Joint Entropy Search
For Maximally-Informed Bayesian Optimization

Carl Hvarfner CARL.HVARFNER@CS.LTH.SE
Lund University

Frank Hutter FH@CS.UNI-FREIBURG.DE
University of Freiburg & Bosch Center for Artificial Intelligence

Luigi Nardi LUIGI.NARDI@CS.LTH.SE
Lund University & Stanford University

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` 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) (Mockus et al., 1978; Shahriari et al., 2016) is a sample-efficient approach, and has been successfully applied to various problems. 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) (Hennig and Schuler, 2012) and *Predictive Entropy Search* (`PES`) (Hernández-Lobato et al., 2014) select queries by maximizing this quantity. A related information-theoretic family of approaches considers the information gain on the optimal objective value (Hoffman and Ghahramani, 2016). *Max-value Entropy Search* (`MES`) (Wang and Jegelka, 2017) considers a one-dimensional density over the output space as

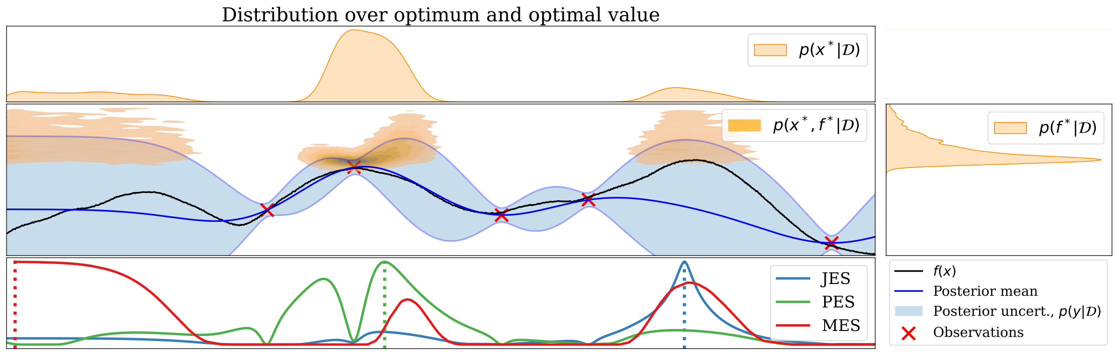

Figure 1: The densities considered by ES/PES (top), MES (right) and JES (center) on a toy function. The multimodal density $p(\boldsymbol{x}^*, f^*)$ is reduced to a heavy-tailed density over f^* for the density used by MES (right), which does not capture the multi-modality of the density over the optimum.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

opposed to a D-dimensional input space, and yields a computationally efficient alternative to ES and PES. However, shortcomings of MES have been highlighted in recent works (Takeno et al., 2020; Nguyen et al., 2022). In particular, MES does not differentiate between the (unobserved) maximal objective value f^* and an observed noisy maximum y_{max}.

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 innovates on previous information-theoretic approaches by: (1) utilizing two sources of information, by considering the entropy over both the optimum and optimal value; (2) Retaining low computational and conceptual complexity by utilizing the full optimal observation, allowing it to rely on standard GP machinery instead of complex approximations; (3) Intrinsically supporting noisy objective functions, making it particularly efficient for these types of tasks.

2. Joint Entropy Search

We now present Joint Entropy Search (JES), a novel information-theoretic acquisition function. Similarly to previous work, JES considers a mutual information quantity. Unlike its predecessors, JES considers the entropy of the optimal distribution over *both* the input and output space. It utilizes a 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*.

The joint density. 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

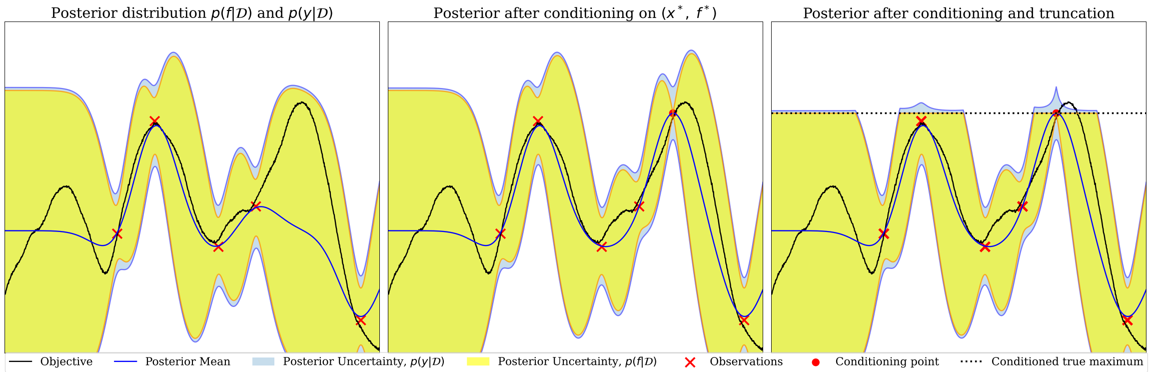

Figure 2: Step-by-step modeling when conditioning on one optimal pair $(\boldsymbol{x}^*, f^*)$. The posterior with noise $p(y|\mathcal{D})$ and without noise $p(f|\mathcal{D})$ are illustrated in blue and yellow, respectively. The GP after 5 (noisy) observations, before conditioning on $(\boldsymbol{x}^*, f^*)$ 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.

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.

The Joint Entropy Search acquisition function. We consider the mutual information between the random variables $(\boldsymbol{x}^*, f^*)$ and a future query $(\boldsymbol{x}, y)$:

$$\begin{aligned}\alpha_{\texttt{JES}}(\boldsymbol{x}) &= I((\boldsymbol{x}, y); (\boldsymbol{x}^*, f^*)|\mathcal{D}) & (1)\\ &= \mathrm{H}[p(y|\mathcal{D}, \boldsymbol{x})] - \mathbb{E}_{(\boldsymbol{x}^*, f^*)}\left[\mathrm{H}[p(y|\mathcal{D}, \boldsymbol{x}, \boldsymbol{x}^*, f^*)]\right] & (2)\\ &= \mathrm{H}[p(y|\mathcal{D}, \boldsymbol{x})] - \mathbb{E}_{(\boldsymbol{x}^*, f^*)}\left[\mathrm{H}[p(y|\mathcal{D} \cup (\boldsymbol{x}^*, f^*), \boldsymbol{x}, f^*)]\right]. & (3)\end{aligned}$$

Eq. 2 is similar to `MES` 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. 3, 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 (Nguyen et al., 2022; Wang and Jegelka, 2017), the resulting effect is to truncate the GP's posterior over f, upper bounding it to f^*. The expectation in Eq. 3 is approximated through MC by sampling L optimal pairs $\{(\boldsymbol{x}_\ell^*, f_\ell^*)\}_{\ell=1}^L$ from $p(\boldsymbol{x}^*, f^*)$ using approximate *Thompson Sampling* (`TS`) (Thompson, 1933). 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 is easily computed exactly, and the latter requires approximation. In Fig. 2, the resulting posterior distribution of the two-step conditioning is shown in detail.

