Inducing Point Allocation for Sparse Gaussian Processes in High-Throughput Bayesian Optimisation

Henry B. Moss
Secondmind.ai

Sebastian W. Ober
Secondmind.ai

Victor Picheny
Secondmind.ai

Abstract

Sparse Gaussian processes are a key component of high-throughput Bayesian optimisation (BO) loops; however, we show that existing methods for allocating their inducing points severely hamper optimisation performance. By exploiting the quality-diversity decomposition of determinantal point processes, we propose the first inducing point allocation strategy designed specifically for use in BO. Unlike existing methods which seek only to reduce global uncertainty in the objective function, our approach provides the local high-fidelity modelling of promising regions required for precise optimisation. More generally, we demonstrate that our proposed framework provides a flexible way to allocate modelling capacity in sparse models and so is suitable for a broad range of downstream sequential decision making tasks.

1 Introduction

Countless design tasks in science, industry and machine learning can be formulated as high-throughput optimisation problems, as characterised by access to substantial evaluation budgets and an ability to make large batches of evaluations in parallel. Prominent examples include high-throughput screening within drug discovery (Hernández-Lobato et al., 2017), DNA sequencing, and experimental design pipelines, where automation allows researchers to efficiently oversee thousands of scientific experiments, field tests and simulations through sensor arrays and cloud compute resources (Kandasamy et al., 2018). However, such design tasks tend to have large search spaces and multi-modal optimisation landscapes such that, even under large optimisation budgets, only a small proportion of candidate solutions can ever be evaluated, and often only with significant levels of observation noise. Consequently, most existing optimisation routines are unsuitable, as brute-force methods require too many evaluations.

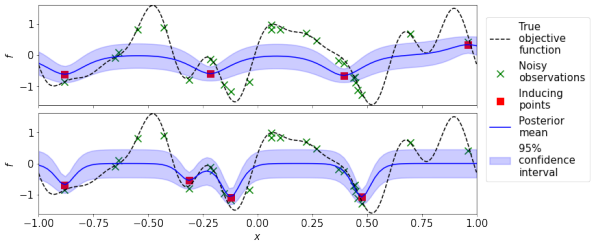


Figure 1: A toy problem showing two sparse GP surrogate models, one with its inducing points chosen using an existing method (top) and the other using one of our proposed BO-specific methods (bottom) that focuses modelling resources into promising areas of the search space. Our model is better suited for assisting BO find this function's minima.

Bayesian optimisation (BO, see Shahriari et al., 2016, for a review) has surfaced as the *de facto* approach for solving noisy black-box optimisation tasks under restricted evaluation budgets, with numerous successful applications across the empirical sciences and industry. However, vanilla BO relies on Gaussian processes (GPs, Rasmussen and Williams, 2006), which incur a significant computational overhead for each individual optimisation step. This cost becomes increasingly unwieldy as data volumes increase, making it unsuitable for the high-throughput tasks motivated above.

Several ways to scale up BO with large data volumes have been explored, including using local models (Eriksson et al., 2019) or neural networks (Hernández-Lobato et al., 2017). Among those alternatives, using sparse GPs (Titsias, 2009) are particularly attractive as they dramatically reduce the computational cost of GPs and have enabled BO to be applied to a range of applications including molecular search (Griffiths and Hernández-Lobato, 2020), laser optimisation (McIntire et al., 2016), model optimisation (Nickson et al., 2014), alloy design (Yang et al., 2021), and risk-adverse optimisation (Picheny et al., 2022).

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

In a nutshell, sparse GPs replace the full set of observations by a smaller representative set of pseudo-observations referred to as *inducing points*. The choice of the inducing point locations has a critical influence on the behaviour of the model, as it encodes local expressivity. However, existing approaches for inducing point allocation (IPA) focus purely on regression tasks, i.e., the global accuracy of models, and so sacrifice high-fidelity (local) modelling of promising regions which is required, as confirmed by our experiments, for effective optimisation (Figure 1). For this reason, there is a need for BO-specific IPA strategies; however, to our knowledge, no such methods exist in the literature.

Our contributions can be summarised as follows:

1. We demonstrate that existing IPA strategies do not support high-precision BO.
2. We introduce the use of quality-diversity decomposed DPPs as an IPA, allowing the trade-off of an IPA's diversity against an underlying preference.
3. We propose a guide for practical BO-specific IPA methods along with several specific recommendations.
4. We show that our methods out-perform established baselines across synthetic and real-world high-throughput optimisation and active learning tasks.

2 Background

Bayesian Optimisation. BO is a highly data-efficient method for finding the optima of a smooth function $f : \mathcal{X} \to \mathbb{R}$. By using a probabilistic surrogate model, typically a GP, coupled with a data acquisition strategy, evaluations are focused into promising areas of the search space $\mathcal{X}$, allowing identification of good solutions within heavily constrained evaluation budgets.

Popular examples of data acquisition strategies include expected improvement (EI, Jones et al., 1998), knowledge gradient (Frazier et al., 2008), entropy search (Hennig and Schuler, 2012), or Thompson sampling (TS, Kandasamy et al., 2018). While our framework is not specific to any acquisition strategy, we focus mainly on Thompson sampling, a simple yet effective strategy that evaluates the maxima (minima) of random samples from the surrogate model when performing black-box maximisation (minimisation). TS is an obvious choice for high-throughput BO due to its natural ability to handle highly parallelised optimisation resources, e.g. for molecular search (Hernández-Lobato et al., 2017) or distributed computing (Kandasamy et al., 2018). Moreover, Vakili et al. (2021) have recently shown that the decoupled TS approach of Wilson et al. (2020) can provide a drastic efficiency gain over traditional TS without significant impact on regret performance.

Gaussian Processes. GP models are a popular choice as surrogate models for BO, as they combine flexibility with reliable uncertainty estimates. A GP can be defined as an infinite collection of random variables, any finite number of which are distributed according to a multivariate Gaussian (Rasmussen and Williams, 2006). Consider a dataset $\mathcal{D} = (X, \mathbf{y})$ consisting of N input-output pairs $(\mathbf{x}_n, y_n)$, where $\mathbf{x} \in \mathcal{X}$ and $y \in \mathbb{R}$. In Gaussian process regression, we model this dataset as a noisy realization of a latent function,

$$y_n = f(\mathbf{x}_n) + \epsilon_n, \quad \epsilon \sim \mathcal{N}(0, \sigma^2),$$

where we have given f a GP prior, $f \sim \mathcal{GP}(\mu_0(\cdot), k(\cdot,\cdot))$, and σ^2 is the noise variance. $\mu_0 : \mathcal{X} \to \mathbb{R}$ is the (prior) mean function, whereas $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive semidefinite covariance function or kernel; taken together, these are sufficient to fully describe the GP prior, which states that $f(X) \sim \mathcal{N}(\mu_0(X), K_X)$, where we have defined $K_X \coloneqq [k(\mathbf{x}, \mathbf{x}')]_{\mathbf{x},\mathbf{x}' \in X}$ (abusing the notation slightly). For notational simplicity, we henceforth assume the mean function to be zero. By conditioning on the observed data, we can compute the exact posterior $p(f|\mathbf{y})$ as a GP with mean and covariance functions

$$\mu(\mathbf{x}) = \mathbf{k}_X(\mathbf{x})^T (K_X + \sigma^2 I_N)^{-1} \mathbf{y} \tag{1}$$
$$\Sigma(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \mathbf{k}_X(\mathbf{x})^T (K_X + \sigma^2 I_N)^{-1} \mathbf{k}_X(\mathbf{x}'),$$

where we have defined $\mathbf{k}_X \coloneqq [k(\mathbf{x}', \mathbf{x})]_{\mathbf{x}' \in X}$ and the identity matrix $I_N \in \mathbb{R}^{N \times N}$. While we can compute the exact posterior predictive using these equations, in practice we are often limited to using small datasets, as computing the required $(K_X + \sigma^2 I_N)^{-1}$ requires $O(N^3)$ computational complexity and $O(N^2)$ memory.

Sparse Variational Gaussian Processes. To mitigate the computational cost of GP modelling and allow for larger datasets, sparse variational approaches (Titsias, 2009; Hensman et al., 2013) have been developed. Instead of conditioning on the N training points, sparse GPs learn a set of $M << N$ *inducing variables* $\mathbf{u} \in \mathbb{R}^M$, defined at *inducing locations* $Z = \{\mathbf{z}_m\}_{m=1}^M, \mathbf{z}_m \in \mathcal{X}$, so that $\mathbf{u} = f(Z)$. By defining an approximate posterior over the inducing variables $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}; \mathbf{m}, S)$ with variational parameters $\mathbf{m} \in \mathbb{R}^M, S \in \mathbb{R}^{M \times M}$, we can simultaneously learn the inducing locations and variational parameters by maximizing the *evidence lower bound (ELBO)*:

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{f})} \left[\log p(\mathbf{y}|\mathbf{f})\right] - \mathrm{KL}(q(\mathbf{u})||p(\mathbf{u})),$$

where $q(\mathbf{f}) = \mathcal{N}(\mathbf{f}; \mu_\mathbf{f}, \Sigma_\mathbf{f})$ is the approximate posterior over the function values defined at the data points implied by conditioning on $\mathbf{u}$,

$$\mu_\mathbf{f} = \mathbf{k}_Z(X)^T K_Z^{-1} \mathbf{m}$$
$$\Sigma_\mathbf{f} = K_X + \mathbf{k}_Z(X)^T K_Z^{-1} (S - K_Z) K_Z^{-1} \mathbf{k}_Z(X),$$

where we have defined $\mathbf{k}_Z(\cdot)$ and K_Z analogously to $\mathbf{k}_X(\cdot)$ and K_X, respectively. We refer to this model as the *sparse*

variational Gaussian process (SVGP). The SVGP model requires $O(M^2\tilde{N})$ computational complexity and $O(M\tilde{N})$ memory, where $\tilde{N}$ is the size of a minibatch, a significant saving over the exact GP.

While the inducing locations can be learned according to the ELBO along with the model hyperparameters and variational parameters, Burt et al. (2020) argues that this yields a very challenging high-dimensional and non-convex optimization task with a complicated dependence structure that is difficult to solve, converges slowly and often provides sub-optimal models. Whereas for regression this may be allowable with sufficient computational resources, for BO we must be able to reliably and quickly fit models, and therefore it would be preferable to allocate Z *a priori* and keep them fixed. Moreover, optimizing Z according to the ELBO encourages the inducing points to approximate the posterior globally (Matthews et al., 2016), which we will argue is wasteful for BO applications. Therefore, we focus the remainder of our work on methods for inducing point allocation (IPA), which we will use to set the inducing points at the start of each BO step. We start by describing prior work for IPA, which focuses on regression, before moving to our BO-specific IPA contributions.

