Gradient-based Sample Selection for Faster Bayesian Optimization

Anonymous authors
Paper under double-blind review

Abstract

Bayesian optimization (BO) is an effective technique for black-box optimization. However, its applicability is typically limited to moderate-budget problems due to the cubic complexity of fitting the Gaussian process (GP) surrogate model. In large-budget scenarios, directly employing the standard GP model faces significant challenges in computational time and resource requirements. In this paper, we propose Gradient-based Sample Selection Bayesian Optimization (GSSBO), a subset-maintenance approach designed to enhance the computational efficiency of BO. Here, "gradient-based" refers to response-gradient sensitivity embeddings induced by the GP log marginal likelihood, not derivative observations of the objective function. The GP model is constructed on a selected set of samples instead of the whole dataset. These samples are selected using GP-specific sensitivity embeddings and a cosine-diversity rule to encourage embedding-space diversity and reduce redundancy in the retained subset. We provide a conditional theoretical analysis of the subset-fitted GP surrogate induced by a fixed retained subset and derive posterior-approximation bounds in terms of subset-dependent residual quantities. Experiments on synthetic and real-world tasks show empirical reductions in GP-fitting cost while achieving competitive optimization performance in the reported settings.

1 Introduction

Bayesian optimization (BO) (Frazier, 2018; Couckuyt et al., 2022) is a successful approach to black-box optimization that has been applied in a wide range of applications, such as hyperparameter optimization (Luo et al., 2021; 2024) and resource exploration (Branke et al., 2024). BO's strength lies in its ability to represent the unknown objective function through a surrogate model and by optimizing an acquisition function (Garnett, 2023; Wang et al., 2023; Zhang et al., 2026). BO consists of a surrogate model, which provides a global prediction for the unknown objective function, and an acquisition function that serves as a criterion to determine the next sample to evaluate. In particular, the Gaussian process (GP) model is often preferred as the surrogate model due to its versatility and reliable uncertainty estimation. However, the GP model often suffers from large data sets, making it more suitable for small-budget scenarios (Binois and Wycoff, 2022). To fit a GP model, the dominant cost comes from factorizing the GP observation covariance matrix and solving the associated linear systems, which scales as $\mathcal{O}(n^3)$, where n is the number of data samples. As the sample set grows, the computational burden increases substantially. This limitation poses a significant challenge for scaling BO to real-world problems with large sample sets.

Despite the various approaches to improve the computational efficiency of BO, including parallel BO (González et al., 2016; Daulton et al., 2021; 2020; Eriksson et al., 2019), kernel or local approximation methods (Kim et al., 2021; Jimenez and Katzfuss, 2023; Williams and Seeger, 2000), and sparse or variational GP surrogates (Lawrence et al., 2002; Hensman et al., 2013; Leibfried et al., 2020; McIntire et al., 2016), the computational overhead of repeated GP updates remains a practical bottleneck in large-budget BO (Shahriari et al., 2015). Existing scalable BO approaches typically reduce this cost either by altering the covariance representation or by constructing an alternative sparse posterior, whereas our method instead maintains an explicit retained subset of observations and refits the GP directly on that subset.

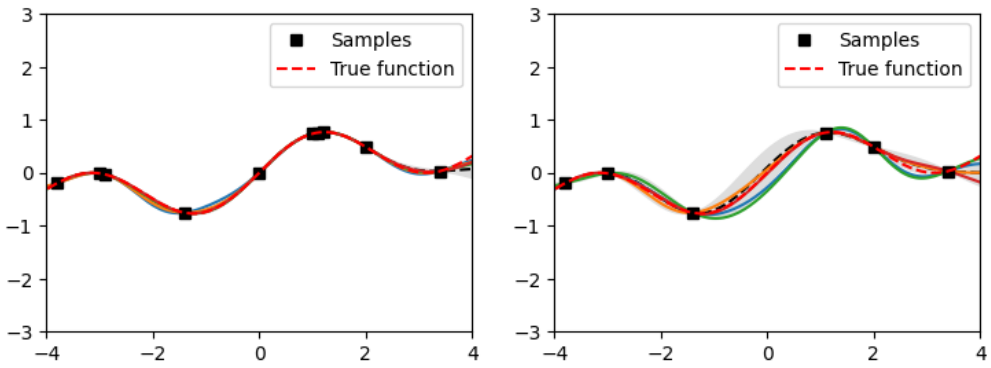

Figure 1: Illustration of GP fitting with sample selection. Left: GP fitted with 10 samples. Right: GP fitted with 6 selected samples. With fewer selected samples, the subset-fitted GP can still provide a useful approximation in this illustrative example and guide subsequent BO evaluations.

During the iterative search process of BO, nearby or redundant observations can become less useful for subsequent model updates as the dataset grows. This motivates subset-maintenance strategies that retain a compact set of observations for GP refitting. Figure 1 gives a toy illustration in which a smaller selected subset can capture the main posterior trend, while the quality of this approximation depends on the retained subset. In this paper, we study a subset-maintenance strategy for BO in which a fixed-size buffer of observations is retained and the GP surrogate is refit only on that subset. Inspired by replay-buffer maintenance ideas from continual learning (Aljundi et al., 2019), we use GP-specific sensitivity embeddings as a heuristic for encouraging diversity and reducing redundancy in the retained buffer. This retained subset is then used to fit the GP model, reducing the cost of GP updates while aiming to retain competitive optimization performance. To the best of our knowledge, this is among the first BO methods to use sensitivity embeddings to maintain a compact observation subset through a cosine-diversity rule and refit the GP only on that subset. We summarize our main contributions as follows:

- **Efficient computations.** We propose Gradient-based Sample Selection Bayesian Optimization (GSSBO), which maintains a fixed-size buffer of observed samples using sensitivity embeddings and a cosine-diversity rule, and refits the GP surrogate only on that subset, thereby reducing the cost of GP updates in large-budget scenarios.
- **Theoretical analysis.** We analyze how subset fitting with a fixed retained subset affects the posterior mean and uncertainty of a GP, and derive approximation-error bounds in terms of subset-dependent residual quantities.
- **Empirical evaluation.** We conduct numerical experiments on synthetic and real-world test problems. These results suggest that the proposed subset-maintenance heuristic can provide a favorable empirical trade-off between BO performance and GP-fitting cost.

2 Related Works

BO with Resource Challenges. In practical applications, BO faces numerous challenges, including high evaluation costs, input-switching costs, resource constraints, and high-dimensional search spaces. Researchers have proposed a variety of methods to address these issues. For instance, parallel BO employs batch sampling to improve efficiency in large-scale or highly concurrent scenarios (González et al., 2016; Daulton et al., 2021; 2020; Eriksson et al., 2019). Kernel approximation methods, such as random Fourier features, map kernels onto lower-dimensional feature spaces, thus accelerating kernel-based approaches (Rahimi and Recht, 2007; Kim et al., 2021). Multi-fidelity BO leverages coarse simulations with a limited number of high-fidelity evaluations to reduce the overall cost (Kandasamy et al., 2016). For high-dimensional tasks, random embeddings or active subspaces help reduce the search dimensionality (Wang et al., 2016; Nayebi et al., 2019). Meanwhile, sparse GP

methods reduce computational complexity through inducing points and variational approximations (Lawrence et al., 2002; Hensman et al., 2013; Leibfried et al., 2020; McIntire et al., 2016; Moss et al., 2023). Canonical variational formulations of sparse GP regression were developed by Titsias (2009), and later work further clarified the relationship between inducing-point variational GPs and Nyström approximations (Wild et al., 2021). Our method is different in that it retains an explicit subset of observations and refits the GP on that subset, rather than constructing an inducing-point posterior approximation. However, these approaches can introduce approximation trade-offs or additional implementation complexity in practical BO settings. Calandriello et al. (2022) scale GP optimization by repeatedly evaluating each selected point until its posterior uncertainty falls below a preset threshold, thus limiting the number of active datasets. However, the dataset still grows with time, and the algorithm's dependence on its initial sample set means that low-value points selected early on remain permanently in the model, potentially inflating computational overhead.