Incorporating optimal pairs. To obtain samples $(\boldsymbol{x}^*, f^*)$, we utilize an approximate variant of TS, originally proposed in PES (Hernández-Lobato et al., 2014). This form allows for fast and dense querying of the sample paths – the arg max and max of which is an approximate draw from $p(\boldsymbol{x}^*, f^*)$. For PES, each sample $\boldsymbol{x}^*_\ell$ and 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 however, only requires the observation $(\boldsymbol{x}^*_\ell, f^*_\ell)$. As such, it can rely on cheap, approximate optimization of these sample paths, e.g., by densely querying sample points. After obtaining a set of optimal pairs $\{(\boldsymbol{x}^*_\ell, y^*_\ell)\}_{\ell=1}^L$, we compute the conditional entropy quantity over y. Concretely, we generate L GPs, each modeling a density $\{p(y|\mathcal{D} \cup (\boldsymbol{x}^*_\ell, f^*_\ell), \boldsymbol{x})\}_{\ell=1}^L$ conditioned on an optimal pair and the observed data $\mathcal{D}$. As each optimal pair is drawn given the current GP hyperparameters, we know that the current hyperparameters are correct even after adding the optimal pair to the data. Thus, we can compute the updated inverse Gram matrix, $(K + \sigma_\epsilon^2)^{-1}$, through a rank-1 update (Sherman and Morrison, 1950) instead of solving a linear system of equations. Updated Gram matrices are then obtained in $\mathcal{O}(n^2)$ per sample, as opposed to $\mathcal{O}(n^3)$ for solving the system of equations.

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 $p(\epsilon) \sim \mathcal{N}(0, \sigma_\epsilon^2)$, where $m_{f|f^*}$ and $\sigma^2_{f|f^*}$ are the mean and variance of the truncated Gaussian posterior $p(f|\mathcal{D}, f^*)$. We can then compute the entropy of the approximate density $\hat{p}_y$ as $\mathrm{H}[\hat{p}(y|\mathcal{D} \cup (\boldsymbol{x}^*, f^*), \boldsymbol{x}, f^*)] = \log(2\pi(\sigma_\epsilon^2 + \sigma^2_{f|f^*}))$. The quality of this approximation is studied in greater detail in Appendix E

Greedy selection to guard against model misspecification. JES, like other information-theoretic approaches, 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 queries 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 on faulty ones.

Putting it all together. For a set $\{(\boldsymbol{x}^*_\ell, y^*_\ell)\}_{\ell=1}^L$ and GPs with mean and covariance functions $\{m^\ell(\boldsymbol{x}), s^\ell(\boldsymbol{x})\}_{\ell=1}^L$, the expression for the JES acquisition function is

$$\alpha_{\texttt{JES}}(\boldsymbol{x}) = \mathrm{H}[p(y|\mathcal{D}, \boldsymbol{x})] - \mathbb{E}_{(\boldsymbol{x}^*, f^*)}\left[\mathrm{H}[p(y|\mathcal{D} \cup (\boldsymbol{x}^*, f^*), \boldsymbol{x}, f^*)]\right] \tag{4}$$

$$\approx \log(2\pi(s(\boldsymbol{x}) + \sigma_\epsilon^2)) - \frac{1}{L}\sum_{\ell=1}^{L} \log(2\pi(\sigma_\epsilon^2 + \sigma^2_{f|f^*}(\boldsymbol{x}; \mathcal{D}, \boldsymbol{x}^*_\ell, f^*_\ell))), \tag{5}$$

where $\sigma^2_{f|f^*}(\boldsymbol{x}; \mathcal{D}, \boldsymbol{x}^*_\ell, f^*_\ell) = \sigma_T^2(f^*; m^\ell(\boldsymbol{x}), s^\ell(\boldsymbol{x}))$ 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. To form our strategy, we interleave queries chosen by $\arg\max \alpha_{\texttt{JES}}(\boldsymbol{x})$ with a fraction γ of greedy queries.

Task	JES-100	MES-100	EI	PES-100
2D	1.40 ± 0.32	1.03 ± 0.19	0.23 ± 0.13	17.39 ± 4.95
4D	1.50 ± 0.37	1.21 ± 0.3	0.3 ± 0.17	34.53 ± 8.3
6D	1.56 ± 0.39	1.26 ± 0.37	0.35 ± 0.2	62.92 ± 13.54

Table 1: Runtime of `JES`, `MES`, `PES` and `EI` on GP sample tasks of varying dimensionalities. JES is only marginally slower than MES, and orders of magnitude faster than PES.

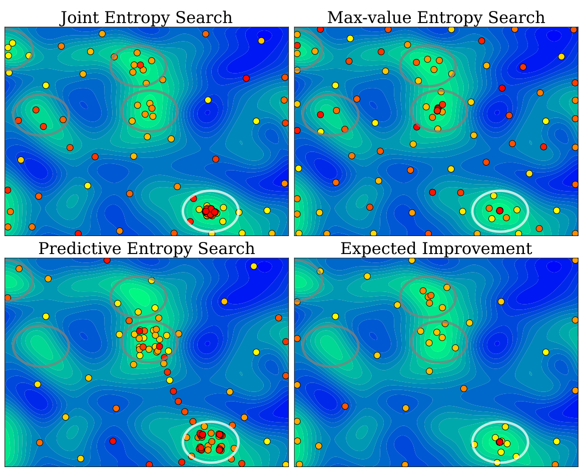

Figure 3: Comparison of queries on a 2D GP sample, 100 function evaluations. The global optimum is circled in white, and four local optima in grey. Earlier queries are colored yellow, and later queries red.