3 Inducing Point Allocation for Regression

Existing IPA strategies include taking a random subset of the data, sampling uniformly across the problem's search space, or using centroids obtained by running a K-means algorithm on the data (Hensman et al., 2013). The remainder of this Section details the recent DPP-based method of Burt et al. (2019), laying out important groundwork for our proposed BO-specific IPA strategies.

Determinantal Point Processes. For regression tasks, a meaningful criterion for IPA would be to have the points spread as uniformly as possible across the input data X. It would also be sensible to have a criterion that takes the kernel and its hyperparameters into account. Burt et al. (2020) showed that one way of achieving these is by using an M-determinantal point process (M-DPP, Kulesza and Taskar, 2012). An M-DPP chooses the M points in Z by sampling them from the data X with probability proportional to the determinant of the Gram matrix K_Z:

$$\mathbb{P}(\mathcal{Z} = Z) \propto \left|K_Z\right|. \tag{2}$$

Notice that this criterion meets our two criteria described above: 1) if two points are close together in Z, the determinant will typically be small since the kernel will have high covariance for those points, giving the M-DPP repulsive properties so that the selected points have a uniform spread, and 2) the determinant clearly depends on the kernel. Using results from the M-DPP literature, Burt et al. (2020) was able to show that sampling inducing points in this way from a DPP will lead to a small expected KL divergence between approximate and true posteriors, $KL[q(f)||p(f|\mathbf{y})]$. Moreover, these results have recently been used to prove regret bounds in BO for sparse GP methods (Vakili et al., 2021).

Conditional Variance Reduction. In practice, sampling from a DPP is computationally expensive. Therefore, Burt et al. (2020) suggests finding the *maximum a posteriori (MAP)* estimate of a DPP, i.e., finding the set of inducing points Z with maximum probability according to Eq. 2. While exact MAP estimation of a DPP is known to be NP-hard (Ko et al., 1995), Chen et al. (2018) provides an algorithm for approximate MAP estimation in $O(M^2N)$, which Burt et al. (2020) uses in practice. This algorithm greedily builds its set of points Z by choosing the j^{th} point from $X \setminus Z_{1:j-1}$ as

$$\mathbf{z}_j = \underset{\mathbf{z}\in X\setminus Z_{1:j-1}}{\operatorname{argmax}} \left|K_{Z_{1:j-1}\bigcup\{\mathbf{z}\}}\right|. \tag{3}$$

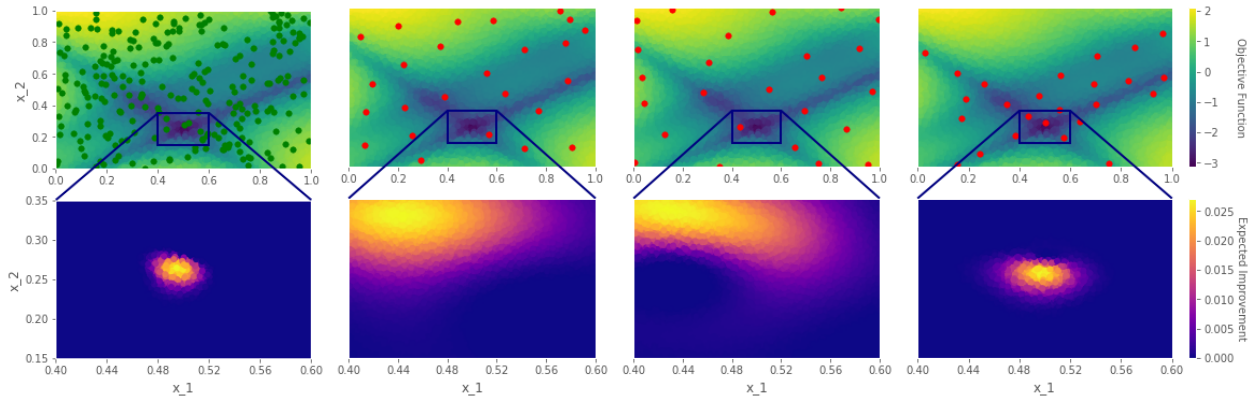
Interestingly, this DPP-based IPA strategy (3) is equivalent to greedily building a set of inducing points by maximising the posterior predictive variance of a noise-free GP model $f \sim \mathcal{GP}(0, k)$ conditioned on previously selected observations, i.e., choosing

$$\mathbf{z}_j = \underset{\mathbf{z}\in X}{\operatorname{argmax}}\, \sigma_{j-1}(\mathbf{z}), \tag{4}$$

where $\sigma^2_{j-1}(\mathbf{z}) = k(\mathbf{z},\mathbf{z}) - \mathbf{k}_{Z_{1:j-1}}(\mathbf{z})^T K^{-1}_{Z_{1:j-1}} \mathbf{k}_{Z_{1:j-1}}(\mathbf{z})$ is the *conditional variance* of the GP (see Hennig and Garnett, 2016; Burt et al., 2019, for detailed description and discussion). Therefore, we refer to this method of selection as conditional variance reduction (CVR), as it selects the datapoint with the highest conditional variance as the next inducing point, in hopes that this variance will be reduced.

4 Inducing Point Allocation for Bayesian Optimisation

BO typically requires updating the surrogate model(s) at each step to leverage the latest information available, so it makes sense to include updating the inducing point locations (see Algorithm 1). In the case of CVR, which requires a kernel, Vakili et al. (2021) use the kernel fitted during the previous BO step. Unfortunately, as we will demonstrate across all our experimental results, regression-inspired IPA strategies are not satisfactory for use within BO loops. While a level of global accuracy is needed to prevent the re-investigation of areas already identified as sub-optimal, Figure 2 shows that accurate modelling in promising areas is necessary to allow the precise identification of the optimum. For a more formal intuition into the unsuitability of existing IPA strategies see Appendix A. For these reasons we now propose DPP-based IPA strategies that are able to change the relative trade-off of local and global modelling capabilities.



(a) Exact GP (b) SVGP with K-means (c) SVGP with CVR (d) SVGP with IMP-DPP

Figure 2: The top row shows (a) 250 available training data points (green) alongside (b,c,d) three different 25 point IPAs (red) chosen for a function minimisation task. Existing approaches which (b) use the centroids from a k-means clustering of the available data or (c) use the CVR strategy provide balanced coverage of the whole search space. In contrast, our IMP-DPP strategy (d) focuses modelling resources into promising central areas. The bottom rows show expected improvement acquisition functions evaluated in the promising region according to (a) an exact GP trained on all available data and those (b,c,d) arising from SVGPs with the IPAs above. Of the SVGPs, only our proposed IMP-DPP's acquisition function agrees with the exact GP.

Algorithm 1: High-throughput BO with SVGPs

Input: Resource Budget R, Batch size B
Initialise $n \leftarrow 0$ and spent resource counter $r \leftarrow 0$
Collect initial design D_0 and fit initial model $\mathcal{M}_0$
while $r \leq R$ **do**
  Begin new iteration $n \leftarrow n + 1$
  Build IPA Z_n using D_{n-1} and $\mathcal{M}_{n-1}$
  Fit model $\mathcal{M}_n$ using IPA Z_n to data D_{n-1}
  Generate B query points $\{\mathbf{x}_i\}_{i=1}^B$
  Collect evaluations $D_n \leftarrow D_{n-1} \bigcup \{(\mathbf{x}_i, y_{\mathbf{x}_i})\}_{i=1}^B$
  Update spent budget $r \leftarrow r + B$
return *Believed optimum across* $\{\boldsymbol{x}_1, ..\boldsymbol{x}_R\}$

We now provide the primary contribution of this work — a general method for IPA suitable for down-stream decision making tasks. Unlike existing IPA strategies, our proposed methods ensure the model focuses its resources on promising (local) areas of the space whilst maintaining a sufficiently accurate global model.

4.1 A General IPA Formulation

Quality-Diversity Decomposition. Although CVR only leverages the repulsive properties of DPPs, it is also possible, through a convenient reparameterisation, to encode a notion of the quality of the sampled points. Consider the DPP defined as in (2) but with K_Z replaced by

$$L_Z = [q(\mathbf{z}_i)k(\mathbf{z}_i, \mathbf{z}_j)q(\mathbf{z}_j)]_{(\mathbf{z}_i,\mathbf{z}_j)\in Z\times Z}\,, \tag{5}$$

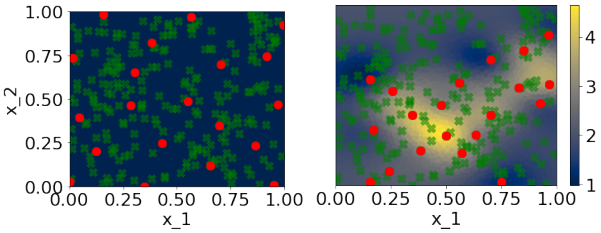


Figure 3: 25 elements (red) chosen from 250 candidates (green) by a DPP with (left) constant and (right) locally varying quality functions (background colour).

where $q : \mathcal{X} \to \mathbb{R}$, i.e., we observe Z with probability $\mathbb{P}(\mathcal{Z} = Z) \propto |L_Z|$. In our case, we can choose $q : \mathcal{X} \to \mathbb{R}^+$ so that it can be seen as a *quality function*, designed to provide large values for points lying in promising areas of the space and low values elsewhere. Indeed, due to the decomposition

$$|L_Z| = |K_Z| * \prod_{i=1}^{N} q(\mathbf{z}_i)^2, \tag{6}$$

as derived in Section 3.1 of Kulesza and Taskar (2012), it is clear that a particular Z will occur with high probability only if it contains points that have large quality scores (as measured by $q(\mathbf{z}_i)$) **and** have a diverse spread (as measured by $|K_Z|$), see Figure 3. Hence, this constitutes an intuitive tool for building IPAs well-suited to the demands of BO (see Figure 2d for a demonstration, and Appendix A for a more formal justification).

Greedy (Approximate) Maximisation. Given a particular

q, we can simply apply the same greedy algorithm used by CVR, just with K_Z replaced by L_Z, to efficiently build a set of BO-specific inducing points as an $O(NM^2)$ approximate MAP estimate of the DPP implied by Eq. 5. Conveniently, this resulting MAP estimate has an intuitive interpretation, as specified in Theorem 1, with proof in Appendix B.