BO with Gradient Information. The availability of derivative information can significantly simplify optimization problems. Ahmed et al. (2016) highlight the potential of incorporating gradient information into BO methods and advocate for its integration into optimization frameworks. Wu and Frazier (2016) introduced the parallel knowledge gradient method for batch BO, achieving faster convergence to global optima. They later introduced d-KG (Wu et al., 2017), a new acquisition function that incorporates derivative observations into BO; in that setting, derivative observations can improve the information value under the assumptions studied there. Rana et al. (2017) incorporated GP priors to enable gradient-based local optimization. Chen et al. (2018) proposed a unified particle-optimization framework using Wasserstein gradient flows for scalable Bayesian sampling. Bilal et al. (2020) demonstrated that BO with gradient-boosted regression trees performed well in cloud configuration tasks. Tamiya and Yamasaki (2022) developed stochastic gradient line BO (SGLBO) for noise-robust quantum circuit optimization. Penubothula et al. (2021) found local critical points by querying where the predicted gradient is zero. Zhang and Rodgers (2024) introduced BO of gradient trajectory (BOGAT) for efficient imaging optimization. Makrygiorgos et al. (2025) integrated exact gradient observations into the Bayesian neural network surrogate's training loss. Although these methods leverage gradient information to improve optimization efficiency and performance, they mainly focus on refining the surrogate model or acquisition rule. In contrast, GSSBO uses sensitivity embeddings only as a retained-subset maintenance heuristic; it does not require derivative observations and keeps the BO acquisition unchanged.

Subset Selection. Subset selection is a key task in fields such as regression, classification, and model selection, aiming to improve efficiency by selecting a subset of features or data. Random subset selection, a simple and widely used method, involves randomly sampling data, often for cross-validation or bootstrap (Hastie, 2009). Importance-based selection focuses on high-value data points, while active learning targets samples that are expected to provide the most information, improving model learning (Quinlan, 1986). Filter methods rank features using statistical measures such as correlation or variance, selecting the top-ranked ones for modeling (Guyon and Elisseeff, 2003). Narendra and Fukunaga (1977) introduced a branch-and-bound algorithm for efficient feature selection. Yang et al. (2022) proposed dataset pruning, an optimization-based sample selection method that identifies the smallest subset of training data to reduce training costs. Ash et al. (2019) employ the k-means++ algorithm in the gradient space for diversity sampling in active learning. Oglic and Gärtner (2017) first map each data point into the reproducing kernel Hilbert space (RKHS), then use a max–min coverage strategy in the RKHS to sequentially sample K landmarks, which are employed to construct the Nyström low-rank approximation. This method is dependent on the quality of the selected landmarks. Hayakawa et al. (2023) provide tighter expected error bounds under a continuous measure for the same underlying idea. However, computing the Mercer decomposition in high dimensions incurs substantial computational cost and suffers from severe error degradation. Because both of these methods perform their approximations in RKHS at high computational expense, they are most suitable for offline, batch-mode resampling. Zhu (2016) proposed a "pilot estimate" to approximate the gradient of the objective function. The core idea is to compute the gradient information corresponding to each data point based on an initial parameter estimate and identify data points with larger gradient values as more "important" samples for subsequent optimization. Despite these advancements, directly applying subset selection methods to BO often yields suboptimal results, necessitating further exploration to integrate subsampling effectively into BO.

3 Preliminaries

3.1 Bayesian Optimization and Gaussian Processes

BO aims to find the global optimum $x^* \in \mathcal{X}$ of an unknown reward function $f : \mathcal{X} \to \mathbb{R}$ over the d-dimensional input space $\mathcal{X} = [0, 1]^d$. Throughout this paper, we consider maximization problems, i.e., we seek $x^* = \arg\max_{x \in \mathcal{X}} f(x)$ as quickly as possible. GPs are one of the fundamental components in BO, providing a theoretical framework for modeling and prediction in a black-box function. In each round, a sample x_t is selected based on the current GP's posterior and acquisition function. The observation (x_t, y_t) is then appended to the accumulated dataset $\mathcal{D}$, and the GP surrogate is updated according to these samples. This iterative process of sampling and updating continues until the optimization objectives are achieved or the available budget is exhausted. The key advantage of GPs lies in their nonparametric nature, allowing them to model complex functions without assuming a specific form. GPs are widely used for regression (Gaussian Process Regression (Schulz et al., 2018), GPR) and classification tasks due to their flexibility and ability to provide uncertainty estimates. Formally, a GP can be defined as: $f(x) \sim \mathcal{GP}(\mu(x), k(x, x'))$, where $\mu(x)$ is the mean function, often assumed to be zero, and $k(x, x')$ is the covariance function, defining the similarity between points x and x'. It should be noted that the algorithmic complexity of GP updates is $\mathcal{O}(n^3)$, where n is the number of observed samples. As the sample set grows, the computational resources required for these updates can become prohibitively expensive, especially in large-scale optimization problems.

3.2 Diversity-based Subset Selection

Due to limited computing resources, properly selecting samples instead of using all samples to fit a model is more efficient in problems with a large sample set. In continual learning, this helps overcome the catastrophic forgetting of previously seen data when faced with online data streams. Suppose that we have a model fitted on observed samples $\mathcal{D} \triangleq \{(x_1, y_1), \ldots, (x_t, y_t)\}$, where $x_i \in \mathcal{X}$ and y_i is the corresponding observation. In the context of subset selection, our goal is to maintain a fixed-size retained subset for subset-fitted GP updates as new observations arrive. Let $\mathbf{g}_t$ denote the sample embedding associated with the sample observed at time t; in GSSBO this will be the sensitivity embedding defined below. Inspired by the gradient-diversity idea of Aljundi et al. (2019), we adopt a diversity-oriented subset-selection surrogate and instantiate it with GP-specific sensitivity embeddings rather than the parameter-gradient constraints used in continual learning. To solve the constraint, we use the geometric properties of these embeddings. Directly optimizing the solid angle subtended by their directions is computationally expensive. Following the gradient-diversity surrogate in (Aljundi et al., 2019), we instead maximize the dispersion of the normalized embedding directions in the fixed-size buffer. This encourages diversity in the embedding space, which we use as a surrogate for reducing redundancy among retained observations. How to determine the buffer size will be detailed in Section 4.3. The previous problem thus becomes a surrogate for selecting a subset $\mathcal{U}$ of the samples that maximizes the diversity of their embeddings:

$$\mathrm{Var}_{i \in \mathcal{U}}\left[\frac{\mathbf{g}_i}{\|\mathbf{g}_i\|}\right] = 1 - \frac{1}{|\mathcal{U}|^2} \sum_{i,j \in \mathcal{U}} \frac{\langle \mathbf{g}_i, \mathbf{g}_j \rangle}{\|\mathbf{g}_i\| \, \|\mathbf{g}_j\|}. \tag{1}$$

In the GSSBO instantiation, $|\mathcal{U}| = M$, where M denotes the buffer size, and $\mathbf{g}_i / \|\mathbf{g}_i\|$ denotes the normalized embedding direction of sample i. The larger the value of this formula, the more dispersed the selected embedding directions are; we therefore interpret it as an embedding-space diversity objective for the retained subset. This empirical surrogate objective is agnostic to how the sample embeddings are computed, making it straightforward to integrate into subset-based methods.

