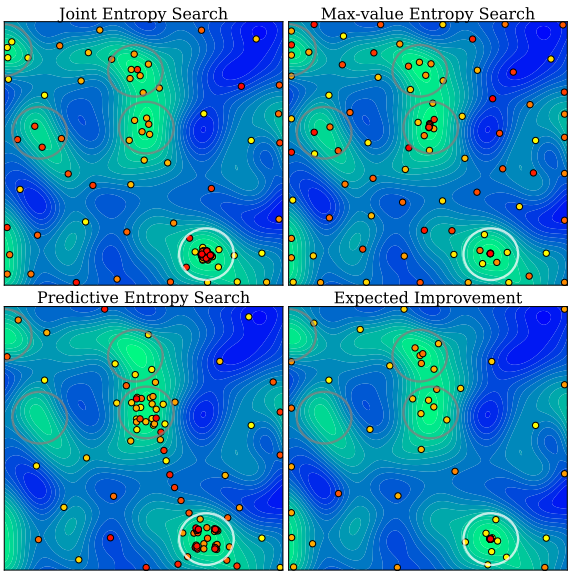

Figure 4: Comparison of queries for JES (top left), MES (top right), PES (bottom left) and EI (bottom right) on a sample of a 2D GP after a 100 function evaluations. The global optimum is circled in white, and four local optima in grey. Earlier queries are colored yellow, and later queries red.

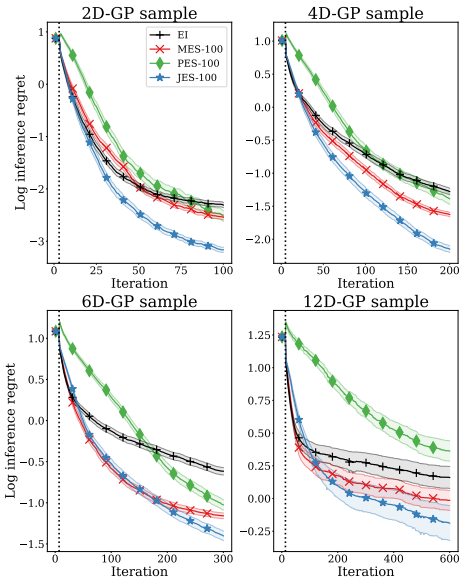

Figure 5: Comparison of JES, MES, PES and EI on GP prior samples. We run 1000 repetitions each for 2, 4 and 6D, and 250 on 12D. Mean and 2 standard errors of log regret are displayed for each acquisition function. The vertical dashed line shows the end of the initial design phase.

Section 3.5 in this case (i.e., we set $\gamma = 0$ in Algorithm 1). For the synthetic and MLP tasks, we marginalize over the GP hyperparameters, and set $\gamma = 0.1$. The hyperparameters of the GP prior experiments can be found in Appendix B, and ablation studies on γ in Appendix C.

Evaluation criteria. We use two types of evaluation criteria as in [43]: *simple regret* and *inference regret*. The simple regret $r_n = \max_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}) - \max_{t\in[1,n]} f(\boldsymbol{x}_t)$ measures the value of the best queried point so far. After a query, we may infer an $\arg\max$ of the function, which is chosen as $\boldsymbol{x}_n^* = \arg\max_{\boldsymbol{x}\in\mathcal{X}} m_n(\boldsymbol{x})$ [14, 15, 43]. We denote the inference regret as $r_n = \max_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}) - f(\boldsymbol{x}_n^*)$. Since information-theoretic approaches do not necessarily seek to query the optimum, but only to know its location, inference regret characterizes how satisfying our belief of the $\arg\max$ is. Notably, this metric is non-monotonic, meaning that the best guess can worsen with time. We use this metric in the ideal model benchmarking setting, when we sample tasks from a GP with known hyperparameters. We use simple regret for the synthetic test functions, as it constitutes a metric that is more robust to surrogate model misspecification. Inference regret for these tasks can be found in Appendix G. For the HPOBench tasks, inference regret is unobtainable.

The experimental setup. We compare against other state-of-the-art information-theoretic approaches: PES [15] and MES [43], as well as EI [20]. The acquisition functions are all run in the same framework written in MATLAB, created for the original PES implementation by Hernández-Lobato et al. [15]. All synthetic experiments were run for $50D$ iterations. In the main paper, we fix the number of MC samples for MES, PES and JES to 100 each. In Appendix D we assess the sensitivity of JES to this number and quantify the computational expense. In Appendix B, we provide all details on our experimental setup.