Theorem 1. *Suppose inducing points $\mathcal{Z}$ are distributed according to a DPP with similarity kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and quality function $q : \mathcal{X} \to \mathbb{R}$, i.e., $\mathbb{P}(\mathcal{Z} = Z) \propto \left|L_Z\right|$. Then, according to the greedy approximation, the j^{th} component of the MAP estimate of $\mathcal{Z}$ is given by*

$$\mathbf{z}_j = \underset{z \in X}{\operatorname{argmax}}\ q(\mathbf{z})\sigma_{j-1}(\mathbf{z}), \tag{7}$$

where $\sigma^2_{j-1}(\mathbf{z})$ is the conditional variance of the noise-free GP model conditioned on the already selected points $Z_{1:j-1}$ (cf. Eq. 4).

4.2 Choosing a Quality Function

While any quality function can be used, in practice q should be carefully chosen to deliver the right quality-diversity trade-off. Intuitively from Eq. 5, the relative amplitudes of the quality function and the similarity kernel are key to this trade off, so for effective IPA, q must be chosen to complement k (rather than dominate or be dominated by it).

In addition, we propose the following four properties to guide our choice of the quality functions:

- **Discriminative:** q should return large values in areas of the space that are worthwhile modelling whilst providing smaller contrasting values elsewhere. This means high values in regions expected to be close to the optimum and/or with large predictive uncertainty.
- **Informative:** $q(\mathbf{z})$ should encode our current knowledge about the objective function f at $\mathbf{z}$, which is available through the already-collected evaluations $y_i = f(\mathbf{x}_i)$ and/or the surrogate model(s) of the previous BO step.
- **Shift invariance:** the resulting IPA should be invariant to adding an offset to the data. Given a GP model, adding an offset should only affect its mean function and leave the kernel unchanged, which means that the quality function must also be insensitive to shift.
- **Scale invariance:** the resulting IPA should be invariant to linear re-scaling of the data. Given a GP model, a multiplicative factor on the data may result in a multiplicative factor on the kernel. Hence, for eq. 7 to deliver identical results, we need q to be invariant to re-scaling, up to a multiplicative (positive) constant.

4.3 A Linear Quality Measure

We propose here a simple and intuitive choice for the quality function that shows strong empirical performance (see Section 6). Many other choices are possible, for example, we also derived a quality function based on information-theoretic considerations. Although well-motivated, we found this entropy-based approach to be less effective, likely due to the computational approximations required, than the simpler choice that we are about to present. To streamline our exposition, the derivation and results of the information-theoretic approach are deferred to Appendix C.

Noise-free evaluations. A natural quality function (for a single-objective maximisation problem) that satisfies the four above-mentioned properties is the following linear function of y_i:

$$q_{\text{Lin}}(\mathbf{z}_i) = y_i - \hat{f}, \qquad \text{with } \hat{f} = \min_i y_i. \tag{8}$$

A linear rescaling of the data will change q by a multiplicative factor only, and subtracting $\hat{f}$ makes it positive and shift invariant. Furthermore, $\hat{f}$ ensures the discriminative property, i.e. that q is zero at the worst observation and largest at the best.

Noisy evaluations. For problems with large observation noise, y_i can give misleading estimates of $f(\mathbf{z}_i)$ and so it is unwise to use (8). However, in these settings, we can make use of the previous BO step's surrogate model $\mathcal{M}_{n-1}$ and instead calculate the expected value of q_{Lin}. Additionally, to ensure positivity, we swap the linear function for the (piece-wise linear) Recitified Linear Unit (ReLU), yielding the quality function

$$q_{\text{IMP}}(\mathbf{z}_i) = \mathbb{E}_{f \sim \mathcal{M}_{n-1}}\left[\max(f(\mathbf{z}_i) - \hat{f}, 0)\right], \tag{9}$$

where the baseline is now the minimal predicted value of the objective function, i.e. $\hat{f} = \min_{\mathbf{x} \in D_n} \mu_{n-1}(\mathbf{x})$ for $\mu_{n-1}(\cdot)$ the posterior mean of $\mathcal{M}_{n-1}$. Note that (9) takes the form of the well-known Expected Improvement (EI), just with a modified baseline, and so can be calculated in closed form (see Jones et al., 1998).

Reassuringly, the performance of (9) is robust to the specific choice of baseline, with Appendix D showing negligible performance differences when using the minima or mean of the predicted objective function values, or even when using a softplus relaxation of the ReLU. However, significantly tightening the baseline to be the maximum of the objective function (as typically used by EI acquisition function) yields a dramatic drop in performance. Indeed, EI is not designed to discern between all our collected points, only to help identify where there could be new maxima.

4.4 Beyond Single Objective BO

The quality function described above is designed for single-objective BO problems. However, our approach based on

the quality-diversity decomposition of DPPs is a general way to ensure that sparse models are accurate in the areas where they will be used and so may apply to a much larger variety of optimisation problems, including those with constraints, multiple objectives, and more generally active learning problems such as level set estimation. The quality function should be tailored to each problem: for instance in level-set estimation, the important regions to model are not regions where the output value is maximal, but where it is close to the targeted level. In Section 6 we demonstrate such extensions.

5 Related Work

Alternative Sparse Surrogate Models. Three other formulations of sparse GPs have been used in BO loops. Firstly, McIntire et al. (2016) propose a compelling modification of sparse online GPs(Csató and Opper, 2002), where they upweight promising areas of the feature space (as measured by the expected improvement of candidate evaluations). However, online GPs, which see only a single pass of the data, provide worse approximations than SVGPs, which have multiple chances to learn from each datapoint. Moreover, due to its requirement of N individual challenging optimisations for each individual model fit, this approach is unsuitable for the high-throughput scenarios tackled in this paper (and consequently was only tested by McIntire et al. (2016) on problems with $M = 30$ and $N = 60$). Another way to alleviate the cost of GP inference is by approximating the spectral density of its kernel (Lázaro-Gredilla et al., 2010). However, spectral approximations are not appropriate for BO as they seek to preserve global structure and, as such, have no way of providing local high-fidelity modelling. Indeed, applying spectral GPs to BO requires expensive and heuristic modifications to its loss function (Yang et al., 2021) and even then fails to match the performance of exact GPs. In contrast to these two alternatives, our proposed approach retains the state-of-the-art computational complexity and performance of SVGPs. Finally, Maddox et al. (2021) (with similar work by Chang et al. (2022)) propose the OVC method for the fast conditioning of SVGPS, allowing efficient calculation of popular look-ahead acquisition functions, albeit those outside of the high-throughput domain.

Additional uses of DPPs in BO. Outside of IPA, DPPs are also commonly used in the context of batch BO, where the goal is to recommend diverse collections of points. Prominent examples include the approaches of Kathuria et al. (2016); Dodge et al. (2017) and Nava et al. (2022), as well as Moss et al. (2021) where, similarly to our ENT-DPP, an information-theoretic motivation is used to inform the construction of the DPP's diversity and quality terms. DPPs have also been used in high-dimensional BO (Wang et al., 2017) to sample diverse subsets of the available search space dimensions.

Scalable BO via Local Models. A popular alternative approach for BO under large evaluation budgets is to use multiple cheaper local models in lieu of a single expensive global model (Gramacy and Apley, 2015; Rullière et al., 2018; Cole et al., 2021, 2022). Particularly powerful BO routines employing local models like TURBO of (Eriksson et al., 2019) and, for multi-objective BO, MORBO (Daulton et al., 2022) are ideal for applications where the only goal is to find a reasonable solution because global modelling (and optimisation) is challenging (e.g., for high-dimensional optimisation problems). However, the local models built by TURBO are not always useful in settings where the goal is to collect data that allows the building of a useful final model, e.g., in the active learning applications we consider below or when we need a rough understanding of global behaviour to ensure global convergence.

6 Experimental Results

We now provide an empirical evaluation of our proposed IPA framework across a suite of high-throughput BO problems using the open-source BO library TRIESTE (Berkeley et al., 2022). We then illustrate the general applicability of our IPA framework, by demonstrating how quality functions can be designed for multi-objective and active learning problems. Additional experimental details are contained in our appendices. Implementations of our IPAs are contained within the TRIESTE (Berkeley et al., 2022) and BOTORCH (Balandat et al., 2020) libraries.

6.1 Single Objective Optimisation

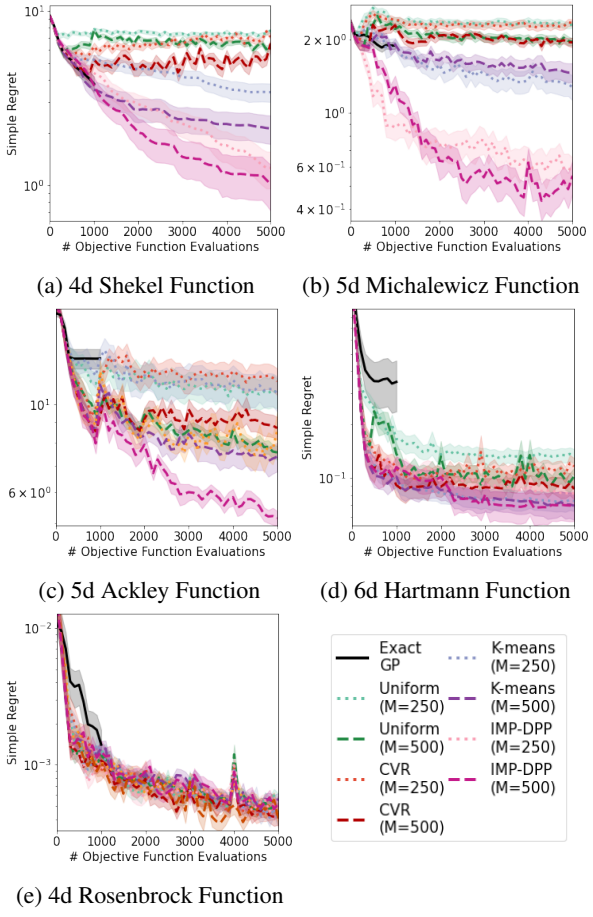
For clarity, all our synthetic experiments follow the same setup. We consider an SVGP model with either $M = 250$ or 500 inducing points using either 1) our proposed IPA strategy with the improvement-based quality function (9) which we call IMP-DPP , 2) the CVR of Burt et al. (2019) (see Section 3), 3) choosing the centroids of a K-means clustering of the data, and 4) choosing inducing points spread uniformly across the search space. SVGP models are fit using an Adam optimiser with learning rate 0.1, using an early stopping criteria with a patience of 50 and a learning rate halving on plateau schedule with a patience of 10.