4 Bayesian Optimization with Gradient-based Sample Selection

4.1 Response-Gradient Scores and Sensitivity Embeddings

In the previous section, we introduced a diversity-based subset-selection objective. In the GP setting, we do not assume access to derivatives of the black-box objective with respect to the input x_i. Instead, we

derive a response-gradient score from the GP log marginal likelihood with respect to the observations: $s_i = \frac{\partial}{\partial y_i} \log p(\mathbf{y} \mid \mathbf{X}, \theta)$. This scalar measures how the fitted log-likelihood responds when the i-th response is perturbed, and serves as the starting point for the sensitivity embedding used below for subset maintenance. In a GP model, given a set of samples $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, let $\mathbf{K}$ be the kernel Gram matrix with entries $K_{ij} = k(x_i, x_j; \theta)$, and define the noisy observation covariance $\mathbf{K}_y \triangleq \mathbf{K} + \sigma_n^2 I$, where θ denotes the kernel hyperparameters and σ_n^2 is the observation-noise variance. Assuming that the observed responses follow a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\mathbf{K}_y$, the probability density function is $p(\mathbf{y} \mid \mathbf{X}, \theta) = \frac{1}{(2\pi)^{n/2}|\mathbf{K}_y|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{K}_y^{-1}(\mathbf{y} - \boldsymbol{\mu})\right)$. Taking the logarithm of it, we derive the log-likelihood function:

$$\begin{aligned} \log p(\mathbf{y} \mid \mathbf{X}, \theta) = &-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{K}_y^{-1}(\mathbf{y} - \boldsymbol{\mu}) \\ &- \frac{1}{2}\log|\mathbf{K}_y| - \frac{n}{2}\log(2\pi). \end{aligned} \tag{2}$$

Remark 4.1. The log-likelihood function in equation 2 comprises three terms. The first term, $-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{K}_y^{-1}(\mathbf{y} - \boldsymbol{\mu})$, represents the sample fit under the observation covariance $\mathbf{K}_y$. The second term, $-\frac{1}{2}\log|\mathbf{K}_y|$, penalizes model complexity through the log-determinant of the covariance matrix. The third term, $-\frac{n}{2}\log(2\pi)$, is constant with respect to the parameters and thus does not affect the gradient calculation.

The derivative of the log marginal likelihood with respect to $\mathbf{y}$ describes how the fitted GP model responds to perturbations in the observed responses. The derivatives in this subsection are partial derivatives with the kernel hyperparameters, mean function, and observation-noise variance fixed at their current fitted values; the sensitivity embeddings do not differentiate through hyperparameter re-estimation. Since the second and third terms in equation 2 do not depend on $\mathbf{y}$, we define the response-gradient vector as $\mathbf{s} \triangleq \frac{\partial \log p(\mathbf{y}|\mathbf{X},\theta)}{\partial \mathbf{y}} = -\mathbf{K}_y^{-1}(\mathbf{y} - \boldsymbol{\mu})$. We denote its i-th component by $s_i = -\left(\mathbf{K}_y^{-1}(\mathbf{y} - \boldsymbol{\mu})\right)_i$. Proposition 5.3 later relates the discarded entries of this scalar score vector to the posterior-mean distortion caused by fitting on a retained subset. Because $\mathbf{K}_y^{-1}$ is generally dense, this scalar quantity is evaluated in the context of the full observation covariance of the observed dataset.

Because the scalar score s_i alone is not a vector representation for diversity-based subset selection, we use the sensitivity of the full response-gradient vector to the i-th response:

$$\mathbf{g}_i \triangleq \frac{\partial \mathbf{s}}{\partial y_i} = -\mathbf{K}_y^{-1}\mathbf{e}_i, \tag{3}$$

where $\mathbf{e}_i$ is the i-th standard basis vector. We call $\mathbf{g}_i$ the sensitivity embedding of sample i: it is the sensitivity of the full response-gradient vector to perturbing y_i. Equivalently, $\mathbf{g}_i$ has the precision-column form $-\mathbf{K}_y^{-1}\mathbf{e}_i$, the negative i-th column of the observation precision matrix $\mathbf{K}_y^{-1}$; whereas the scalar score s_i depends on the residual $\mathbf{y} - \boldsymbol{\mu}$ through $\mathbf{K}_y^{-1}(\mathbf{y} - \boldsymbol{\mu})$, the embedding $\mathbf{g}_i$ is determined by the corresponding precision column. We use these embeddings to promote diversity in the retained subset. More generally, when the embedding is computed on a finite ambient set $\mathcal{A}$, we write

$$\mathbf{g}_i^{(\mathcal{A})} = -\left(\mathbf{K}_y^{(\mathcal{A})}\right)^{-1}\mathbf{e}_i^{(\mathcal{A})}, \qquad i \in \mathcal{A}, \tag{4}$$

where $\mathbf{K}_y^{(\mathcal{A})} = \mathbf{K}_{\mathcal{A}\mathcal{A}} + \sigma_n^2 I_{|\mathcal{A}|}$ is the observation covariance restricted to $\mathcal{A}$, and $\mathbf{e}_i^{(\mathcal{A})} \in \mathbb{R}^{|\mathcal{A}|}$. When the ambient set is clear from context, we write $\mathbf{g}_i$. After obtaining the sensitivity embedding for each sample, we can use the directional dispersion of these normalized embedding directions as an alternative diversity objective, i.e. promoting embedding-direction diversity in the retained subset. Constructing these embeddings requires access to the candidate observations' inverse observation-covariance structure and introduces additional overhead beyond the subset-GP refit. We account for this overhead separately in the complexity summary below.

4.2 Gradient-based Sample Selection

As the number of observed samples increases, fitting a GP model can become prohibitively expensive, especially in large-scale scenarios. A common remedy is to work with a subset of samples of size $M \ll N$, thereby

reducing the computational cost of GP updates. The efficiency and effectiveness of GP model fitting in BO are closely related to the quality of the chosen subset. This raises the question: *How do we choose a compact retained subset that is useful for subset-fitted GP updates?* Inspired by the success of gradient-based subset selection methods in machine learning, we use GP response-sensitivity embeddings to guide the maintenance of such subsets within BO. To this end, we introduce a sensitivity-embedding-based subset-maintenance methodology that aims to improve retained-subset coverage within a limited sample buffer size. By using sensitivity embeddings, the approach seeks to reduce computational burden while empirically maintaining surrogate quality as the sample set grows. We begin by modeling the objective function f with a GP and setting a buffer size M. Initially, the algorithm observes f at n_0 samples and uses all initial observations before subset maintenance is activated. After each subsequent evaluation, if the number of samples exceeds M, we perform a sensitivity-embedding-based subset selection step to keep a subset of size M for the next GP update.