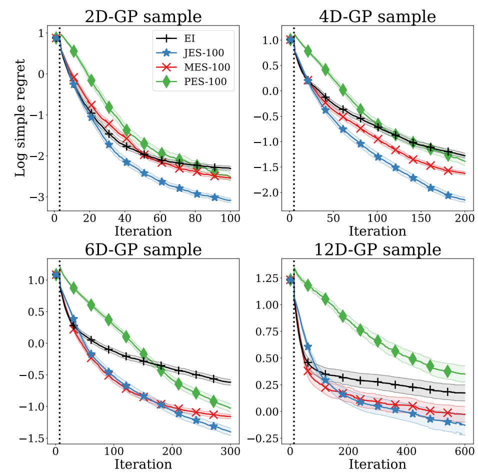

Figure 4: Comparison on GP prior samples. We run 1000 repetitions each for 2, 4 and 6D, and 250 on 12D. Mean and 2 SE of log regret are displayed.

3. Experimental evaluation

Benchmarks. We now evaluate `JES` on a suite of diverse tasks: Samples from a GP prior, synthetic test functions, and MLP tuning from HPOBench (Eggensperger et al., 2021). We do not use the inverse γ-greedy approach for the GP tasks (i.e., we set $\gamma = 0$) and report mean and standard error (SE) of *inference regret*. For the synthetic and MLP tasks, we marginalize over the GP hyperparameters, and set $\gamma = 0.1$, and report *simple regret*. The experimental setup can be found in Appendix B, and ablation studies on γ in Appendix C.

GP prior samples. We consider samples from a GP prior for four different dimensionalities: 2D, 4D, 6D, and 12D. In this optimal setting with fixed GP hyperparameters, `JES` outperforms all other acquisition functions by substantial amounts on all tasks. In Tab. 1, we display runtimes for all acquisition functions when GP hyperparameters are marginalized over.

Fig. 3 compares `JES` queries (top left) against `PES`, `MES` and `EI` for one repetition on a two-dimensional GP sample task. `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. `EI` (bottom right) fails to query two local optima. `PES` (bottom left) also ignores two local optima, and circles the (perceived) optimum densely, which is costly in terms of number of evaluations. Lastly, `MES` (top right) successfully queries all attractive regions of the space, but also samples regions that are evidently

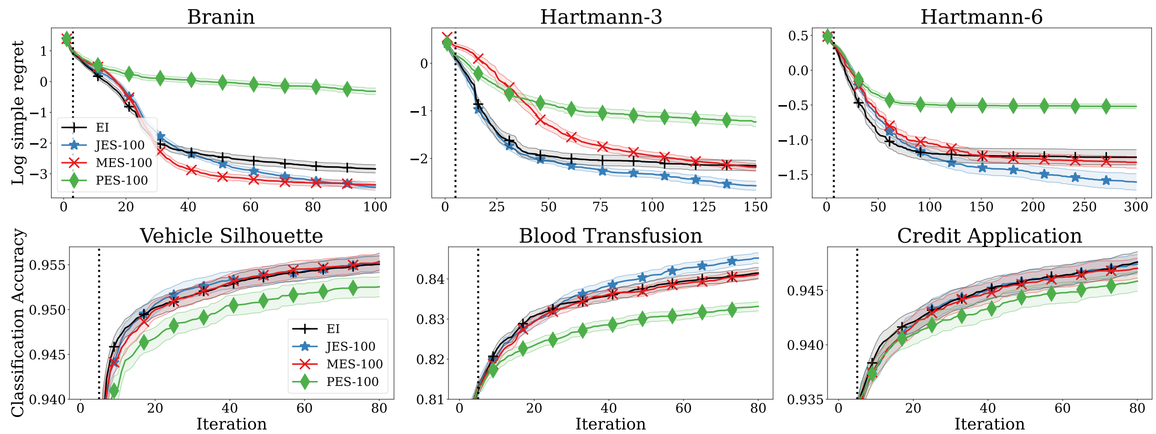

Figure 6: Comparison on synthetic test functions (top) and HPOBench (bottom). Mean and 2 (top) and 1 (bottom) SE are displayed for each acquisition function across 100/150 repetitions.

poor the most densely. `JES` successfully excludes the apparent suboptimal 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. While `MES` and `PES` slow down approximately at the halfway point for both tasks, `JES` steadily improves for the entire length of the run.

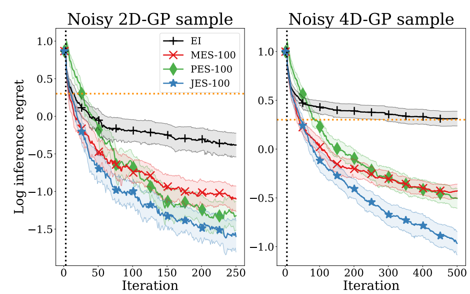

Figure 5: Evaluation on noisy ($\sigma_\epsilon^2 = 4$, orange) GP sample tasks across 100 repetitions. Mean and 2 SE of log regret.

Synthetic test functions. In Fig. 6, we study `JES` on Branin (2D), Hartmann (3D) and Hartmann (6D). On Branin, `JES` starts out slightly slower than MES but ultimately reaches the same performance; and on the two Hartmann functions, JES performs amongst the best in the beginning and clearly best in the end.

MLP tasks. Lastly, we evaluate `JES` on the tuning an MLP model's 4 hyperparameters on three datasets. These tasks are part of the OpenML (Vanschoren et al., 2014) library of tasks. The tasks are classification problems with tabular data. We observe that `JES` performs almost identically to `MES` and `EI` on two tasks, and substantially better on one.

4. Conclusion

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.

References

Katharina Eggensperger, Philipp Müller, Neeratyoy Mallik, Matthias Feurer, Rene Sass, Aaron Klein, Noor Awad, Marius Lindauer, and Frank 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-`.

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.

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`.

Matthew W. Hoffman and Zoubin Ghahramani. Output-space predictive entropy search for flexible global optimization. 2016.