4.1 GP prior samples

We consider samples from a GP prior for four different dimensionalities: 2D, 4D, 6D, and 12D, with a noise standard deviation of 0.1 for a range of outputs spanning roughly $[-10, 10]$. These tasks constitute an optimal setting for each algorithm, as the surrogate perfectly models the task at hand. JES demonstrates empirically the value of the additional source of information, finishing at least half an order of magnitude ahead on all tasks except for 12D.

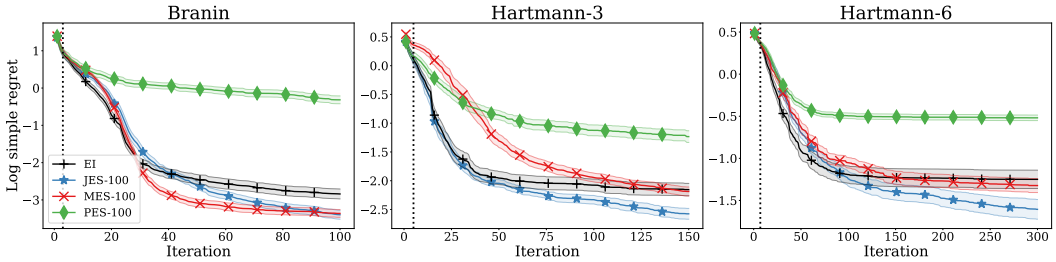

Figure 7: Comparison of JES, MES, PES and EI on Branin and Hartmann-6, $\sigma_n^2 = 0.10$. Mean and 2 standard errors of log regret are displayed for each acquisition function across 100 repetitions. The vertical dashed line represents the end of the initial design phase.

Fig. 4 compares JES (top left) against PES, MES and EI in terms of point selection for one repetition on a two-dimensional sample task, where all runs are initialized with $D+1$ identical random samples. We observe that JES succeeds in finding all attractive regions of the search space, and queries the region around the optimum densely, which is sensible in a noisy setting. We further notice that EI (bottom right) fails to query the two circled local optima. PES (bottom left) also ignores two local optima to various degrees, and tends to circle the (perceived) optimum densely, which is expensive in terms of number of evaluations. We believe this showcases a shortcoming of only considering the density over the optimum: PES circles the optimum, but does not query its value. Lastly, MES (top right) successfully queries all attractive regions of the space, but also samples regions that are evidently poor the most densely out of the four approaches, despite information given by earlier (brighter) samples. Since JES considers the information conveyed by both MES and PES, it successfully excludes the apparent sub-optimal regions of the space, finds all relevant optima, and queries these optima in a desirable manner.

We additionally evaluate the performance of all approaches on GP sample tasks that have a substantial amount of noise - its standard deviation roughly accounting for 10% of the total output range. We run these tasks for a larger number of iterations, 125D, to display the stagnation of some approaches. While MES and PES slow down approximately at the halfway point for both tasks, JES steadily improves for the entire length of the run. This robustness to large noise magnitudes highlights the importance of intrinsically handling noisy objectives in JES.

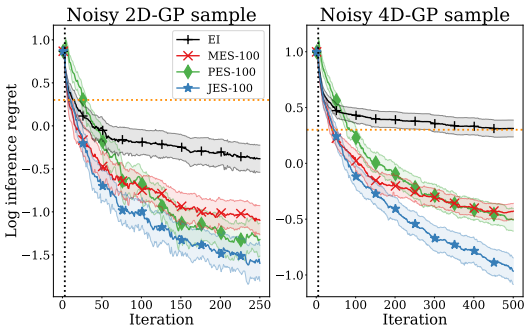

Figure 6: Evaluation of JES, MES, PES and EI on noisy ($\sigma_\epsilon^2 = 4$, orange) GP sample tasks across 100 repetitions. Mean and 2 standard errors of log regret are displayed for each acquisition function.

4.2 Synthetic test functions

Next, in Fig. 7, we study the performance of JES on three optimization test functions: Branin (2D), Hartmann (3D) and Hartmann (6D). For these tasks, we follow convention [15, 29] and marginalize over GP hyperparameters. On Branin, JES starts out slightly slower than MES but reaches the same performance in 100 iterations; and on the two Hartmann functions, JES performs amongst the best in the beginning and clearly best in the end. We note that PES experienced numerical issues on Branin, and as such, we acknowledge that its performance should be better than what is reported.