For all DPP-based IPAs we follow the thoroughly tested approach of Burt et al. (2020) and Vakili et al. (2021) and use the kernel of the previous BO step's model to allocate the IPA and then refit the kernel when training the current BO step's model on the new data and the chosen IPA. We allocate a total evaluation budget of $N = 5{,}000$ evaluations split across 50 BO steps in batches of 100 points. When the total number of queried points is less than the desired number of inducing points (e.g. for the first 4 optimisation steps when $M = 500$), we use just the N available training points as our IPA. Subsequent batches are collected using the decoupled Thompson sampling scheme presented in

Vakili et al. (2021). We use 100 random Fourier features to build a Fourier representation of samples and maximise each using an L-BFGS optimiser starting from the best of a random sample of 10,000 points. BO using an exact GP model is included as a baseline; however, we can report only the first 10 optimisation steps, after which it became prohibitively expensive (i.e., for $N > 1{,}000$).

Figure 4 demonstrates optimisation performance across the 4d Shekel, 5d Michalewicz, 5d Ackley, 6d Hartmann, and 4d Rosenbrock functions (see Appendix E for definitions), where we have contaminated the evaluations of each with Gaussian noise of variance 0.01, except for the easier Hartmann where we consider a larger variance of 0.1. Note that we re-scaled these baselines so that they have a variance of 1.0 (under random samples across the search space) and so these noise levels are large, resulting in challenging optimisation tasks. Unsurprisingly, greater performance is achieved when using larger number of inducing points for all the considered methods, except for the easier Rosenbrock function, where all methods perform equally well. For the Shekel and Michalewicz functions, only IMP-DPP achieves precise optimisation, even when using just $M = 250$ inducing points. For the Michalewicz function, IMP-DPP with $M = 500$ provides a dramatic improvement over the other methods. In contrast, on the Hartmann function we see all $M = 500$ approaches, as well as IMP-DPP with $M = 250$, achieve comparable optimisation performance. In addition to performing improved optimisation, we show (in Appendix F) that our SVGP-based approaches incur significantly lower computational overheads than exact GPs. Moreover, unlike the exact GP, the SVGP approaches maintain a constant overhead as BO progresses.

Interestingly, for some of the more challenging functions considered in Figure 4, the exact GP leads to optimisation that gets stuck in local minima, whereas the SVGP approaches are able to fully converge. We hypothesise that SVGPs have an advantage in these non-stationary settings as they are able to ignore promising yet not optimal areas of the space that would otherwise mislead the algorithm — a helpful consequence of their limited modelling resources. Similar behaviour is noted by Maddox et al. (2021) when also using SVGPs for online decision making.



(a) 4d Shekel Function (b) 5d Michalewicz Function

(c) 5d Ackley Function (d) 6d Hartmann Function

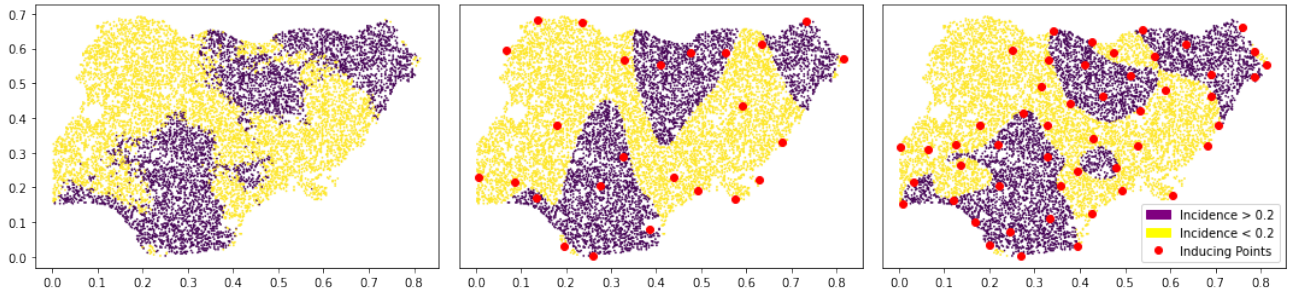
(e) 4d Rosenbrock Function

Figure 4: Results are averaged over 50 runs and we report the mean and its 95% confidence intervals for the simple regret of the maximiser of the posterior mean across previously queried points. Our proposed IMP-DPP is the only IPA strategy that provides consistently high performance.

6.2 Active Learning

To demonstrate the generality of our proposed IPA framework, we now depart from single-objective BO and instead consider an active learning task inspired by Balandat et al. (2020). We wish to learn which spatial locations in Nigeria have rates of a malaria-causing parasite *Plasmodium Falciparum* over a critical threshold.

We model the occurrence of a breach in the critical threshold at location $\mathbf{x}$ through a Bernoulli likelihood $y_{\mathbf{x}}|f \sim \mathcal{B}(\Phi(f(\mathbf{x})))$ where f denotes a latent sparse GP with 50



(a) Ground truth of test set (b) SVGP using CVR (c) SVGP using our active learning IPA

Figure 5: Incidence threshold breaches predicted by two surrogate models each fine-tuned over 10 steps of high-throughput active learning. (a) Performance is evaluated across a randomly sampled held-out test set. (b) CVR fails to accurately learn the complex classification boundary and obtains an accuracy of only 79%. (c) In contrast, our proposed IPA focuses inducing points (red dots) along the classification boundary, yielding a improved model with an accuracy of 89%.

inducing points and $\Phi : \mathbb{R} \to [0, 1]$ is the inverse probit function (see Hensman et al., 2015, for details). Starting from a random initial design of 100 evaluations, we then use the BALD acquisition function of Houlsby et al. (2011) to sequentially improve our model over 10 data acquisition steps, each time collecting evaluations at 100 informative locations then updating the classification surrogate models.

As the performance of the classifier is determined by the accuracy of its classification boundary (i.e., where $f \approx 0$), it is natural to consider a quality function that encourages the placing of inducing points where $|f|$ is small. To this end, we consider the active learning quality function

$$q_{\text{AL}}(\mathbf{z}) = \mathbb{E}_f \left[\hat{f} - |f(\mathbf{z})| \right], \tag{10}$$

where $\hat{f} = \max(|\max(f)|, |\min(f)|)$ is the largest absolute value obtained by the latent GP. This quality function has maximal score at $f = 0$, i.e., the level set of the latent GP corresponding to the classification boundary. Figure 5 demonstrates the benefit of using this custom quality function to drive IPA in the considered active learning problem.

6.3 Multi-objective Optimisation

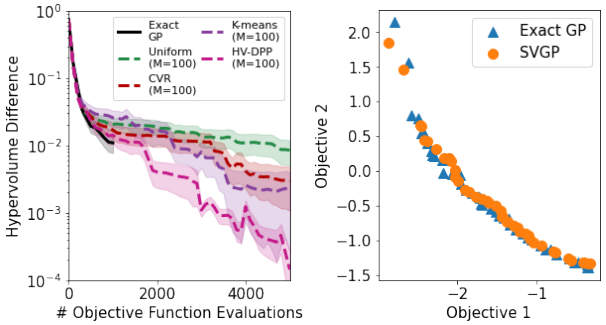
In multi-objective optimisation (MOO) we seek to find high-performing solutions according to K (≥ 2) competing objective functions $f^1(\mathbf{x}), \ldots, f^K(\mathbf{x})$. In these tasks, where improvements in one objective may harm another, the ability to characterise trade-offs between these competing objectives becomes crucial. Consequently, multi-objective optimisation corresponds to finding the so-called Pareto set which contains all locations representing optimal trade-offs, i.e., those that cannot be perturbed to yield an improved score in a single objective without a deterioration in the score of another objective (see Emmerich (2005) for an introduction). Therefore, when using sparse models as surrogate models for MOO, it is no longer sufficient to focus modelling resources into the "best" areas of the space; rather, we want to focus coverage around the Pareto front. Therefore we consider the quality function

$$q_{\text{HV}}(\mathbf{z}) = \mathbb{E}_{f_1,\ldots,f_K} \left[\prod_{k=1}^{K} \max(f_k(\mathbf{z}_i) - \hat{f}_k, 0) \right], \tag{11}$$

where f_k represents the model of the k^{th} objective function and $\hat{f}_k$ its minimal value, i.e., we consider a product of our single objective quality functions. As Eq. 11 can be interpreted as the Hyper-Volume (HV) of the set containing all the previously collected points that are dominated by $\mathbf{z}$, we refer to the IPA resulting from this quality function as HV-DPP. We allocate inducing points for each model separately but use the same shared quality function (that uses information from all the models) to encourage the allocation of points along the Pareto front. Although q_{HV} has a strong bias for points in the central area of the front, it is fast to evaluate and we found it adequate for enabling effective high-throughput BO. Future work will build a more sophisticated quality function that provides an even focus along the whole Pareto front.

Synthetic benchmark. Figure 6 demonstrates high-throughput optimisation of a noisy variant of the 4-dimensional ZDT3 problem (see Appendix E for a problem description). We start with 100 random evaluations and use the Chebyshev scalarisation acquisition function described by Paria et al. (2020) to collect 50 batches of 100 evaluations for the sparse methods (each with 100 inducing points), and 10 batches for the exact GP. As is standard practice in multi-objective optimisation, we measure performance in terms of the difference between the hyper-volume dominated by the true Pareto optimal front and the one found by BO.

Real-world problem. For our final example, we turn to the problem of designing an effective yet light-weight automotive heat exchanger (radiator), as considered by Paleyes et al. (2022) (see Appendix E for a full description). This challenging 9d problem has two objectives and three constraints, so requires five surrogate models, each of which will need



(a) 4d ZDT3 (b) 9d Heat Exchanger Design

Figure 6: Demonstration of high-throughput multi-objective optimisation. (a) Optimisation of the 4-dimensional ZDT3 problem over 50 runs, where only our proposed HV-DPP matches and then builds upon the performance of the exact GP. (b) The Pareto fronts found for the challenging heat exchanger design task by 1) an SVGP with our IPA strategy on only 100 inducing points alongside, and 2) exact GPs.