4.3 Gradient-based Sample Selection BO

The following outlines the GSSBO implementation details and practical considerations. *We highlight the key insight of this subsection: we tackle the scalability of BO by maintaining a compact subset selected by the proposed subset-maintenance heuristic.*

Algorithmic Overview. Algorithm 1 summarizes the online subset-maintenance procedure. In each iteration, GSSBO selects a new point using the current GP posterior and acquisition function, evaluates the corresponding observation, and updates the dataset with the new sample. If $|\mathcal{D}| \leq M$, the GP is updated using all observed samples. Once the buffer limit is exceeded, the method switches to subset maintenance: the newly acquired sample (x_t, y_t) is retained, the remaining $M-1$ samples are selected from a selection pool $\mathcal{P}_t \subseteq \mathcal{D}_t$ according to the cosine-diversity rule over normalized sensitivity embeddings in Eq. 1, and the GP surrogate is refit on the resulting subset $\mathcal{U}_t$. For a finite ambient set $\mathcal{A}$, the sensitivity embedding in Eq. 3 is computed using the observation covariance restricted to $\mathcal{A}$; taking $\mathcal{A} = \mathcal{P}_t$ gives the pool-specific embeddings used by the selection step, while $\mathcal{A} = \mathcal{D}_t$ recovers the full-history embedding definition.

(1) Dynamic Buffer Size. In practice, a suitable buffer size may be difficult to choose in advance. We therefore use a simple runtime-triggered rule to initialize M. Let $\bar{T}$ be the average wall-clock time of the initial iterations and T_{current} the current iteration time. When T_{current} first exceeds the user-specified threshold $Z \times \bar{T}$, we set $M = |\mathcal{D}|$ at that iteration and keep this value fixed thereafter. In this way, the transition itself does not discard previously collected observations, while subsequent iterations proceed with subset maintenance under the fixed buffer size. This rule provides a simple way to balance computational cost and retained-data coverage without requiring M to be chosen a priori.

(2) Retaining Latest Observations. During the procedure, the newly acquired sample, (x_t, y_t), is always included in the subset. This makes the most recent observation available to the next subset-fitted GP update, which can help the retained subset track newly explored regions. Additionally, this design addresses a practical limitation of fixed-size representations in BO (McIntire et al., 2016): a constrained representation size may otherwise delay the use of new observations in the model. It may also reduce repeated refitting on subsets concentrated around the same current best region, since the most recent observation is always retained in the next iteration.

Here we highlight the difference between our method and SparseGP. SparseGP methods introduce inducing-point approximations together with additional posterior-correction mechanisms. In contrast, our method maintains a compact subset of observed samples via the subset-maintenance heuristic and refits the GP only on that retained subset. The computational gain in GSSBO therefore comes from subset-of-data GP updates rather than from inducing-point posterior corrections.

Algorithm 1 Gradient-based Sample Selection BO

1: **Initialization:** Obtain n_0 initial samples $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n_0}$, and fit an initial GP model. Set buffer size $M > n_0$, total budget N, average initial iteration time $\bar{T}$, and threshold factor Z. Initialize `switched` $\leftarrow$ `false`.
2: **for** $t = n_0 + 1$ to N **do**

Select $x_t = \arg\max_x \ \alpha\big(x; p_t(f)\big)$, where $p_t(f)$ is the current fitted GP posterior and α is the acquisition function.
Evaluate $y_t = f(x_t)$ and set $\mathcal{D} \leftarrow \mathcal{D} \cup \{(x_t, y_t)\}$.
if not switched **then**
  Let T_{current} be the current iteration time.
  if $T_{\text{current}} > Z \times \bar{T}$ **then**
    Set `switched` $\leftarrow$ `true`, $\quad M \leftarrow |\mathcal{D}|$.
  end if
end if
if switched **then**
  Choose a selection pool $\mathcal{P}_t \subseteq \mathcal{D}_t$ with $(x_t, y_t) \in \mathcal{P}_t$, $|\mathcal{P}_t| \geq M$, and $q_t = |\mathcal{P}_t|$.
  Compute or update sensitivity embeddings $\mathbf{g}_i^{(\mathcal{P}_t)}$ for samples in $\mathcal{P}_t$.
  Form subset $\mathcal{U}_t$ by forcing in (x_t, y_t).
  Select the remaining $M - 1$ samples from $\mathcal{P}_t \setminus \{(x_t, y_t)\}$ according to the cosine-diversity rule in Eq. 1, using these pool-specific embeddings.
  Update the subset-fitted GP using the M samples in $\mathcal{U}_t$.
else
  Update the GP using all samples in $\mathcal{D}$.
end if
end for

The dominant computational trade-off in GSSBO is between refitting the GP on a retained subset of size M and maintaining that subset over time. At iteration t, let $n_t = |\mathcal{D}_t|$ be the number of accumulated observations and let $q_t = |\mathcal{P}_t|$ be the size of the selection pool. In standard GP-UCB, the full-data GP refit/factorization cost is $\mathcal{O}(n_t^3)$. Once subset maintenance is activated, the subset-fitted GP update in GSSBO uses M retained observations, so the GP refit/factorization term is $\mathcal{O}(M^3)$ for fixed hyper-parameters. This term should be read as the subset-GP refitting cost after the switch, not as a complete end-to-end per-iteration bound for GSSBO. The total update also includes the cost of refreshing the sensitivity embeddings, $C_{\text{emb}}(n_t, q_t)$, and the cost of the cosine-diversity subset-maintenance step, $C_{\text{div}}(q_t, M)$. Thus, compared with the $\mathcal{O}(n_t^3)$ refit/factorization term of full GP-UCB, GSSBO caps the subset-GP refit term at $\mathcal{O}(M^3)$ after the switch while making the additional subset-maintenance overhead explicit. RSSBO has the same subset-GP refit term but uses random subset maintenance instead of the sensitivity-embedding and cosine-diversity step.

The terms C_{emb} and C_{div} depend on the embedding-maintenance and cosine-diversity implementation. If exact full-data precision-column embeddings are recomputed over all accumulated observations, then $q_t = n_t$ and the precision-structure construction is dominated by $\mathcal{O}(n_t^3)$ factorization. If the observation precision matrix is maintained under fixed hyper-parameters using a rank-one block update, the refresh can be $\mathcal{O}(n_t^2)$ per new observation with $\mathcal{O}(n_t^2)$ memory; when hyper-parameters are refit, this cache may need to be rebuilt. Given cached normalized embeddings or cached pairwise cosine scores, a greedy cosine-diversity implementation costs $\mathcal{O}(q_t M)$; without cached similarities, forming all pairwise cosine scores costs $\mathcal{O}(d_{g,t} q_t^2)$, where $d_{g,t}$ is the embedding dimension for that iteration. For pool-specific precision-column embeddings, $d_{g,t} = q_t$, so naive pairwise cosine construction costs $\mathcal{O}(q_t^3)$; for full-history embeddings evaluated on q_t pool candidates, $d_{g,t} = n_t$. Restricted-pool implementations correspond to smaller q_t, while the full-history case is obtained by taking $\mathcal{P}_t = \mathcal{D}_t$.

5 Theoretical Analysis

We analyze the posterior distortion caused by fitting the GP on a retained subset $\mathcal{U}$ rather than on the full dataset $\mathcal{D}$. Throughout this section and the appendix, vectors and matrices are not boldfaced; vector norms are Euclidean and matrix norms are operator norms. We use the standard noisy GP model with observation-noise variance $\sigma_n^2 > 0$, write the results in zero-mean form, and compare full-data and subset-fitted posteriors under shared kernel hyper-parameters and observation-noise variance. The results are conditional diagnostics for a fixed retained subset, not unconditional guarantees for the cosine-diversity selection rule.