4.3 MLP tasks

Lastly, we evaluate the performance of JES on the tuning an MLP model's 4 hyperparameters for $20D$ iterations on three datasets. These tasks are part of the OpenML[1] library of tasks, and the HPO benchmark is provided through the HPOBench [9] suite. The tasks, which are classification problems with tabular data, regard classifying: (1) vehicles into one of four different categories, (2) blood donors for a blood transfusion service center, and (3) anonymous credit card applications.

[1] https://www.openml.org/

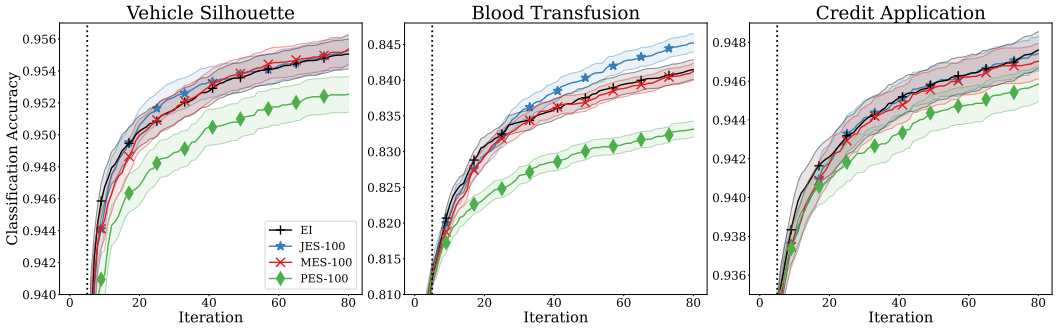

Figure 8: Comparison of JES, MES, PES and EI on three different MLP tuning tasks from the HPOBench suite. Mean and 1 standard error of best observed accuracy are displayed for each acquisition function across 150 repetitions. The vertical dashed line represents the end of the initial design phase.

We measure the best observed classification accuracy. Notably, these tasks have a large amount of noise, which causes the performance to fluctuate substantially between repetitions. We observe that JES performs almost identically to MES and EI on vehicle silhouette and credit application, and substantially better on blood transfusion. PES lags behind significantly on all tasks.

5 Conclusions

We have presented Joint Entropy Search, an information-theoretic acquisition function that considers an entirely new quantity, namely the joint density over the optimum and optimal value. By utilizing the entropy reduction from fantasized optimal observations, JES obtains a simple form for the entropy reduction regarding the joint distribution. As such, the additional information considered comes with minimal computational overhead, avoids restrictive assumptions on the objective, and yields state-of-the-art performance along with superior decision-making. We believe JES to be a new go-to acquisition function for BO, and to establish a new standard for subsequent information-theoretic techniques.

6 Limitations and Future Work

The main contribution of this paper is to provide a novel information-theoretic acquisition function which, given a sufficiently accurate model, yields impressive results. However, the non-myopic, speculative nature of information-theoretic approaches lend them to be susceptible to model misspecification, such as a poor choice of GP kernel or GP hyperparameters. In our view, information-theoretic approaches are possibly more susceptible to this issue than their myopic counterparts (EI, UCB, TS). While we propose a remedy to stabilize and improve the acquisition function under model misspecification with the inverse γ-greedy approach, this technique only serves to *discover* misspecification and adjust accordingly, but not to inherently fix the misspecification. We believe misspecification can only be remedied by altering the surrogate model. It is thus very promising to combine advanced modelling techniques with information-theoretic acquisition functions, as already done with the additive GP approach utilized in conjunction with MES by Wang and Jegelka [43]; further promising additions would be to tackle heterogeneous noise and input warping as done by HEBO [8].

We also note that, since JES computes the entropy reduction from conditioning on the optimal pair, it relies on some level of noise in the objective. A surrogate model with zero noise will result in an infinite information gain for every optimal pair, which (by utilizing some random tie-breaking strategy) would make JES equivalent to TS. However, if JES is to be used in a completely noiseless setting, we argue that a small noise term should be added as a remedy. As this is done by default in many prominent GP frameworks [12, 13, 41], we do not view this as a major limitation of our approach. Nevertheless, improving upon this strategy would be interesting in future work.

For future work, we also envision work on the adaptation of JES to various different domains, such as multi-fidelity [44] and multi-objective optimization [1], as well as the integration of user prior knowledge over the location of the optimum [19] to accelerate optimization.

References

[1] S. Belakaria, A. Deshwal, and J. R. Doppa. Max-value entropy search for multi-objective bayesian optimization. *Advances in neural information processing systems*, 32, 2019.

[2] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for Hyper-Parameter Optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 24. Curran Associates, Inc., 2011.