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

Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet. Rectified max-value entropy search for bayesian optimization, 2022. URL `https://arxiv.org/abs/2202.13597`.

Ali Rahimi and Benjamin 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.

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.

Jack Sherman and Winifred 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`.

Shion Takeno, Hitoshi Fukuoka, Yuhki Tsukada, Toshiyuki Koyama, Motoki Shiga, Ichiro Takeuchi, and Masayuki Karasuyama. Multi-fidelity Bayesian optimization with max-value entropy search and its parallelization. In Hal Daumé III and Aarti 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`.

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.

Jarno Vanhatalo, Jaakko Riihimäki, Jouni Hartikainen, Pasi Jylänki, Ville Tolvanen, and Aki Vehtari. Gpstuff: Bayesian modeling with gaussian processes. *Journal of Machine Learning Research*, 14:1175–1179, April 2013. ISSN 1532-4435.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luís Torgo. Openml: networked science in machine learning. *CoRR*, abs/1407.7722, 2014. URL `http://arxiv.org/abs/1407.7722`.

Zi Wang and Stefanie Jegelka. Max-value entropy search for efficient bayesian optimization. In *International Conference on Machine Learning (ICML)*, 2017.

D	θ_d	σ^2	σ_ϵ^2	Range
2	0.1	10	0.01	$[-9, 9]$
4	0.2	10	0.01	$[-11, 11]$
6	0.3	10	0.01	$[-13, 13]$
12	0.6	10	0.01	$[-18, 18]$

Table 2: Hyperparameters for the generated GP sample tasks.

Appendix A. Broader impact

Our work proposes a novel acquisition function for Bayesian optimization. The approach is foundational and does not have direct societal or ethical consequences. However, `JES` will be used in the development of applications for a wide range of areas and thus indirectly contribute to their impacts on society. As an algorithm that can be used for HPO, `JES` intends to cut resource expenditure associated with model training, while increasing their performance. This can help reduce the environmental footprint of machine learning research.

Appendix B. Experimental Setup

Frameworks. For all tasks and acquisition functions, we use the original `PES` implementation in MATLAB by Hernández-Lobato et al. (2014), which uses the GPStuff (Vanhatalo et al., 2013) library. The implementation optimizes the acquisition function, and the posterior mean, by sampling a dense grid of points, and uses a gradient-based optimizer to further optimize the single best point. For better accuracy, we substantially increased the number of grid points. The `JES` implementation can be found at `https://github.com/jointentropysearch/JointEntropySearch`.

GP Sample tasks. To generate the GP sample tasks, we use a random Fourier feature (Rahimi and Recht, 2007) with weights drawn from the spectral density of a squared exponential kernel. For dimension-wise length scale θ_d, output scale σ^2, and noise variance σ_ϵ^2, the hyperparameters per task are shown in Table 2. The range in the last column is a rough approximation of the magnitude of the output spanned by each GP sample. The length scales of the samples are gradually increased with each dimensionality to maintain a reasonable level of difficulty for all tasks. Since the optimal values for these tasks are unavailable, they are approximated through a dense random search, followed by local search on the most promising subset of points.

Runtime tests. For the runtime tests, we utilized each acquisition function's original MATLAB implementation. Each acquisition function is timed on the whole procedure of selecting the next query – starting from the pre-computation involved with `JES`, `MES` and `PES`, until the subsequent query is ultimately selected. As such, the time spent sampling GP hyperparameters, inverting the Gram matrix, and evaluating the objective function is excluded. In Tab. 3, all relevant hyperparameters for the runtime of each acquisition function are displayed: The number of sampled sets of hyperparameters (No. HP sets), the number of maxima per sampled set of hyperparameters (No. maxima per set), the number of randomly

Approach	No. HP sets	No. maxima per set	No. RS	No. features
JES	10	10	10000	1000
MES	10	10	10000	N/A
EI	10	N/A	10000	N/A
PES	10	10	1000	1000

Table 3: Settings for the runtime tests. The JES settings are chosen to mimic MES/PES as closely as possible.

Task	No. HP sets	No. maxima per set	Update frequency	σ_ϵ^2
Branin	20	5	5	0.01
Hartmann (3D)	20	5	10	0.01
Hartmann (6D)	20	5	10	0.01
MLP classification	20	5	5	0.0

Table 4: GP Hyperparameter sets and updates for the synthetic test functions and MLP tasks.

sampled points evaluated on the acquisition function (No. RS) and the number of Fourier features (No. features) employed when approximately sampling optima.

Synthetic test functions. For the synthetic test functions, 100 sampled optimal pairs are used for each acquisition function. GP hyperparameters are marginalized over for these tasks, so an equal number of optimal pairs are sampled for each hyperparameter set. The hyperparameters are re-sampled on a fixed schedule throughout the run. Naturally, the sampled maxima were updated at each iteration. Moreover, each test function was also given a fixed amount of noise. Regret was computed not from the noisy observed value, but from the true, noiseless function value.

MLP classification tasks. All the classification tasks have a substantial amount of noise. As noiseless objective values are unavailable, we report the observed classification accuracy. 5 hyperparameters are available in HPOBench for these tasks. However, one of them (number of layers, *depth*) is held fixed due to its small integer-valued domain and the lack of integer hyperparameter support in the MATLAB framework. As seen in Tab. 5, the two other integer-valued hyperparameters batch size and *width* have orders of magnitude larger domains, and can therefore reasonably be treated as continuous hyperparameters. All tasks are optimized in the [0, 1] range, and are scaled, transformed, and rounded to the nearest integer in the objective function where necessary. All non-fixed parameters are evaluated in log scale. The three tasks evaluated are *Australian*, *Blood-transfusion-service-center*, and *Vehicle*, with HPOBench task numbers #146818, #10101, #53, respectively.

Compute resources. All experiments are carried out on *Intel Xeon Gold 6130* CPUs. Each repetition is run on a single core. In total, approximately 50,000 core hours are used for the experiments in the main paper, and an additional 20,000 for the appendix.

Name	Type	Range
Alpha (L2)	Continuous	$[10^{-8}, 10^{-3}]$
Batch size	Integer	$[2^2, 2^8]$
Depth	Fixed to 2	$\{1, 2, 3\}$
Initial learning rate	Continuous	$[10^{-5}, 1]$
Width	Integer	$[2^4, 2^{10}]$

Table 5: Search space for the MLP tasks.