Fix a partition $\mathcal{D} = \mathcal{U} \cup \mathcal{R}$, where $\mathcal{U}$ is retained and $\mathcal{R}$ is discarded. Define $A = K_{\mathcal{U}\mathcal{U}} + \sigma_n^2 I$, $B = K_{\mathcal{U}\mathcal{R}}$, $C = K_{\mathcal{R}\mathcal{R}} + \sigma_n^2 I$, and $S = C - B^\top A^{-1} B$. For a test point x, let $k_{\mathcal{U}}(x) = K_{\mathcal{U}x}$, $k_{\mathcal{R}}(x) = K_{\mathcal{R}x}$, $r_{\mathcal{U}} = y_{\mathcal{R}} - B^\top A^{-1} y_{\mathcal{U}}$, and $c_{\mathcal{U}}(x) = k_{\mathcal{R}}(x) - B^\top A^{-1} k_{\mathcal{U}}(x)$. Let $(\mu_{\mathcal{D}}, \sigma_{\mathcal{D}}^2)$ and $(\mu_{\mathcal{U}}, \sigma_{\mathcal{U}}^2)$ be the full-data and retained-subset posteriors. Here $m = |\mathcal{U}|$ and $r = |\mathcal{R}|$, so $A \in \mathbb{R}^{m\times m}$, $B \in \mathbb{R}^{m\times r}$, $C, S \in \mathbb{R}^{r\times r}$, $k_{\mathcal{U}}(x) \in \mathbb{R}^m$, and $k_{\mathcal{R}}(x), c_{\mathcal{U}}(x) \in \mathbb{R}^r$.

Posterior gap. The following theorem gives exact posterior-gap identities for this fixed partition.

Theorem 5.1 (Posterior gap for a fixed subset-fitted GP under shared hyper-parameters)**.** *With the notation above, for every x, $\mu_{\mathcal{D}}(x) - \mu_{\mathcal{U}}(x) = c_{\mathcal{U}}(x)^\top S^{-1} r_{\mathcal{U}}$ and $\sigma_{\mathcal{U}}^2(x) - \sigma_{\mathcal{D}}^2(x) = c_{\mathcal{U}}(x)^\top S^{-1} c_{\mathcal{U}}(x) \geq 0$. Consequently, $|\mu_{\mathcal{D}}(x) - \mu_{\mathcal{U}}(x)| \leq \|c_{\mathcal{U}}(x)\| \, \|S^{-1}\| \, \|r_{\mathcal{U}}\|$ and $0 \leq \sigma_{\mathcal{U}}^2(x) - \sigma_{\mathcal{D}}^2(x) \leq \|c_{\mathcal{U}}(x)\|^2 \, \|S^{-1}\|$. Since $S \succeq \sigma_n^2 I$, the coarser bounds $|\mu_{\mathcal{D}}(x) - \mu_{\mathcal{U}}(x)| \leq \|c_{\mathcal{U}}(x)\| \, \|r_{\mathcal{U}}\| / \sigma_n^2$ and $\sigma_{\mathcal{U}}^2(x) - \sigma_{\mathcal{D}}^2(x) \leq \|c_{\mathcal{U}}(x)\|^2 / \sigma_n^2$ also hold.*

Remark 5.2 (Interpretation). Theorem 5.1 shows that posterior distortion is governed by the discarded-response residual $r_{\mathcal{U}}$ and the conditional-correlation residual $c_{\mathcal{U}}(x)$. Small residuals imply a small subset/full posterior gap, but the theorem does not prove that the GSSBO cosine-diversity rule always makes these residuals small.

The next proposition connects the scalar response-gradient score used in GSSBO's diagnostics with the posterior-mean distortion above.

Proposition 5.3 (Residual bridge for scalar response scores)**.** *Under the assumptions and notation of Theorem 5.1, let $z_{\mathcal{D}}$ denote the response vector centered by the chosen GP mean function. Under the zero-mean notation of the theorem, $z_{\mathcal{D}} = y_{\mathcal{D}}$. Define $s = -(K_{\mathcal{D}\mathcal{D}} + \sigma_n^2 I)^{-1} z_{\mathcal{D}}$, and let $s_{\mathcal{R}}$ be the entries indexed by $\mathcal{R}$. Then $s_{\mathcal{R}} = -S^{-1} r_{\mathcal{U}}$ and $r_{\mathcal{U}} = -S s_{\mathcal{R}}$. Consequently, for every x, $\mu_{\mathcal{D}}(x) - \mu_{\mathcal{U}}(x) = -c_{\mathcal{U}}(x)^\top s_{\mathcal{R}}$, and $|\mu_{\mathcal{D}}(x) - \mu_{\mathcal{U}}(x)| \leq \|c_{\mathcal{U}}(x)\| \, \|s_{\mathcal{R}}\|$. If $\rho_{\mathcal{U}} := \sup_{x\in\mathcal{X}} \|c_{\mathcal{U}}(x)\| < \infty$, then $\sup_{x\in\mathcal{X}} |\mu_{\mathcal{D}}(x) - \mu_{\mathcal{U}}(x)| \leq \rho_{\mathcal{U}} \|s_{\mathcal{R}}\|$.*

This proposition gives the scalar response scores a posterior-distortion interpretation: the discarded full-data scores determine the posterior-mean distortion of a fixed retained subset through the conditional-correlation residual $c_{\mathcal{U}}(x)$.

For acquisition stability, define the full-data and subset-fitted UCB values with the same $\beta_t \geq 0$: $\alpha_{\mathcal{D}}^{\mathrm{UCB}}(x) = \mu_{\mathcal{D}}(x) + \sqrt{\beta_t}\, \sigma_{\mathcal{D}}(x)$ and $\alpha_{\mathcal{U}}^{\mathrm{UCB}}(x) = \mu_{\mathcal{U}}(x) + \sqrt{\beta_t}\, \sigma_{\mathcal{U}}(x)$, where $\sigma_{\mathcal{D}}(x) = \sqrt{\sigma_{\mathcal{D}}^2(x)}$ and $\sigma_{\mathcal{U}}(x) = \sqrt{\sigma_{\mathcal{U}}^2(x)}$.

Corollary 5.4 (UCB acquisition stability for a fixed retained subset)**.** *Under the assumptions of Theorem 5.1, for every $x \in \mathcal{X}$, $|\alpha_{\mathcal{D}}^{\mathrm{UCB}}(x) - \alpha_{\mathcal{U}}^{\mathrm{UCB}}(x)| \leq \|c_{\mathcal{U}}(x)\| \, \|s_{\mathcal{R}}\| + \sqrt{\beta_t}\, \|c_{\mathcal{U}}(x)\| \sqrt{\|S^{-1}\|}$. If $\rho_{\mathcal{U}} := \sup_{x\in\mathcal{X}} \|c_{\mathcal{U}}(x)\| < \infty$, then $\sup_{x\in\mathcal{X}} |\alpha_{\mathcal{D}}^{\mathrm{UCB}}(x) - \alpha_{\mathcal{U}}^{\mathrm{UCB}}(x)| \leq \rho_{\mathcal{U}} \|s_{\mathcal{R}}\| + \sqrt{\beta_t} \rho_{\mathcal{U}} \sqrt{\|S^{-1}\|}$.*

Remark. Theorem 5.1, Proposition 5.3, and Corollary 5.4 are conditional diagnostic and stability statements for a fixed retained subset under shared hyper-parameters and observation-noise variance. They do not prove that the GSSBO selection rule necessarily makes $\|s_{\mathcal{R}}\|$, $\rho_{\mathcal{U}}$, $\|r_{\mathcal{U}}\|$, or $\|S^{-1}\|$ small, nor do they provide cumulative-regret, argmax-equivalence, or unconditional optimality guarantees. The experiments evaluate downstream optimization performance, runtime, and an auxiliary RMSE diagnostic for the retained-subset GP fit; direct measurement of the residual and acquisition-level quantities in the bounds is left to future work. Full proofs are deferred to the appendix.