[3] F. Berkenkamp, A. Krause, and A. Schoellig. Bayesian optimization with safety constraints: Safe and automatic parameter tuning in robotics. *Machine Learning*, 06 2021. doi: 10.1007/s10994-021-06019-1.

[4] S. Bochner et al. *Lectures on Fourier integrals*, volume 42. Princeton University Press, 1959.

[5] A. D. Bull. Convergence rates of efficient global optimization algorithms. 12:2879–2904, 2011.

[6] R. Calandra, N. Gopalan, A. Seyfarth, J. Peters, and M. Deisenroth. Bayesian gait optimization for bipedal locomotion. In P. Pardalos and M. Resende, editors, *Proceedings of the Eighth International Conference on Learning and Intelligent Optimization (LION'14)*, 2014.

[7] Y. Chen, A. Huang, Z. Wang, I. Antonoglou, J. Schrittwieser, D. Silver, and N. de Freitas. Bayesian optimization in alphago. *CoRR*, abs/1812.06855, 2018. URL `http://arxiv.org/abs/1812.06855`.

[8] A. I. Cowen-Rivers, W. Lyu, Z. Wang, R. Tutunov, J. Hao, J. Wang, and H. Bou-Ammar. HEBO: heteroscedastic evolutionary bayesian optimisation. *CoRR*, abs/2012.03826, 2020. URL `https://arxiv.org/abs/2012.03826`.

[9] K. Eggensperger, P. Müller, N. Mallik, M. Feurer, R. Sass, A. Klein, N. Awad, M. Lindauer, and F. Hutter. HPOBench: A collection of reproducible multi-fidelity benchmark problems for HPO. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL `https://openreview.net/forum?id=1k4rJYEwda-`.

[10] A. Ejjeh, L. Medvinsky, A. Councilman, H. Nehra, S. Sharma, V. Adve, L. Nardi, E. Nurvitadhi, and R. A. Rutenbar. Hpvm2fpga: Enabling true hardware-agnostic fpga programming. In *Proceedings of the 33rd IEEE International Conference on Application-specific Systems, Architectures, and Processors*, 2022.

[11] P. I. Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[12] J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson. Gpytorch: Black-box matrix-matrix gaussian process inference with gpu acceleration. In *Advances in Neural Information Processing Systems*, 2018.

[13] GPy. GPy: A gaussian process framework in python. `http://github.com/SheffieldML/GPy`, since 2012.

[14] P. Hennig and C. J. Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13(1):1809–1837, June 2012. ISSN 1532-4435.

[15] J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in Neural Information Processing Systems*, 2014. URL `https://proceedings.neurips.cc/paper/2014/file/069d3bb002acd8d7dd095917f9efe4cb-Paper.pdf`.

[16] J. M. Hernández-Lobato, M. Gelbart, M. Hoffman, R. Adams, and Z. Ghahramani. Predictive entropy search for bayesian optimization with unknown constraints. In *International conference on machine learning*, pages 1699–1707. PMLR, 2015.

[17] M. W. Hoffman and Z. Ghahramani. Output-space predictive entropy search for flexible global optimization. 2016.

[18] F. Hutter, H. H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *Learning and Intelligent Optimization*, 2011.

[19] C. Hvarfner, D. Stoll, A. Souza, L. Nardi, M. Lindauer, and F. Hutter. PiBO: Augmenting Acquisition Functions with User Beliefs for Bayesian Optimization. In *International Conference on Learning Representations*, 2022.

[20] D. Jones, M. Schonlau, and W. Welch. Efficient global optimization of expensive black box functions. 13:455–492, 1998.

[21] H. J. Kushner. A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise. *Journal of Basic Engineering*, 86(1):97–106, 03 1964. ISSN 0021-9223. doi: 10.1115/1.3653121. URL `https://doi.org/10.1115/1.3653121`.

[22] B. Matern. *Spatial Variation*. Lecture Notes in Statistics. Springer New York, 2013. ISBN 9781461578925. URL `https://books.google.se/books?id=HrbSBwAAQBAJ`.

[23] M. Mayr, F. Ahmad, K. I. Chatzilygeroudis, L. Nardi, and V. Krüger. Skill-based Multi-objective Reinforcement Learning of Industrial Robot Tasks with Planning and Knowledge Integration. *CoRR*, abs/2203.10033, 2022. URL `https://doi.org/10.48550/arXiv.2203.10033`.

[24] J. Mockus, V. Tiesis, and A. Zilinskas. The application of Bayesian methods for seeking the extremum. *Towards Global Optimization*, 2(117-129):2, 1978.