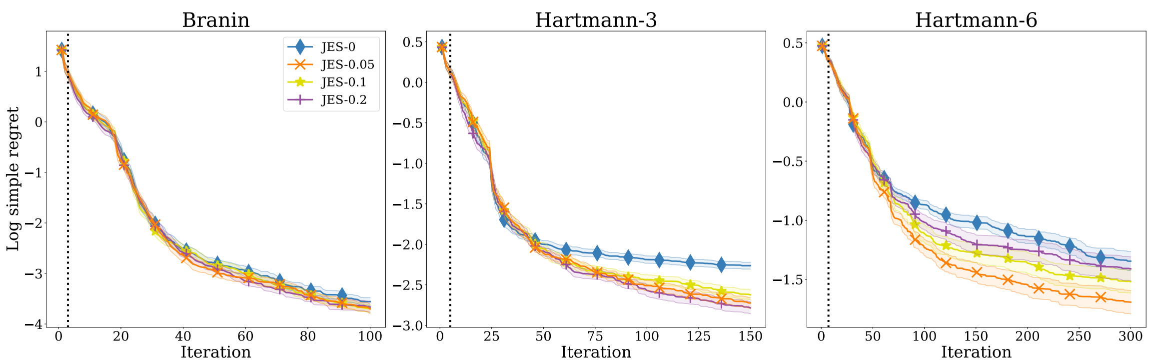

Figure 7: Comparison of JES with varying fraction γ of inverse greedy selections. Mean and 1 standard error of log simple regret is displayed for all tasks.

Appendix C. Ablation Studies

We provide ablation studies for the hyperparameter controlling the ratio of inverse greedy selection γ and the noise variance σ_ϵ^2.

C.1 Ablation study on γ

We provide an ablation study on γ in terms of both the simple and inference regret of JES. While simple regret may be a more practically relevant metric, inference regret helps understand the ability of the acquisition function to successfully locate the optimum. Fig. 7 shows that $\gamma > 0$ improves simple regret, which is to be expected from its occasional greedy selection. However, as is shown in Fig. 8, a moderate fraction $\gamma \in \{0.05, 0.1\}$ also yields comparable or even improved inference regret on all tasks. Notably, $\gamma = 0.2$ yields slightly worse performance on Hartmann (6D), but yields marginally improved performance on Branin and Hartmann (3D). As such, the inverse γ-greedy approach not only improves performance in terms of simple regret, but yields improved inference as well.

C.2 Ablation study on σ_ϵ^2

For the noise variance ablation study, we once again consider the GP sample tasks. We fix the GP noise hyperparameter σ_ϵ^2 to the correct value prior to the start of the experiment. In Fig. 9, we show that the performance of `JES` is robust with respect to the noise level, while the performance of `MES` and `PES` decrease more drastically as the level of noise increases.

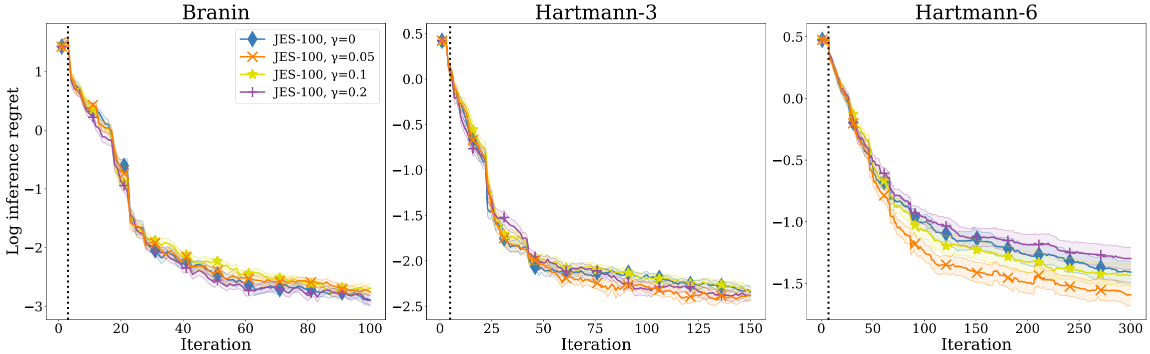

Figure 8: Comparison of JES with varying fraction γ of inverse greedy selections. Mean and 1 standard error of log inference regret is displayed for all tasks.