IPA. For the constraint models, we use the q_{AL} quality function (as presented for the active learning task) and for the objective models we use the q_{HV} quality function. We start with 100 random evaluations and use Paleyes et al. (2022)'s HIPPO acquisition function to allocate 10 batches of 100 further evaluations. Figure 6 shows that an SVGP surrogate model using our proposed IPA and only 100 inducing points is able to find a comparable Pareto front to an expensive exact GP. Moreover, in Appendix F, we provide wall-clock timing for these experiments, demonstrating that the SVGP incurs order-of-magnitude lower optimisation overheads than the exact GP.

7 Conclusions and Further Work

We have proposed the first BO-specific methods for selecting the locations of inducing points in sparse GPs. By exploiting the quality-diversity decomposition of DPPs, we are able to dramatically improve DPP-based IPA, transforming what is often, in the context of BO, a poorly performing IPA (the conditional variance reduction of Burt et al., 2019) to the best (our IMP-DPP). Moreover, we have shown that our proposed framework provides a general framework for ensuring that sparse GPs are accurate in key areas, and so has applications across a range of down-stream tasks.

In future work we will apply our BO-specific IPAs to real-world problems where sparse GPs are already being used, e.g., quantile optimisation (Torossian et al., 2020). We will also investigate their applicability to other inducing point-based methods that are also used in decision making loops, like deep GPs (Damianou and Lawrence, 2013). Moreover, our SVGPs could also be applied to high-dimensional optimisation problems by extending single-model trust region approaches (Diouane et al., 2022) to support large optimisation budgets. Finally, note that our proposed IPA does not require Euclidean input spaces (unlike standard SVGP formulations which optimise inducing point locations using gradient descent). Therefore, we also wish to use our proposed scalable surrogate models to enable high-throughput versions of active learning loops over discrete structures that can be modelled with GPs, e.g., genes (Moss et al., 2020a) and molecules (Moss and Griffiths, 2020; Thawani et al., 2020; Griffiths et al., 2022; Ranković et al., 2022).

References

Balandat, M., Karrer, B., Jiang, D., Daulton, S., Letham, B., Wilson, A. G., and Bakshy, E. (2020). BoTorch: A framework for efficient Monte-Carlo Bayesian optimization. *Advances in Neural Information Processing Systems*.

Belabbas, M.-A. and Wolfe, P. J. (2009). Spectral methods in machine learning and new strategies for very large datasets. *Proceedings of the National Academy of Sciences*, 106(2):369–374.

Berkeley, J., Moss, H. B., Artemev, A., Pascual-Diaz, S., Granta, U., Stojic, H., Couckuyt, I., Qing, J., Loka, N., Paleyes, A., Ober, S. W., and Picheny, V. (2022). Trieste. *https://github.com/secondmind-labs/trieste*.

Burt, D., Rasmussen, C. E., and Van Der Wilk, M. (2019). Rates of convergence for sparse variational Gaussian process regression. In *International Conference on Machine Learning (ICML)*.

Burt, D. R., Rasmussen, C. E., and van der Wilk, M. (2020). Convergence of sparse variational inference in Gaussian processes regression. *Journal of Machine Learning Research (JMLR)*, 21:131:1–131:63.

Chang, P. E., Verma, P., John, S., Picheny, V., Moss, H., and Solin, A. (2022). Fantasizing with dual GPs in Bayesian optimization and active learning. *arXiv preprint arXiv:2211.01053*.

Chen, L., Zhang, G., and Zhou, H. (2018). Fast greedy MAP inference for determinantal point process to improve recommendation diversity. In *Advances Neural Information Processing Systems*.

Cole, D. A., Christianson, R. B., and Gramacy, R. B. (2021). Locally induced Gaussian processes for large-scale simulation experiments. *Statistics and Computing*, 31(3):1–21.

Cole, D. A., Gramacy, R. B., and Ludkovski, M. (2022). Large-scale local surrogate modeling of stochastic simulation experiments. *Computational Statistics & Data Analysis*, page 107537.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Csató, L. and Opper, M. (2002). Sparse on-line Gaussian processes. *Neural computation*, 14(3):641–668.

Damianou, A. and Lawrence, N. D. (2013). Deep Gaussian processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Daulton, S., Eriksson, D., Balandat, M., and Bakshy, E. (2022). Multi-objective Bayesian optimization over high-dimensional search spaces. In *Conference on Uncertainty in Artificial Intelligence (UAI)*.

Diouane, Y., Picheny, V., Riche, R. L., and Perrotolo, A. S. D. (2022). TREGO: A trust-region framework for efficient global optimization. *Journal of Global Optimization*, pages 1–23.

Dodge, J., Jamieson, K., and Smith, N. A. (2017). Open loop hyperparameter optimization and determinantal point processes. *arXiv preprint arXiv:1706.01566*.

Emmerich, M. (2005). *Single-and multi-objective evolutionary design optimization assisted by Gaussian random field metamodels*. PhD thesis, Dortmund, Univ., Diss., 2005.

Eriksson, D., Pearce, M., Gardner, J., Turner, R. D., and Poloczek, M. (2019). Scalable global optimization via local Bayesian optimization. *Advances in Neural Information Processing Systems*.

Frazier, P. I., Powell, W. B., and Dayanik, S. (2008). A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439.

Gramacy, R. B. and Apley, D. W. (2015). Local Gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics*, 24(2):561–578.

Griffiths, R.-R. and Hernández-Lobato, J. M. (2020). Constrained Bayesian optimization for automatic chemical design using variational autoencoders. *Chemical science*, 11(2):577–586.

Griffiths, R.-R., Klarner, L., Moss, H., Ravuri, A., Truong, S. T., Rankovic, B., Du, Y., Jamasb, A. R., Schwartz, J., Tripp, A., et al. (2022). GAUCHE: A library for Gaussian processes in chemistry. In *ICML 2022 2nd AI for Science Workshop*.

Hennig, P. and Garnett, R. (2016). Exact sampling from determinantal point processes. *arXiv preprint arXiv:1609.06840*.

Hennig, P. and Schuler, C. J. (2012). Entropy search for information-efficient global optimization. *Journal of Machine Learning Research (JMLR)*, 13(6).

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*.

Hensman, J., Matthews, A., and Ghahramani, Z. (2015). Scalable variational Gaussian process classification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. (2014). Predictive entropy search for efficient global optimization of black-box functions. *arXiv preprint arXiv:1406.2541*.

Hernández-Lobato, J. M., Requeima, J., Pyzer-Knapp, E. O., and Aspuru-Guzik, A. (2017). Parallel and distributed Thompson sampling for large-scale accelerated exploration of chemical space. In *International Conference on Machine Learning (ICML)*.

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. (2011). Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

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

Kandasamy, K., Krishnamurthy, A., Schneider, J., and Póczos, B. (2018). Parallelised Bayesian optimisation via Thompson sampling. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Kathuria, T., Deshpande, A., and Kohli, P. (2016). Batched Gaussian process bandit optimization via determinantal point processes. *Advances in Neural Information Processing Systems*.

Ko, C.-W., Lee, J., and Queyranne, M. (1995). An exact algorithm for maximum entropy sampling. *Operations Research*, 43(4):684–691.

Kulesza, A. and Taskar, B. (2012). Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*.

Lázaro-Gredilla, M., Quinonero-Candela, J., Rasmussen, C. E., and Figueiras-Vidal, A. R. (2010). Sparse spectrum Gaussian process regression. *The Journal of Machine Learning Research (JMLR)*, 11:1865–1881.

Maddox, W. J., Stanton, S., and Wilson, A. G. (2021). Conditioning sparse variational Gaussian processes for online decision-making. *Advances in Neural Information Processing Systems*.

Matthews, A. G. d. G., Hensman, J., Turner, R., and Ghahramani, Z. (2016). On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

McIntire, M., Ratner, D., and Ermon, S. (2016). Sparse Gaussian processes for Bayesian optimization. In *Conference on Uncertainty in Artificial Intelligence (UAI)*.

Moss, H., Leslie, D., Beck, D., Gonzalez, J., and Rayson, P. (2020a). BOSS: Bayesian optimization over string spaces. *Advances in Neural Information Processing Systems*.

Moss, H. B. and Griffiths, R.-R. (2020). Gaussian process molecule property prediction with FlowMO. *arXiv preprint arXiv:2010.01118*.

Moss, H. B., Leslie, D. S., Gonzalez, J., and Rayson, P. (2021). GIBBON: General-purpose information-based Bayesian optimisation. *arXiv preprint arXiv:2102.03324*.

Moss, H. B., Leslie, D. S., and Rayson, P. (2020b). BOSH: Bayesian optimization by sampling hierarchically. *arXiv preprint arXiv:2007.00939*.

Moss, H. B., Leslie, D. S., and Rayson, P. (2020c). MUMBO: Multi-task max-value Bayesian optimization. *arXiv preprint arXiv:2006.12093*.

Moss, H. B., Ober, S. W., and Picheny, V. (2022). Information-theoretic inducing point placement for high-throughput Bayesian optimisation. *arXiv preprint arXiv:2206.02437*.

Nava, E., Mutny, M., and Krause, A. (2022). Diversified sampling for batched Bayesian optimization with determinantal point processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Nickson, T., Osborne, M. A., Reece, S., and Roberts, S. J. (2014). Automated machine learning on big data using stochastic algorithm tuning. *arXiv preprint arXiv:1407.7969*.

Paleyes, A., Moss, H. B., Picheny, V., Zulawski, P., and Newman, F. (2022). A penalisation method for batch multi-objective Bayesian optimisation with application in heat exchanger design. *arXiv preprint arXiv:2206.13326*.

Paria, B., Kandasamy, K., and Póczos, B. (2020). A flexible framework for multi-objective Bayesian optimization using random scalarizations. In *Conference on Uncertainty in Artificial Intelligence (UAI)*.

Picheny, V., Moss, H., Torossian, L., and Durrande, N. (2022). Bayesian quantile and expectile optimisation. In *Conference on Uncertainty in Artificial Intelligence (UAI)*.

Qing, J., Moss, H. B., Dhaene, T., and Couckuyt, I. (2022). Parallel feasible pareto frontier entropy search for multi-objective Bayesian optimization under unknown constraints. *arXiv preprint arXiv:2204.05411*.

Ranković, B., Griffiths, R.-R., Moss, H. B., and Schwaller, P. (2022). Bayesian optimisation for additive screening and yield improvements in chemical reactions–beyond one-hot encodings. In *2022 ELLIS Ml4Molecules Workshop*.

Rasmussen, C. E. and Williams, C. K. (2006). Gaussian Processes for Machine Learning. *ISBN-13 978-0-262-18253-9*.