6 Experiments

In this section, we conduct numerical experiments to evaluate the practical efficiency of GSSBO. The objective of the numerical experiments is threefold: (1) to evaluate computational efficiency; (2) to assess optimization performance; and (3) to report an auxiliary RMSE diagnostic for the retained-subset GP fit. To assess the performance of our proposed methods, we test five benchmark functions (Eggholder2, Hart6, Levy20, Powell50, Rastrigin100) and a real-world diabetes hyperparameter optimization task.

Experimental Setup. We choose UCB as the acquisition function in GSSBO, and compare GSSBO with the following benchmarks: *(1) Standard GP-UCB* (Srinivas et al., 2009); *(2) Random Sample Selection GP-UCB (RSSBO, our ablation)*, which mirrors our approach in restricting the sample set size but chooses which samples randomly; *(3) VecchiaBO* (Jimenez and Katzfuss, 2023), which utilizes the Vecchia approximation

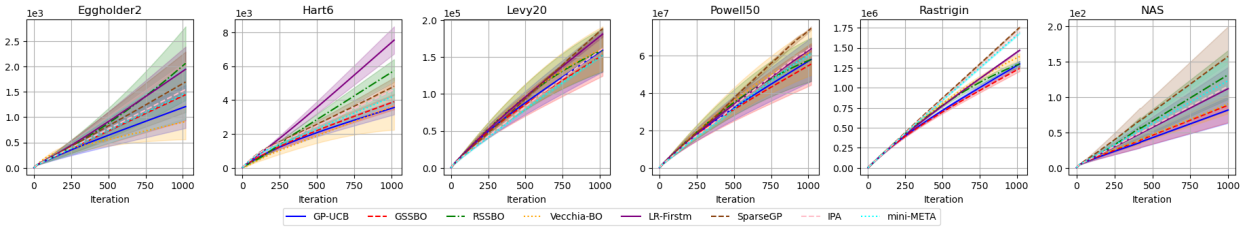

Figure 2: Cumulative regret of algorithms on synthetic benchmarks and the diabetes hyperparameter optimization task. Solid lines denote the mean over 50 runs; shaded regions indicate ±1 standard deviation.

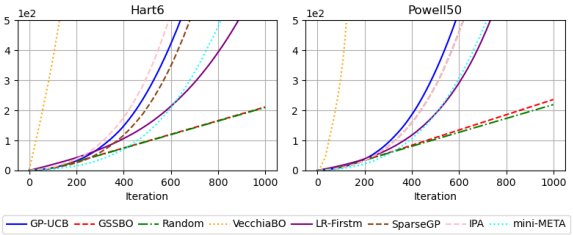

Figure 3: Cumulative time cost of algorithms.

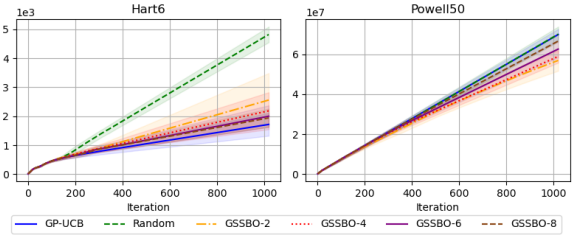

Figure 4: Sensitivity analysis of Z.

method in BO; *(4) LR-First m* (Williams and Seeger, 2000), which uses a low-rank approximation based on the first m samples. *(5) SparseGP* (Lawrence et al., 2002), which introduces a small number of inducing points to obtain a sparse approximation of GP; *(6) IPA* (Moss et al., 2023), which multiplies an expected-improvement-based quality function by the GP posterior and selects inducing points regarding quality-diversity; and *(7) mini-META* (Calandriello et al., 2022), which scales GP optimization by repeatedly evaluating each selected point until its posterior uncertainty falls below a preset threshold. (Additional baselines and high-dimensional tasks are reported in the appendix.) We employ a Matérn 5/2 kernel for the GP, with hyperparameters learned via maximum likelihood estimation. Both GSSBO and RSSBO use the same buffer size M, dynamically adjusted by a parameter $Z = 4$. M is usually around 100 and will change with Z. The size of the initial set is 20. Each experiment is repeated 50 times, and the total number of iterations is 1000. All experiments were conducted on a MacBook Pro with Apple M2 Pro (10-core CPU, 16 GB unified RAM). For maximization tasks, we report cumulative regret over a horizon T (with $T = N$ in the experiments below, where N is the total budget in Algorithm 1): $R_T = \sum_{t=1}^{T} \big(f(x^\star) - f(x_t)\big)$. For the real-world hyperparameter optimization task, where the raw objective is a validation classification error to be minimized, we equivalently maximize its negative so that the same regret convention applies.

6.1 Synthetic Test Problems

Computational Efficiency Analysis. Figure 3 compares the cumulative runtime (in seconds) over 1000 iterations on low-dimensional Hart6 and high-dimensional Powell50 (results for other test functions are provided in the appendix due to space constraints). In both plots, VecchiaBO incurs a rapidly accelerating runtime, whereas GSSBO and RSSBO remain notably lower than GP-UCB. While VecchiaBO reduces the cost of GP fitting by conditioning on nearest neighbors, its runtime is dominated by the costly maintenance of a structured neighbor graph, which scales poorly with dimensionality and sample size. The running speeds of LR-First, SparseGP, and mini-META are all improved compared to the standard GP-UCB. IPA has no obvious advantage in terms of time consumption because it needs to calculate the sample quality. In contrast, GSSBO and RSSBO have much lower runtime in these experiments. Once the active subset is restricted to size $M \ll n$, the subset-GP refit component in GSSBO is capped at $\mathcal{O}(M^3)$, while the total runtime also includes the implemented sensitivity-embedding construction and cosine-diversity subset-selection overhead described above. GSSBO and RSSBO are often similar in runtime, though GSSBO can be slightly higher due to the overhead of the subset-maintenance heuristic. By iteration 1000, the cumulative runtime of GSSBO on both Hart6 and Powell50 is about 10% of that of standard GP-UCB. This empirical comparison suggests

that GP refitting is a major computational burden in the large-budget setting considered here. Under this setup, the runtime gap may become larger as n increases.

Optimization Performance Analysis. Figure 2 compares methods on multiple functions, evaluating cumulative regret. Overall, GSSBO achieves cumulative regret comparable to standard GP-UCB and lower than RSSBO in most reported settings. In low-dimensional problems such as Eggholder2 and Hart6, GSSBO has a small gap with standard GP-UCB and VecchiaBO while remaining competitive with the other baselines in Figure 2. In high-dimensional settings, such as on Levy20, Powell50, and Rastrigin100, GSSBO obtains among the lowest mean cumulative regrets among the compared methods in Figure 2. From the experimental results, the cumulative regret of GSSBO remains competitive over the tested horizon. In particular, GSSBO achieves these results with a substantial reduction in computation time in our experiments, as shown in Figure 3. These results show a favorable empirical trade-off between regret and runtime in the reported settings. A plausible explanation for the stronger performance of GP-UCB in Figure 2 is that it uses full-data GP updates, whereas the other baselines rely on approximation or subset-restriction mechanisms that may reduce surrogate fidelity in some tasks.