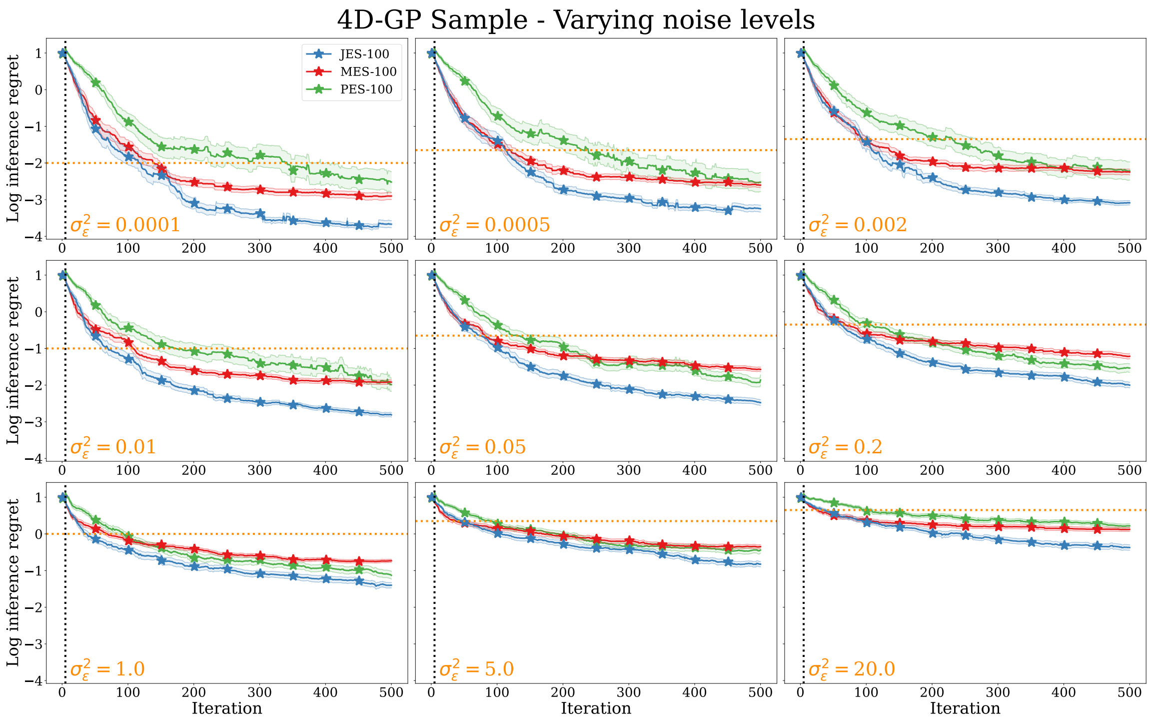

Figure 9: Evaluation of `JES`, `MES` and `PES` on noisy 4D GP sample tasks across 50 repetitions for 9 different noise levels. The noise variance σ_ϵ^2 ranges from 10^{-4} (top left) to 20 (bottom right). Log noise standard deviation $\log(\sigma_\epsilon)$ is marked in dashed orange.

Appendix D. Dependence on the number of MC samples

We show in Fig. 10 the dependence of `JES` on the number of MC samples for the GP sample tasks. `JES` displays a slight dependence on the number of GP samples, as initial performance improves marginally for larger number of samples. This is more prominent for higher-dimensional tasks, where a lower number of samples causes a substantially slower start. This can be explained by the fact that a larger number of sampled optimal pairs

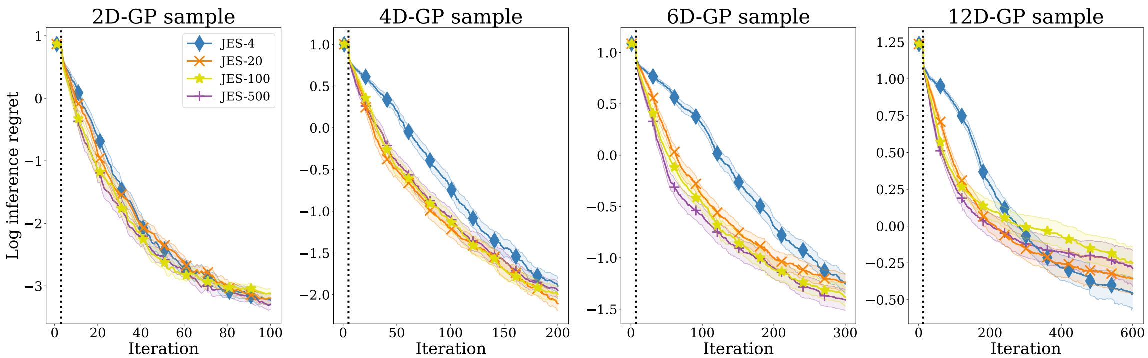

Figure 10: Comparison of JES with varying numbers of MC samples on GP tasks of varying dimensionalities - 10 repetitions on 10 unique tasks, for a total of 100 repetitions. Mean and 1 standard error of log inference regret is displayed for all tasks.

are required to accurately model a higher-dimensional density. Notably, the final regret on the 12-D benchmark is better for lower numbers of samples, such as `JES`-4. One potential explanation for this is its inability to model the joint distribution accurately. If all realized optimal pairs are close to the perceived optimum, `JES` is almost certainly going to sample there. Moreover, since the information gain for each sample is relatively low for samples in a well-explored region, the information gain of a few samples in an unexplored region may outweight the information gain of a much larger number of samples in a well-explored region. For `JES`-4 and `JES`-20, it is more likely that all of the sampled optimal pairs are close to the optimum in this manner, while there is still small positive density on other parts of the search space. As such, it is possible that `JES`-4 and `JES`-20 over-exploit slightly in the 12-dimensional benchmark.

Appendix E. Approximation Quality

We approximate the entropy of the posterior over observations conditioned on the data and the optimal pair $\mathrm{H}\left[p(y|\mathcal{D} \cup (\boldsymbol{x}^*, f^*), \boldsymbol{x}, f^*)\right]$ – the entropy of the sum of a Gaussian and a truncated Gaussian variable – by moment matching of the truncated distribution. We now show the quality of this approximation, and what impact it can potentially have on query selection. To do so, we utilize the results from Nguyen et al. (2022) regarding the density of $p(y|\mathcal{D}, \boldsymbol{x}, f^*)$. We note that the approximation regards the *truncation* from knowing the optimal value f^*, which constitutes an additional reduction in entropy after having conditioned the GP on the new observation. As such, the approximation considers only a fraction of the total entropy reduction, as visualized in Fig. **??**.

To establish the quality of the approximation, we compare our approach to approximating the entropy by MC. Naturally, the MC approach is more computationally expensive than moment matching, but yields an asymptotically correct result. In Fig. 11, we show for varying noise levels, expressed as the noise variance ratio of the total variance, and truncation quantiles $\Phi^{-1}(\alpha)$, the difference in entropy between an MC and a moment matching approach. For example, a truncation quantile of 10^{-2} means that the upper 99% of the density of the posterior distribution is removed as a result of truncation. We see that the approximate