Ru, B., Osborne, M. A., McLeod, M., and Granziol, D. (2018). Fast information-theoretic Bayesian optimisation. In *International Conference on Machine Learning (ICML)*.

Rullière, D., Durrande, N., Bachoc, F., and Chevalier, C. (2018). Nested kriging predictions for datasets with a large number of observations. *Statistics and Computing*, 28(4):849–867.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and De Freitas, N. (2016). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.

Takeno, S., Fukuoka, H., Tsukada, Y., Koyama, T., Shiga, M., Takeuchi, I., and Karasuyama, M. (2019). Multi-fidelity Bayesian optimization with max-value entropy search. *arXiv preprint arXiv:1901.08275*.

Thawani, A. R., Griffiths, R.-R., Jamasb, A., Bourached, A., Jones, P., McCorkindale, W., Aldrick, A. A., and Lee, A. A. (2020). The photoswitch dataset: A molecular machine learning benchmark for the advancement of synthetic chemistry. *arXiv preprint arXiv:2008.03226*.

Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Torossian, L., Picheny, V., and Durrande, N. (2020). Bayesian quantile and expectile optimisation. *arXiv preprint arXiv:2001.04833*.

Vakili, S., Moss, H., Artemev, A., Dutordoir, V., and Picheny, V. (2021). Scalable Thompson sampling using sparse Gaussian process models. *Advances in Neural Information Processing Systems*.

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

Wang, Z., Li, C., Jegelka, S., and Kohli, P. (2017). Batched high-dimensional Bayesian optimization via structural kernel learning. In *International Conference on Machine Learning (ICML)*.

Wilson, J., Borovitskiy, V., Terenin, A., Mostowsky, P., and Deisenroth, M. (2020). Efficiently sampling functions from Gaussian process posteriors. In *International Conference on Machine Learning (ICML)*.

Yang, A., Li, C., Rana, S., Gupta, S., and Venkatesh, S. (2021). Sparse spectrum Gaussian process for Bayesian optimization. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 257–268. Springer.

Zitzler, E., Deb, K., and Thiele, L. (2000). Comparison of multiobjective evolutionary algorithms: Empirical results. *Evolutionary computation*, 8(2):173–195.

A Theoretical Justification

For a more formal intuition into the unsuitability of existing IPA strategies for BO and the suitability of our proposed method we can apply Lemma 4 of Burt et al. (2020), which bounds the approximation error between an exact GP and its sparse approximation and is restated below for convenience.

Theorem 2. ***(Burt et al., 2020, Lemma 4)*** *Suppose $\mathbf{y}|X, Z \sim \mathcal{N}(0, K_X + \sigma^2 I_N)$. Then, for our exact posterior P and our variational posterior Q, we have that for any X and Z*

$$t(Z)/(2\sigma^2) \leq \mathbb{E}\left[KL[Q||P]|Z, X\right] \leq t(Z)/\sigma^2,$$

where $KL[Q||P]$ denotes the Kullback-Leibler divergence and $t(Z) = tr(K_X - Q_X(Z))$, for $Q_X(Z) := K_Z(X)^T K_Z^{-1} K_Z(X)$.

Now, we can state our Corollary 1 which bounds the maximal and minimal discrepancy between an SVGP and an exact GP over a localised area.

Corollary 1. *Suppose $\mathbf{y}|\boldsymbol{X}, \boldsymbol{Z} \sim \mathcal{N}(0, K_X + \sigma^2 I_N)$. Then, for our exact posterior P and our variational posterior Q, we have that for any subspace $A \subseteq \mathcal{X}$*

$$t_A(Z)/(2\sigma^2) \leq \mathbb{E}\left[KL_A[Q||P]|Z, X\right] \leq t_A(Z)/\sigma^2.$$

Here $KL_A[Q||P]$ denotes the Kullback-Leibler divergence calculated only over datapoints contained in the subspace $A \subseteq \mathcal{X}$ and $t_A(Z) = tr(K_{X_A} - Q_{X_A}(Z))$, where X_A denotes the subset of data X lying in A and $Q_{X_A}(Z) := K_Z(X_A)^T K_Z^{-1} K_Z(X_A)$.

Proof. Noting that

$$KL_A[Q||P] = KL\left[\mathcal{N}(0, K_{X_A} + \sigma^2 I)||\mathcal{N}(0, Q_{X_A} + \sigma^2 I)\right],$$

we can directly apply Lemma 4 of Burt et al. (2020) but with K_{X_A} instead of K_X and Q_{X_A} instead of Q_X to get the required result. □

By applying Corollary 1 in the case $A = \mathcal{X}$, Burt et al. (2019) justify that the goal of IPA strategies should be to minimise the trace term $t_{\mathcal{X}}(Z)$, as minimising $t_{\mathcal{X}}(Z)$ ensures a small divergence between our SVGP and the exact GP (a notion formalised in Matthews et al., 2016), which in turn (through Proposition 1 of Burt et al., 2019) ensures a small approximation error across the whole search space. Sampling inducing points from a DPP can then be justified as, under such a sampling scheme, the resulting $t_{\mathcal{X}}(Z)$ lying within a constant factor of its minimal achievable value (see Theorem 1 of Belabbas and Wolfe, 2009).

Now suppose that, as is the case in BO (see Figure 2), we want to ensure our variational approximation is especially accurate in a particular region $A \subseteq \mathcal{X}$, i.e., we seek IPAs with small values of $t_A(Z)$. However, noting the decomposition $t_{\mathcal{X}}(Z) = t_A(Z) + t_{\mathcal{X}/A}(Z)$, it is clear that the minimisation of $t_A(Z)$ forms only a small part of the overall IPA objective targeted by DPP sampling, especially in BO applications where the promising areas A often contains only a small proportion of the search space.

We can now understand the benefit of including a quality function in our DPP through Corollary 2, an extension of Lemma 1 of Belabbas and Wolfe (2009). Before presenting our Corollary, we restate this Lemma as it is used by Burt et al. (2020) to justify using a DPP for IPA.

Theorem 3. ***(Belabbas and Wolfe, 2009, Theorem 1)*** *Let $\eta_1 \geq ... \geq \eta_N \geq 0$ be the eigenvalues of the SPSD matrix K_X. Suppose a set of points Z are sampled according to an M-determinantal point process with kernel matrix K. Define the matrix $Q_X(Z) = K_Z(X)^T K_Z^{-1} K_Z(X)$ and the trace term $t(Z) = tr(K_X - Q_X(Z))$. Then,*

$$\mathbb{E}\left[t(Z))\right] \leq (M+1) \sum_{m=M+1}^{N} \eta_m. \tag{12}$$

Theorem 3 tells us that using an M-DPP to perfom IPA will make $t(Z)$ relatively close to its minimal value of $\sum_{m=M+1}^{N} \eta_m$ (see Section 4.2.1 of Burt et al., 2020, for a discussion of why $\min_{Z \in \mathcal{X}^M} t(Z) = \sum_{m=M+1}^{N} \eta_m$). Therefore, in the context of

Theorem 2, Theorem 3 guarantees low KL divergence between an exact GP and the sparse approximation arising from a DPP IPA strategy.

Now we can present our Corollary 2, which adapts the above results for M-DPPs with quality functions (i.e., $K \to L$). In particular, we show that when sampling from a DPP with a quality function q, we are guaranteed (in expectation) to have $t^q(Z)$ lying within a constant factor of its minimal achievable value. Therefore, by increasing q in our promising region A and reducing it elsewhere, we increase the contribution of the individual components corresponding to the terms of $t_A(Z)$ in the overall objective $t^q(Z)$, thus ensuring low KL divergence between our SVGP and the exact posterior in A.

Corollary 2. *Suppose that Z is sampled from a DPP with quality function $q : \mathcal{X} \to \mathbb{R}^+$. Then*

$$\mathbb{E}\left[t^q(Z)\right] \leq (M+1)\hat{t}^q$$

for a q-weighted trace term $t^q(Z) = \sum_{i=1}^N q(\boldsymbol{x}_i)^2 \left([K_X]_{i,i} - [Q_X(Z)]_{i,i}\right)$ and its minimal achievable value $\hat{t}^q = \min_{Z' \in \mathcal{X}^M} t^q(Z')$.

Proof. A direct application of Theorem 3 but with K_X replaced by

$$L_X = [q(\mathbf{x}_i)k(\mathbf{x}_i, \mathbf{x}_j)q(\mathbf{x}_j)]_{(\mathbf{x}_i, \mathbf{x}_j) \in X \times X}$$

yields

$$\mathbb{E}\left[t_*(Z)\right] \leq (M+1) \min_{Z \in \mathcal{X}^M} \hat{t}_*(Z),$$

for $t_*(Z) = tr(L_X - L_Z(X)^T L_Z^{-1} L_Z(X))$. All that remains is to show that $t^q(Z) = t_*(Z)$.

Note that we can write $L_Z(X) = D_q(Z) K_Z(X) D_q(X)$, where $D_q(X)$ is a diagonal matrix with non-zero entries given by the vector $[q(\mathbf{x})]_{\mathbf{x} \in \mathcal{X}}$. Therefore, after routine algebraic manipulations, we have

$$\begin{aligned}
t_*(Z) &= tr(L_X - L_Z(X)^T L_Z^{-1} L_Z(X)) \\
&= tr(D_q(X) K_X D_q(X) - D_q(X) Q_X(Z) D_q(X)) \\
&= \sum_{i=1}^N q(\mathbf{x}_i)^2 \left([K_X]_{i,i} - [Q_X(Z)]_{i,i}\right) \\
&= t_q(Z)
\end{aligned}$$

□

To help explain why Corollary 2 justifies the use of DPPs with quality functions as IPA strategies in BO, consider the simple demonstrative binary quality function

$$q(\mathbf{z}) = \begin{cases} \sqrt{\beta} & \text{for } \mathbf{z} \in A, \\ \sqrt{1-\beta} & \text{otherwise,} \end{cases}$$

under which $t^q(Z) = \beta^2 * t_A(Z) + (1-\beta)^2 * t_{\mathcal{X}/A}(Z)$. Therefore, by varying the quality function (through β), we re-weight the contributions $t_A(Z)$ and $t_{\mathcal{X}/A}(Z)$ in $t^q(Z)$, i.e., we change the relative trade-off allocated by our DPP on local (inside A) and global (outside A) modelling. Although our practical recommendations for quality functions in Section 4 are much more sophisticated than this binary function, the intuition remains the same.