Sensitivity analysis of Hyperparameter Z. We further examine how the dynamic buffer parameter Z affects GSSBO. Figure 4 presents results on two functions: Hart6 and Powell50. For RSSBO, Z remains fixed at 4, whereas for GSSBO, we vary $Z \in \{2, 4, 6, 8\}$. On Hart6, larger Z tends to move GSSBO closer to standard GP-UCB, while RSSBO has higher cumulative regret in these runs. In contrast, in Powell50, smaller Z gives slightly better GSSBO performance in this experiment, which is consistent with a possible benefit of more aggressive subset updates in this setting. In these two cases, Hart6 benefits from a larger Z, whereas Powell50 performs slightly better with more aggressive subset limiting. This suggests that the preferred Z may depend on the problem structure.

6.2 Real-World Application

To assess the applicability of GSSBO beyond analytic test functions, we also consider hyperparameter optimization for a diabetes-detection model built from the UCI diabetes dataset (Dua and Graff, 2017). In this task, each BO query (x_t, y_t) corresponds to training the model with a proposed hyperparameter configuration and observing its validation classification error. The search space contains four hyperparameters: batch size in $[32, 128]$, learning rate in $[10^{-6}, 1.0]$, learning-rate decay in $[10^{-6}, 1.0]$, and hidden dimension in $[1, 8]$. Since the raw objective is an error to be minimized, we optimize its negative when applying the common maximization-based regret convention described above. This experiment tests whether the subset-maintenance strategy remains useful when the objective evaluations are generated by a learning pipeline rather than by a closed-form synthetic benchmark. The results in Figure 2 show that GSSBO performs favorably relative to the compared baselines on this task while using the same subset-maintenance protocol as in the synthetic experiments.

7 Conclusion

BO is known to be effective for optimization in settings where the objective function is expensive to evaluate. In large-budget scenarios, the use of a full GP model can slow the convergence of BO, leading to poor scaling in these cases. In this paper, we investigated the use of sensitivity-embedding-based subset maintenance to accelerate BO. The results show empirically that the GSSBO-selected retained subset can often support BO performance close to full-data GP-UCB while substantially reducing GP fitting cost in the reported settings. Across the reported synthetic and real-world benchmarks, GSSBO reduces runtime while retaining competitive optimization performance. The sensitivity analysis suggests that the preferred Z can depend on the problem setting. Overall, these findings support response-sensitivity-based subset maintenance as a practical route for addressing BO scaling challenges. Our theoretical results should be read as conditional posterior- and acquisition-stability diagnostics for fixed retained subsets; the empirical results evaluate how the proposed selection rule behaves in the reported settings. Future work includes designing subset-selection rules that more directly control the residual quantities appearing in the posterior-gap bounds.

References

M. O. Ahmed, B. Shahriari, and M. Schmidt. Do we need "harmless" bayesian optimization and "first-order" bayesian optimization. *NIPS BayesOpt*, 5:21, 2016.

R. Aljundi, M. Lin, B. Goujaud, and Y. Bengio. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32, 2019.

J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

M. Bilal, M. Serafini, M. Canini, and R. Rodrigues. Do the best cloud configurations grow on trees? an experimental evaluation of black box algorithms for optimizing cloud workloads. *Proceedings of the VLDB Endowment*, 13(12):2563–2575, 2020.

M. Binois and N. Wycoff. A survey on high-dimensional gaussian process modeling with application to bayesian optimization. *ACM Transactions on Evolutionary Learning and Optimization*, 2(2):1–26, 2022.

J. Branke, S. Rojas-Gonzalez, and I. Couckuyt. Bayesian optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 772–790, 2024.

D. Calandriello, L. Carratino, A. Lazaric, M. Valko, and L. Rosasco. Scaling gaussian process optimization by evaluating a few unique candidates multiple times. In *International Conference on Machine Learning*, pages 2523–2541. PMLR, 2022.

C. Chen, R. Zhang, W. Wang, B. Li, and L. Chen. A unified particle-optimization framework for scalable bayesian sampling. *arXiv preprint arXiv:1805.11659*, 2018.

I. Couckuyt, S. R. Gonzalez, and J. Branke. Bayesian optimization: tutorial. In *Proceedings of the genetic and evolutionary computation conference companion*, pages 843–863, 2022.

S. Daulton, M. Balandat, and E. Bakshy. Differentiable expected hypervolume improvement for parallel multi-objective bayesian optimization. *Advances in Neural Information Processing Systems*, 33:9851–9864, 2020.

S. Daulton, M. Balandat, and E. Bakshy. Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement. *Advances in Neural Information Processing Systems*, 34:2187–2200, 2021.

D. Dua and C. Graff. Uci machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

D. Eriksson, M. Pearce, J. Gardner, R. D. Turner, and M. Poloczek. Scalable global optimization via local bayesian optimization. *Advances in neural information processing systems*, 32, 2019.

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

R. Garnett. *Bayesian optimization*. Cambridge University Press, 2023.

J. González, Z. Dai, P. Hennig, and N. Lawrence. Batch bayesian optimization via local penalization. In *Artificial intelligence and statistics*, pages 648–657. PMLR, 2016.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

T. Hastie. The elements of statistical learning: Data mining, inference, and prediction, 2009.

S. Hayakawa, H. Oberhauser, and T. Lyons. Sampling-based nyström approximation and kernel quadrature. In *International Conference on Machine Learning*, pages 12678–12699. PMLR, 2023.

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

F. Jimenez and M. Katzfuss. Scalable bayesian optimization using vecchia approximations of gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pages 1492–1512. PMLR, 2023.

K. Kandasamy, G. Dasarathy, J. B. Oliva, J. Schneider, and B. Póczos. Gaussian process bandit optimisation with multi-fidelity evaluations. *Advances in neural information processing systems*, 29, 2016.

J. Kim, M. McCourt, T. You, S. Kim, and S. Choi. Bayesian optimization with approximate set kernels. *Machine Learning*, 110:857–879, 2021.

N. Lawrence, M. Seeger, and R. Herbrich. Fast sparse gaussian process methods: The informative vector machine. *Advances in neural information processing systems*, 15, 2002.

F. Leibfried, V. Dutordoir, S. John, and N. Durrande. A tutorial on sparse gaussian processes and variational inference. *arXiv preprint arXiv:2012.13962*, 2020.

H. Luo, J. W. Demmel, Y. Cho, X. S. Li, and Y. Liu. Non-smooth bayesian optimization in tuning problems. *arXiv preprint arXiv:2109.07563*, 2021.

H. Luo, Y. Cho, J. W. Demmel, X. S. Li, and Y. Liu. Hybrid parameter search and dynamic model selection for mixed-variable bayesian optimization. *Journal of Computational and Graphical Statistics*, 33(3):855–868, 2024.

G. Makrygiorgos, J. H. S. Ip, and A. Mesbah. Towards scalable bayesian optimization via gradient-informed bayesian neural networks. *arXiv preprint arXiv:2504.10076*, 2025.