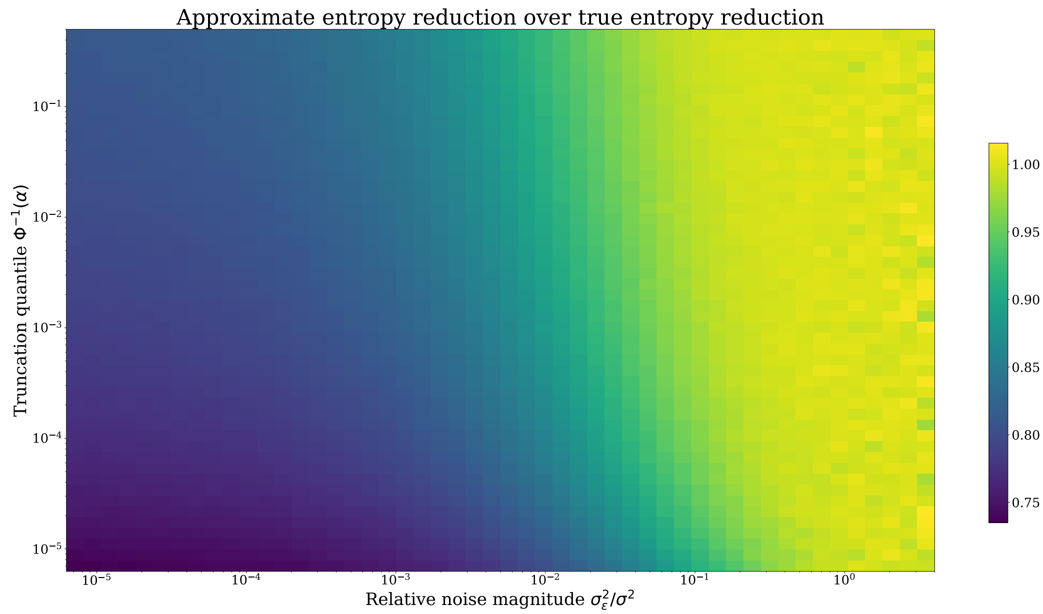

Figure 11: Visualization of approximation error from the moment matching approach compared to an asymptotically exact MC approach. The colormap represents the fraction of the entropy reduction resulting from truncation as approximated from moment matching divided by the entropy reduction as computed through MC. The inconsistencies in coloring in the rightmost part of the image are caused by inconsistent MC approximation.

entropy reduction from moment matching is consistently lower than for the asymptotically exact MC approach. Moreover, the approximation error seemingly increases logarithmically with the truncation quantile. As such, we can expect modest underestimates of the entropy reduction when we truncate the posterior to an extreme degree, and the level of noise is low. In the right image of Fig 2, the posterior is severely truncated left of the conditioned location. However, the noise variance constitutes a large fraction of the total variance at this location, which means that the approximation is still accurate with respect to the true entropy reduction. The scenario represented in the bottom left corner of the grid, where we severely truncate the posterior *and* the noise variance is low, does not reasonably occur in practice. The aforementioned scenario would entail sampling an optimal pair which is several orders of magnitudes *worse* than the mean of the GP at uncertain (in the sense that $\sigma >> \sigma_\epsilon$) locations. Lastly, the blue region in the upper left corner of the grid represents a region where we have less truncation, and the noise is a relatively insignificant part of the total variance. Fig. 11 shows that the entropy reduction from truncation in this region is underestimated by approximately 15%. As such, the approximation error leads to a slightly less explorative strategy than what a strategy with exact computation of the truncation term would provide.

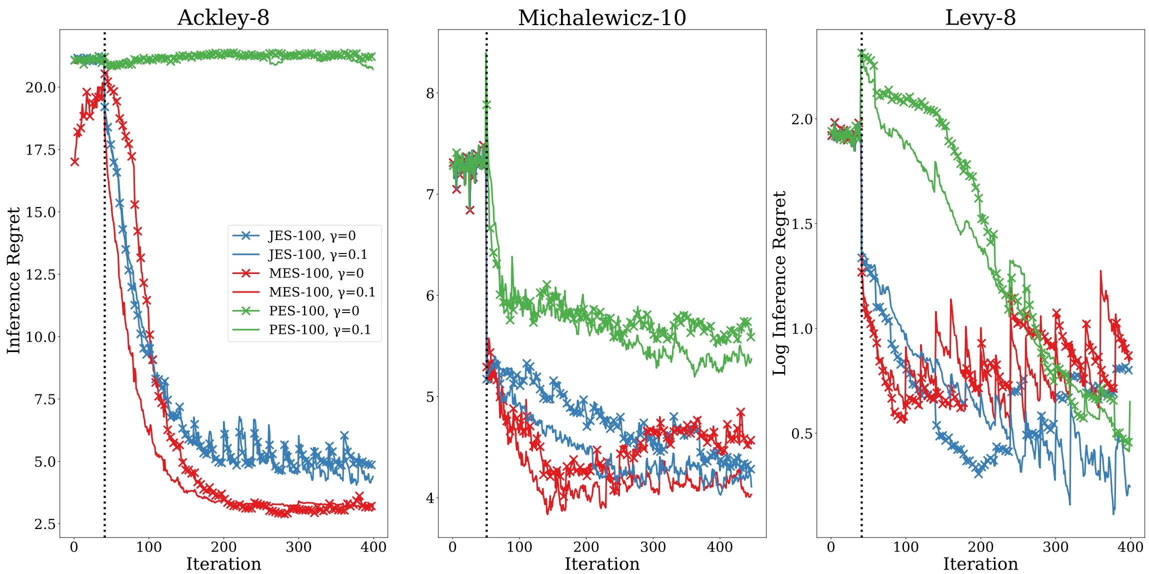

Figure 12: Mean of inference regret on high-dimensional synthetic functions for vanilla inverse γ-greedy versions of `JES`, `MES` and `PES`. Upward spikes signify points where the GP hyperparameters are re-sampled, and the inference regret getting worse as a result. Error bars are omitted to increase legibility.

Appendix F. Model Misspecification

The performance of information-theoretic methods can suffer substantially from model misspecification. In Fig. 12, we show for `JES`, `PES` and `MES` how the inverse γ-greedy approach helps stabilize inference and improve inference regret, and yields substantially better simple regret for all methods. For Michalewicz (10D), we observe that the inference regret of `MES` gets substantially worse after iteration 150. With the inverse γ-greedy approach, this issue is severely reduced. In Fig. 13, the corresponding simple regrets for `MES` deviate at iteration 150. The same behavior can be observed for `JES` on Levy (8D) around iteration 200. Across all test functions in Fig. 12 and Fig. 13, an inverse γ-greedy strategy yields comparable or improved inference regret, and strictly improved simple regret for all acquisition functions. Moreover, we observe that `MES` is generally the top-performing acquisition function, implying that it is the most robust to model misspecification. Using the inverse γ-greedy approach, the acquisition function verifies whether the belief over the location of the optimum is correct under the current model hyperparameters. If it is not, it re-calibrates its belief.

Appendix G. Regret Measures

We display the regret measures for the GP sample and synthetic test functions.