B Proof of Theorem 1

We now restate and prove Theorem 1 that demonstrates the effect of the choice of q on the inducing points chosen by the IPA. This Theorem is used in the main paper to explain the constraints imposed upon q in order to achieve scale and translation invariant IPA strategies.

Theorem 1. *Suppose inducing points $\mathcal{Z}$ are distributed according to a DPP with similarity kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and quality function $q : \mathcal{X} \to \mathbb{R}$, i.e., $\mathbb{P}(\mathcal{Z} = Z) \propto \left|L_Z\right|$. Then, according to the greedy approximation, the j^{th} component of the MAP estimate of $\mathcal{Z}$ is given by*

$$\mathbf{z}_j = \underset{z \in X}{\operatorname{argmax}}\ q(\mathbf{z})\sigma_{j-1}(\mathbf{z}), \tag{7}$$

where $\sigma^2_{j-1}(\mathbf{z})$ is the conditional variance of the noise-free GP model conditioned on the already selected points $Z_{1:j-1}$ (cf. Eq. 4).

Proof. As derived in Eq. (4) of Hennig and Garnett (2016), greedy maximisation of a DPP with a kernel k corresponds to setting the j^{th} component of the MAP estimate of as

$$\mathbf{z}_j = \underset{\mathbf{z} \in X}{\operatorname{argmax}}\ \sigma_{j-1}(\mathbf{z}),$$

where $\sigma^2_{j-1}(\mathbf{z}) = k(\mathbf{z}, \mathbf{z}) - \mathbf{k}_{Z_{1:j-1}}(\mathbf{z})^T K^{-1}_{Z_{1:j-1}} \mathbf{k}_{Z_{1:j-1}}(\mathbf{z})$.

We can apply exactly the same derivation to a DPP with a similarity kernel $l(\mathbf{x}, \mathbf{x}') = q(\mathbf{x})k(\mathbf{x}, \mathbf{x}')q(\mathbf{x}')$, i.e., as arising from including a quality function q, to get

$$\mathbf{z}_j = \underset{\mathbf{z} \in X}{\operatorname{argmax}}\ \hat{\sigma}_{j-1}(\mathbf{z}), \tag{13}$$

where $\hat{\sigma}^2_{j-1}(\mathbf{z}) = l(\mathbf{z}, \mathbf{z}) - \mathbf{l}_{Z_{1:j-1}}(\mathbf{z})^T L^{-1}_{Z_{1:j-1}} \mathbf{l}_{Z_{1:j-1}}(\mathbf{z})$ for $\mathbf{l}_{Z_{1:j-1}}(\mathbf{z}) = [l(\mathbf{z}', \mathbf{z})]_{\mathbf{z}' \in \mathcal{Z}_{1:j-1}}$ and $L_{Z_{1:j-1}} = [l(\mathbf{z}, \mathbf{z}')]_{\mathbf{z}, \mathbf{z}' \in Z_{1:j-1} \times Z_{1:j-1}}$.

Note that we can expand $\mathbf{l}_{Z_{1:j-1}}(\mathbf{z}) = D_q(Z_{1:j-1})\mathbf{k}_{Z_{1:j-1}}(\mathbf{z})q(\mathbf{z})$ and $L_{Z_{1:j-1}} = D_q(Z_{1:j-1})K_{Z_{1:j-1}}D_q(Z_{1:j-1})$ for a diagonal matrix $D_q(Z_{1:j-1})$ with non-zero entries given by the vector $[q(\mathbf{z}_i)]_{i=1}^{j-1}$. Therefore, we can expand Equation 13 as

$$\begin{aligned}
\hat{\sigma}^2_{j-1}(\mathbf{z}) =& q(\mathbf{z})k(\mathbf{z}, \mathbf{z})q(\mathbf{z}) \\
&- q(\mathbf{z})\mathbf{k}^T_{Z_{1:j-1}}(\mathbf{z})D_q(Z_{1:j-1}) \\
&* D_q(Z_{1:j-1})^{-1}K^{-1}_{Z_{1:j-1}}D_q(Z_{1:j-1})^{-1} \\
&* D_q(Z_{1:j-1})\mathbf{k}_{Z_{1:j-1}}(\mathbf{z})D_q(Z_{1:j-1})q(\mathbf{z}) \\
=& q(\mathbf{z})^2\sigma^2_{j-1}(\mathbf{z}).
\end{aligned}$$

Therefore, for the greedy MAP estimate of a DPP with quality function q, we have

$$\begin{aligned}
\mathbf{z}_j &= \underset{\mathbf{z} \in X}{\operatorname{argmax}}\ \hat{\sigma}_{j-1}(\mathbf{z}) \\
&= \underset{\mathbf{z} \in X}{\operatorname{argmax}}\ q(\mathbf{z})\sigma_{j-1}(\mathbf{z}).
\end{aligned} \tag{14}$$

□

C An Information-theoretic Approach

We now propose our second quality function, this time chosen such that the resulting IPA, which we name ENT-DPP, can be viewed as maximising an intuitive information-theoretic quantity well-aligned with the goal of BO. This work is also available in the non-archival technical report Moss et al. (2022).

Reducing global uncertainty. For GP regression tasks it is sufficient to choose points that provide maximal information about the whole objective function f. In fact, the arguments of Srinivas et al. (2009) and Hennig and Garnett (2016) demonstrate that providing the maximal reduction in uncertainty corresponds exactly to the DPP MAP objective (2) targeted by the CVR IPA strategy. We repeat this result below as Theorem 4. See Srinivas et al. (2009) for a discussion of information theory in the context of GP learning and Cover (1999) for a general introduction.

Theorem 4. *Under a GP prior $f \sim \mathcal{GP}(0, k)$, maximising the M-DPP objective (2) is equivalent to choosing our inducing points Z as the datapoints with evaluations $\mathbf{y}_Z$ that provide the largest gain in information about the unknown function f*

$$Z = \underset{Z' \subseteq X : |Z'| = M}{\operatorname{argmax}} \mathit{IG}(\mathbf{y}_{Z'}; f).$$

where $\mathit{IG}(\mathbf{y}_Z, f) = H(f) - H(f|\mathbf{y}_Z)$ quantifies the reduction in the differential entropy H of f provided by revealing the evaluations $\mathbf{y}_Z$.

Reducing Uncertainty in f^*. In BO it is natural to also consider how much information is provided about the function's maximum value $f^* = \max f(\mathbf{x})$. Taking inspiration from the empirical success of max-value entropy search acquisition functions (Wang and Jegelka, 2017; Takeno et al., 2019; Moss et al., 2020c,b; Qing et al., 2022), where query points in BO loops are chosen to reduce our current uncertainty in f^*, we thus propose choosing inducing points Z that maximise the following trade-off criterion:

$$C_\alpha(Z) = \alpha \times \text{IG}(\mathbf{y}_Z; f^*) + (1 - \alpha) \times \text{IG}(\mathbf{y}_Z; f). \tag{15}$$

where the information gain $\text{IG}(\mathbf{y}_Z, f^*) = H(f^*) - H(f^*|\mathbf{y}_Z)$ quantifies the reduction in uncertainty provided by these inducing points about the maximum f^* (see Wang and Jegelka, 2017, for a derivation), and $\text{IG}(\mathbf{y}_Z; f)$ is as in Theorem 4. $\alpha \in [0, 1]$ controls the balance of modelling resources spent on global variations and those spent in areas around potential maxima. Note that setting $\alpha = 1$ returns the criterion targeted by CVR.

Approximation with a DPP. The lack of closed-form expression for the distribution of f^* renders the calculation of this criterion (15) challenging, and it is unclear how this objective fits into our DPP framework. Fortunately, there exists a rich literature of methods for approximately optimising similar information-theoretic quantities (Hennig and Schuler, 2012; Hernández-Lobato et al., 2014). In particular, we can follow the ideas of Moss et al. (2021) and use common information-theoretic inequalities to replace our desired criterion $C_\alpha(Z)$ with a simpler lower bound that takes the form of a DPP. Specifically, we use the well-known inequality $H(\mathbf{y}_Z|f^*) \leq \sum_{i=1}^M H(y_{\mathbf{z}_i}|f^*)$. After simple mathematical manipulations, the resulting lower bound for the trade-off in entropy can be expressed in the following form:

$$C_\alpha(Z) \geq \frac{1 - \alpha}{2} \log \left| L_Z \right|, \tag{16}$$

where L_Z is defined as in (5) but with quality function

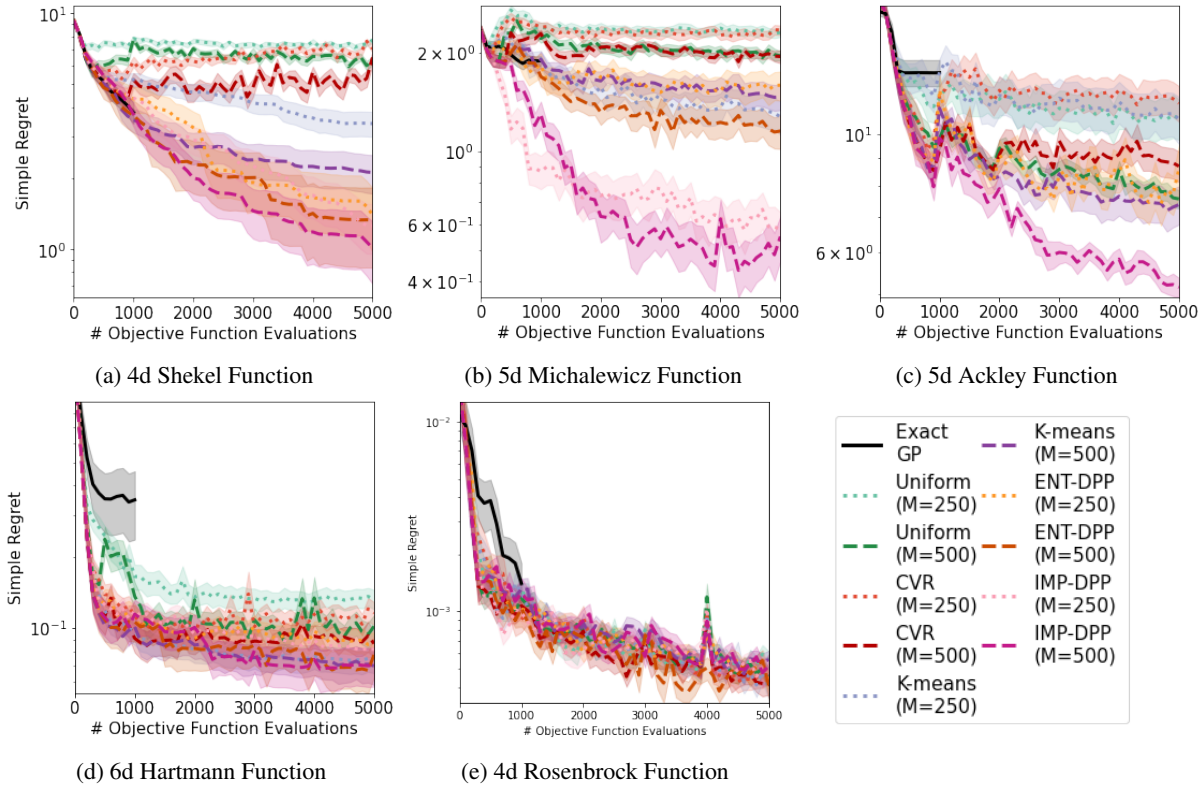
$$q(\mathbf{z}) = \exp\left(\frac{1}{M} \frac{\alpha}{1 - \alpha} \text{IG}(y_{\mathbf{z}}; f^*) \right), \tag{17}$$