[25] L. Nardi, D. Koeplinger, and K. Olukotun. Practical design space exploration. In *2019 IEEE 27th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, pages 347–358. IEEE, 2019.

[26] Q. P. Nguyen, B. K. H. Low, and P. Jaillet. Rectified max-value entropy search for bayesian optimization, 2022. URL `https://arxiv.org/abs/2202.13597`.

[27] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, Advances in Neural Information Processing Systems, page 1177–1184, Red Hook, NY, USA, 2007. Curran Associates Inc. ISBN 9781605603520.

[28] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[29] B. Ru, M. A. Osborne, M. Mcleod, and D. Granziol. Fast information-theoretic Bayesian optimisation. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4384–4392. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/ru18a.html`.

[30] D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. *Advances in Neural Information Processing Systems*, 27, 2014.

[31] K. Šehić, A. Gramfort, J. Salmon, and L. Nardi. LassoBench: A High-Dimensional Hyperparameter Optimization Benchmark Suite for Lasso. *arXiv preprint arXiv:2111.02790*, 2021.

[32] A. Shah and Z. Ghahramani. Parallel predictive entropy search for batch global optimization of expensive objective functions. *Advances in neural information processing systems*, 28, 2015.

[33] B. Shahriari, K. Swersky, Z. Wang, R. Adams, and N. de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.

[34] J. Sherman and W. J. Morrison. Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix. *The Annals of Mathematical Statistics*, 21(1):124 – 127, 1950. doi: 10.1214/aoms/1177729893. URL `https://doi.org/10.1214/aoms/1177729893`.

[35] J. Snoek, H. Larochelle, and R. Adams. Practical Bayesian optimization of machine learning algorithms. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Proceedings of the 26th International Conference on Advances in Neural Information Processing Systems (NeurIPS'12)*, pages 2960–2968, 2012.

[36] J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. Patwary, Prabhat, and R. Adams. Scalable Bayesian optimization using deep neural networks. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML'15)*, volume 37, pages 2171–2180. Omnipress, 2015.

[37] J. Springenberg, A. Klein, S. Falkner, and F. Hutter. Bayesian optimization with robust Bayesian neural networks. In D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Proceedings of the 30th International Conference on Advances in Neural Information Processing Systems (NeurIPS'16)*, 2016.

[38] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, May 2012. ISSN 1557-9654. doi: 10.1109/tit.2011.2182033. URL `http://dx.doi.org/10.1109/TIT.2011.2182033`.

[39] S. Takeno, H. Fukuoka, Y. Tsukada, T. Koyama, M. Shiga, I. Takeuchi, and M. Karasuyama. Multi-fidelity Bayesian optimization with max-value entropy search and its parallelization. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9334–9345. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/takeno20a.html`.

[40] W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[41] J. Vanhatalo, J. Riihimäki, J. Hartikainen, P. Jylänki, V. Tolvanen, and A. Vehtari. Gpstuff: Bayesian modeling with gaussian processes. *Journal of Machine Learning Research*, 14: 1175–1179, Apr. 2013. ISSN 1532-4435.

[42] J. Villemonteix, E. Vazquez, and E. Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization*, 44, 12 2006. doi: 10.1007/s10898-008-9354-2.

[43] Z. Wang and S. Jegelka. Max-value entropy search for efficient bayesian optimization. In *International Conference on Machine Learning (ICML)*, 2017.

[44] Y. Zhang, T. N. Hoang, B. K. H. Low, and M. Kankanhalli. Information-based multi-fidelity bayesian optimization. In *NeurIPS Workshop on Bayesian Optimization*, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Our abstract and introduction accurately reflect our paper.
   (b) Did you describe the limitations of your work? [Yes] See Section 6
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section A
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We discuss the ethics guidelines in Section A
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] As our method is proposed as an heuristic, we do not include theoretical results.
   (b) Did you include complete proofs of all theoretical results? [N/A] As our method is proposed as an heuristic, we do not include theoretical results.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The URL to the repository where all experiments were run is included in Section 1
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] All experimental details are reported in Appendix B
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All results have error bars include.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section A
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] All the repositories that we have used or built on have been cited.
   (b) Did you mention the license of the assets? [Yes] The licenses of the assets are mentioned in Appendix B
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] e include new surrogate benchmarks and frameworks in the supplementary material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We did not use or release any datasets with personal data.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The data we are using does not contain personal information or offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not run experiments with human subjects.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We did not run experiments with human subjects.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We did not run experiments with human subjects.