G.1 GP sample tasks

In Fig. 14, we show the simple regret for the GP sample tasks. We note that the simple and inference regrets for `MES` are approximately equal, while there is a substantial difference for

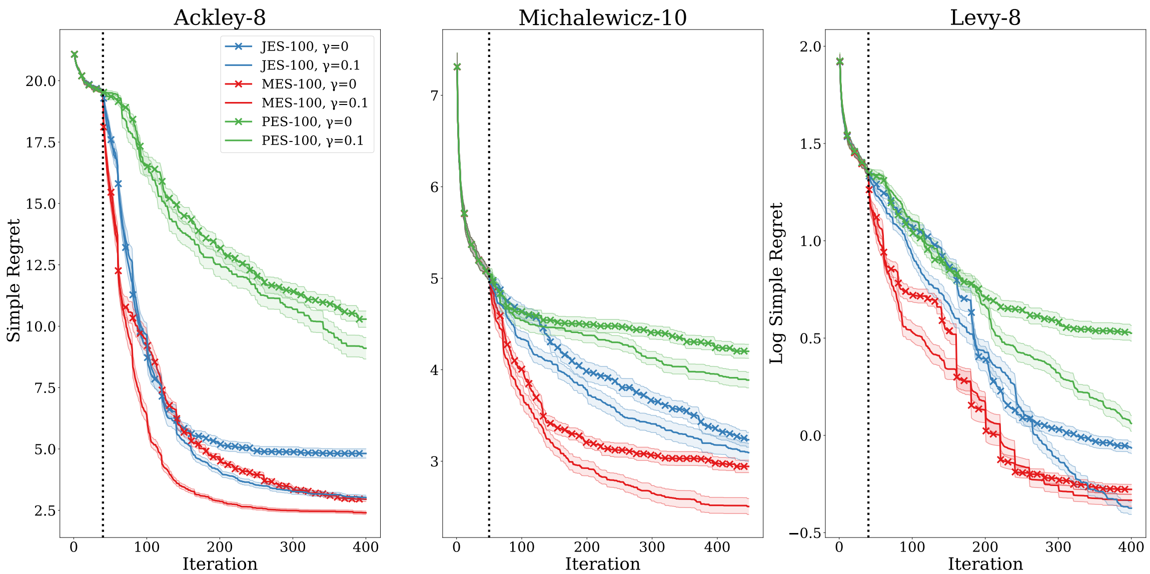

Figure 13: Mean and 1 standard error of simple regret on high-dimensional synthetic functions for vanilla inverse γ-greedy versions of JES, MES and PES.

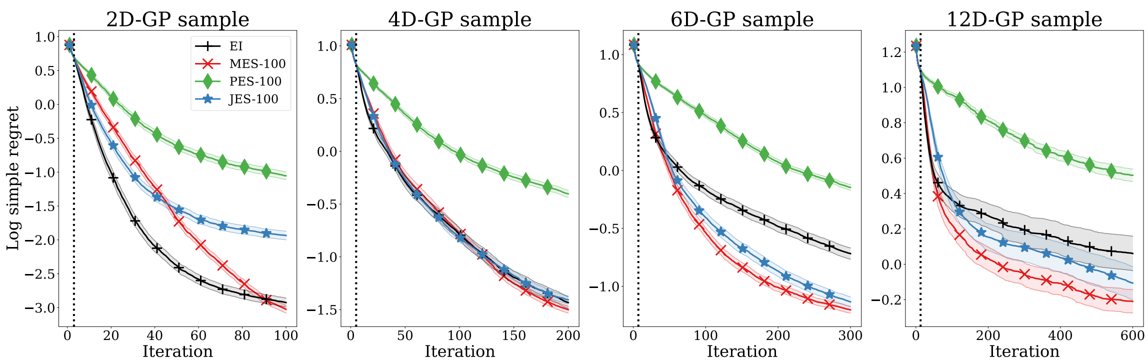

Figure 14: Comparison of JES, MES, PES and EI on GP prior samples using simple regret. We run 1000 repetitions each for 2, 4 and 6D, and 250 on 12D. Mean and 2 standard errors of log regret are displayed. The vertical dashed line shows the end of the initial design phase.

JES. For PES, the difference between simple and inference regret is the most pronounced at approximately two order of magnitude at the most.

G.2 Synthetic test functions

Next, we show the inference regret for the synthetic test functions in Fig. 15. We note that he simple and inference regrets for JES are approximately equal. For PES, the difference between simple and inference regret is once again very pronounced. Notably, the simple regret is significantly better than the inference regret for MES on Branin, implying that it does not yet have full knowledge of where the optimum is. We once again note the numerical issues of PES on Branin. Moreover, the inference regret of MES gets marginally worse for

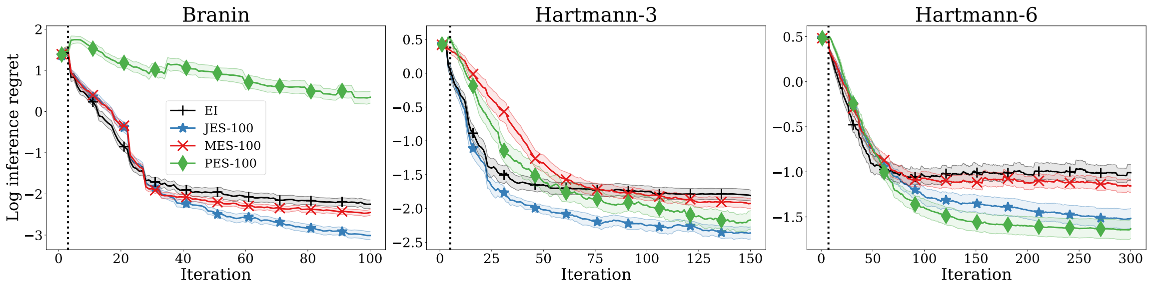

Figure 15: Comparison of `JES`, `MES`, `PES` and `EI` on Branin and Hartmann-6, $\sigma_\epsilon^2 = 0.10$. Mean and 2 standard errors of log regret are displayed across 100 repetitions. The vertical dashed line represents the end of the initial design phase.

Hartmann (6D) from approximately iteration 100 until the end of the run, implying that its knowledge about the location of the optimum gets worse.