a function increasing in $\text{IG}(y_{\mathbf{z}}; f^*)$ — the reduction in uncertainty provided about f^* by a single observation $y_{\mathbf{z}}$. Although we cannot calculate $\text{IG}(y_{\mathbf{z}}; f^*)$ in (17) exactly, we can follow Ru et al. (2018) and approximate $q(\mathbf{z})$ with moment-matching.

Experimental Results. We now test the IPA resulting from the entropy-based quality function 17, which we name ENT-DPP. All our experiments use $\alpha = 0.5$ as we found performance to be unaffected by all but extreme choices of α (i.e., those very close to 0 or 1). In Figure 7, we repeat all the single-objective benchmark experiments presented in the main body of the paper. We see that, although providing a performance boost over existing IPA, our well-motivated ENT-DPP is not as performant as our (simpler) IMP-DPP. Indeed, the superiority of IMP-DPP over this carefully constructed ENT-DPP provides yet further justification for the practical use of IMP-DPP. He believe that the poor performance of ENT-DPP is due to the required approximations, which degrade as we increase the number of inducing points. More precisely, the bound $H(\mathbf{y}_Z; f^*) \leq \sum_{i=1}^M H(y_{\mathbf{z}_i}; f^*)$ becomes looser as we increase the number M of inducing points $\mathbf{z}_i$ and they become closer together (on average).

D DPP-IMP Ablation Study

In order to justify the quality function used for our proposed IMP-DPP IPA strategy, we now perform an ablation study where we demonstrate the robustness of this IPA to small changes to its quality function. We compare the IPA resulting from the proposed quality function (the expected improvement with respect to the minimal predicted value of the function) with some other reasonable choices. In particular, Figure 8 demonstrates that IMP-DPP still outperforms existing IPA methods even when swapping its ReLU with a softmax approximation or when using the mean predicted value of the function instead of the minimal predicted value as its baseline.



(a) 4d Shekel Function (b) 5d Michalewicz Function (c) 5d Ackley Function

(d) 6d Hartmann Function (e) 4d Rosenbrock Function

Figure 7: Results are averaged over 50 runs and we report the mean and its 95% confidence intervals for the simple regret of the maximiser of the posterior mean across previously queried points. Although ENT-DPP does improve over existing approaches across most of the test functions, our IMP-DPP achieves much more impressive and consistent performance.

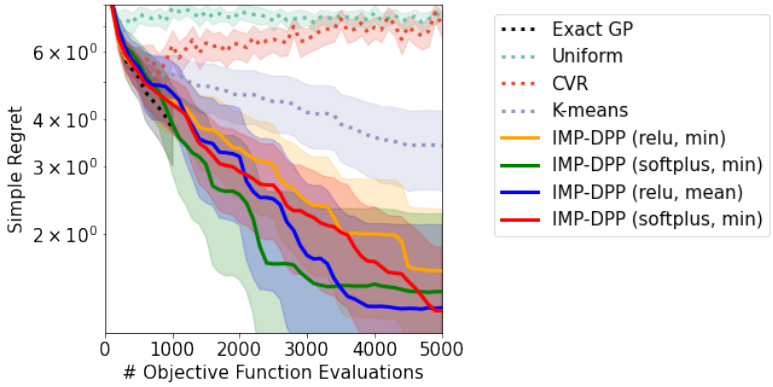


Figure 8: Regret achieved on the 4d Shekel function for an exact GP and SVGPs with $M = 250$ inducing points. All four variants of our IMP-DPP outperfom exisiting IPA strategies and achieve statistically similar results.

E Additional Experimental Details

Synthetic baselines. For details on the Shekel, Hartmann and Ackley functions see Appendix C of Vakili et al. (2021), and for the multi-objective ZDT3 see Zitzler et al. (2000).

We now present the analytical forms of the remaining Michalewicz and Rosenbrock functions.

Michalewicz function. A five-dimensional function with $5!$ local minima and a single global minimum defined on $\mathcal{X} \in [0, \pi]^5$:

$$f(\mathbf{x}) = -\sum_{i=1}^{d} \sin(x_i) \sin^{20}(\frac{ix_i^2}{\pi}).$$

Rosenbrock function. A unimodal four-dimensional function a single global minimum lying in a narrow valley defined on $\mathcal{X} \in [-5, 10]^5$:

$$f(\mathbf{x}) = \sum_{i=1}^{3} \left[100(x_{i+1} - x_i^2) + (x_i - 1)^2\right].$$

Heat exchanger design. For our final experiment, we considered the real-world constrained multi-objective problem of designing a both effective and light-weight heat exchanger (radiator) for a car, as parameterised by its dimensions and internal geometry. Performance is simulated using areo-thermal analysis as part of an expensive-to-evaluate and highly non-linear digital twin. There are 6 continuous and 3 discrete inputs and three constraints that ensure a minimal standard of performance and safety. All our models use Matérn-5/2 kernels and a constant mean function. See Paleyes et al. (2022) for further details about this problem.

F Wall-clock Timings

6d Hartmann. Figure 9 demonstrates the significant reduction in overhead (i.e., the computation required to generate the next set of query points) provide by using SVGPs instead of their exact counterparts. The experiments for Figure 9 were performed on a quad core 2.40GHz Intel Xeon CPU. We highlight that the GP is more expensive than the SVGP even when $N < M$. This perhaps surprising observation is due to the different optimisers used which have different early-stopping tolerances, i.e. the GP uses LBFGS whereas the SVGP uses Adam. Although the GP could likely be made a bit cheaper through appropriate tuning of the LBFGS stopping criterion, it will still be significantly more expensive than the SVGP when $M << N$.

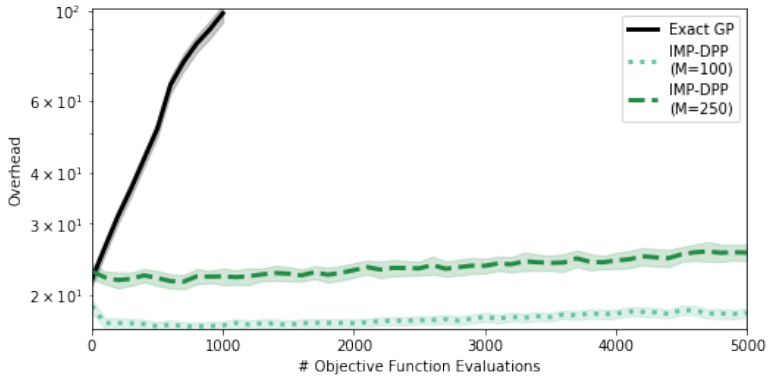


Figure 9: The computational overhead incurred by using different surrogate models when optimising the 6d Hartmann function. We see that our SVGP-based approaches have significantly lower overheads than the prohibitively expensive exact GP. Moreover, unlike the exact GP, the SVGP approaches maintain a constant overhead as BO progresses.

Heat exchanger design. In the main body of the paper we showed that an SVGP surrogate model using our proposed IPA and only 100 inducing points is able to find a comparable Pareto front to an expensive exact GP. We now present additional

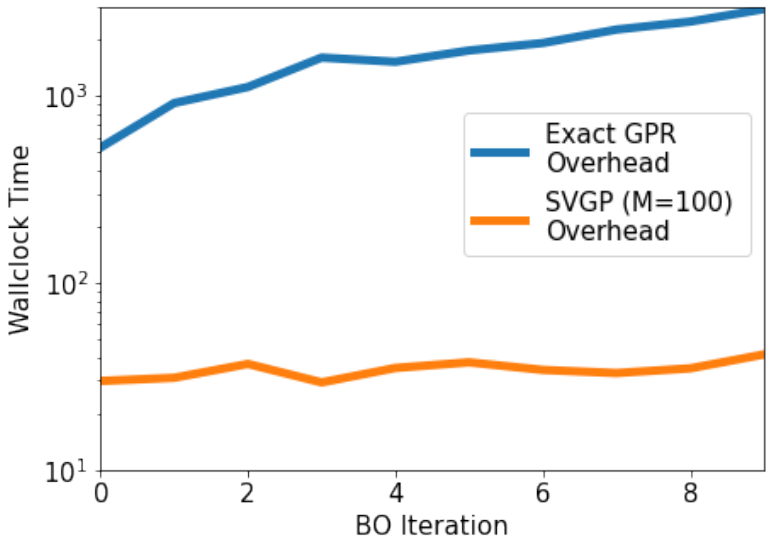


Figure 10: Time taken (on a log scale) to fit the surrogate models and maximise acquisition functions for each BO iteration of our heat exchanger design problem. BO with an SVGP incurs order of magnitude lower costs than BO with an exact GP. Note also that the exact GP gets increasingly expensive as the optimisation progresses, whereas the SVGP's overhead remains roughly the same.

insights through Figure 10 which shows the total time spent running BO for each experiment. These wall-clock timings only measure the time spent fitting models and maximising acquisition functions, not the time spent evaluating the heat exchanger simulator. Figure 10 demonstrates that the SVGP incurs order-of-magnitude lower optimisation overheads than the exact GP, opening up the feasibility of using BO on only moderately expensive functions, not just those with function query costs sufficiently large to overshadow the very significant optimisation overheads incurred by the exact GP models. These experiments were performed on a single NVIDIA GeForce GTX 1070 GPU.