M. McIntire, D. Ratner, and S. Ermon. Sparse gaussian processes for bayesian optimization. In *UAI*, volume 3, page 4, 2016.

H. B. Moss, S. W. Ober, and V. Picheny. Inducing point allocation for sparse gaussian processes in high-throughput bayesian optimisation. In *International Conference on Artificial Intelligence and Statistics*, pages 5213–5230. PMLR, 2023.

V. Mullachery, A. Khera, and A. Husain. Bayesian neural networks. *arXiv preprint arXiv:1801.07710*, 2018.

Narendra and Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Transactions on computers*, 100(9):917–922, 1977.

A. Nayebi, A. Munteanu, and M. Poloczek. A framework for bayesian optimization in embedded subspaces. In *International Conference on Machine Learning*, pages 4752–4761. PMLR, 2019.

D. Oglic and T. Gärtner. Nyström method with kernel k-means++ samples as landmarks. In *International Conference on Machine Learning*, pages 2652–2660. PMLR, 2017.

S. Penubothula, C. Kamanchi, and S. Bhatnagar. Novel first order bayesian optimization with an application to reinforcement learning. *Applied Intelligence*, 51(3):1565–1579, 2021.

J. R. Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

S. Rana, C. Li, S. Gupta, V. Nguyen, and S. Venkatesh. High dimensional bayesian optimization with elastic gaussian process. In *International conference on machine learning*, pages 2883–2891. PMLR, 2017.

E. Schulz, M. Speekenbrink, and A. Krause. A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of mathematical psychology*, 85:1–16, 2018.

B. Shahriari, K. Swersky, Z. Wang, R. P. 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, 2015.

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

S. Tamiya and H. Yamasaki. Stochastic gradient line bayesian optimization for efficient noise-robust optimization of parameterized quantum circuits. *npj Quantum Information*, 8(1):90, 2022.

M. Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574, 2009.

S. Wang and S. H. Ng. Partition-based bayesian optimization for stochastic simulations. In *2020 Winter Simulation Conference (WSC)*, pages 2832–2843. IEEE, 2020.

X. Wang, Y. Jin, S. Schmitt, and M. Olhofer. Recent advances in bayesian optimization. *ACM Computing Surveys*, 55(13s):1–36, 2023.

Z. Wang, F. Hutter, M. Zoghi, D. Matheson, and N. De Feitas. Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research*, 55:361–387, 2016.

V. Wild, M. Kanagawa, and D. Sejdinovic. Connections and equivalences between the nyström method and sparse variational gaussian processes. *CoRR*, abs/2106.01121, 2021. URL https://arxiv.org/abs/2106.01121.

C. Williams and M. Seeger. Using the nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13, 2000.

A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing. Deep kernel learning. In *Artificial intelligence and statistics*, pages 370–378. PMLR, 2016.

J. Wu and P. Frazier. The parallel knowledge gradient method for batch bayesian optimization. *Advances in neural information processing systems*, 29, 2016.

J. Wu, M. Poloczek, A. G. Wilson, and P. Frazier. Bayesian optimization with gradients. *Advances in neural information processing systems*, 30, 2017.

S. Yang, Z. Xie, H. Peng, M. Xu, M. Sun, and P. Li. Dataset pruning: Reducing training data by examining generalization influence. *arXiv preprint arXiv:2205.09329*, 2022.

M. Zhang and C. T. Rodgers. Bayesian optimization of gradient trajectory for parallel-transmit pulse design. *Magnetic Resonance in Medicine*, 91(6):2358–2373, 2024.

Z. Zhang, T. Ren, and Y. Peng. Nonparametric bayesian optimization for general rewards. *arXiv preprint arXiv:2602.07411*, 2026.

R. Zhu. Gradient-based sampling: An adaptive importance sampling for least-squares. *Advances in neural information processing systems*, 29, 2016.

A Appendix Theoretical Analysis

A.1 Posterior gap for subset-fitted GP under shared hyper-parameters

In this appendix, we analyze the effect of subset-of-data fitting itself. Throughout this section, the full-data GP posterior and the subset-fitted GP posterior are compared under the same kernel hyper-parameters and the same observation-noise variance. This isolates the effect of subset fitting and does not account for the additional error introduced by re-estimating hyper-parameters on the retained subset.

We adopt the standard noisy GP regression model and state the derivation in the zero-mean form; equivalently, the same proof applies after centering the observations with respect to the chosen GP mean function.

Unless otherwise stated, vector norms are Euclidean norms and matrix norms are operator norms.

Theorem 5.1 is the main posterior-gap result used in the remainder of this appendix.

A.2 Proof of Theorem 5.1

We first introduce the block notation used in the proof. The noisy observations are partitioned according to

$$y_{\mathcal{D}} = \begin{bmatrix} y_{\mathcal{U}} \\ y_{\mathcal{R}} \end{bmatrix}, \qquad k_{\mathcal{D}}(x) = \begin{bmatrix} k_{\mathcal{U}}(x) \\ k_{\mathcal{R}}(x) \end{bmatrix}.$$

The full observation covariance matrix is

$$K_{\mathcal{D}\mathcal{D}} + \sigma_n^2 I = \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix},$$

where

$$A = K_{\mathcal{U}\mathcal{U}} + \sigma_n^2 I, \qquad B = K_{\mathcal{U}\mathcal{R}}, \qquad C = K_{\mathcal{R}\mathcal{R}} + \sigma_n^2 I.$$

Since $K_{\mathcal{D}\mathcal{D}} \succeq 0$ and $\sigma_n^2 > 0$, we have $A \succ 0$. The Schur complement is

$$S = C - B^\top A^{-1} B.$$

The subset-fitted posterior mean and variance are

$$\mu_{\mathcal{U}}(x) = k_{\mathcal{U}}(x)^\top A^{-1} y_{\mathcal{U}},$$
$$\sigma_{\mathcal{U}}^2(x) = k(x,x) - k_{\mathcal{U}}(x)^\top A^{-1} k_{\mathcal{U}}(x).$$

The full-data posterior mean and variance are

$$\mu_{\mathcal{D}}(x) = k_{\mathcal{D}}(x)^\top \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}^{-1} y_{\mathcal{D}},$$

$$\sigma_{\mathcal{D}}^2(x) = k(x,x) - k_{\mathcal{D}}(x)^\top \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}^{-1} k_{\mathcal{D}}(x).$$

By the block inverse formula,

$$\begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}BS^{-1}B^\top A^{-1} & -A^{-1}BS^{-1} \\ -S^{-1}B^\top A^{-1} & S^{-1} \end{bmatrix}.$$

We now derive the posterior mean gap. Expanding $\mu_{\mathcal{D}}(x)$,

$$\begin{aligned} \mu_{\mathcal{D}}(x) &= \begin{bmatrix} k_{\mathcal{U}}(x) \\ k_{\mathcal{R}}(x) \end{bmatrix}^\top \begin{bmatrix} A^{-1} + A^{-1}BS^{-1}B^\top A^{-1} & -A^{-1}BS^{-1} \\ -S^{-1}B^\top A^{-1} & S^{-1} \end{bmatrix} \begin{bmatrix} y_{\mathcal{U}} \\ y_{\mathcal{R}} \end{bmatrix} \\ &= k_{\mathcal{U}}(x)^\top A^{-1} y_{\mathcal{U}} + \big(k_{\mathcal{R}}(x) - B^\top A^{-1} k_{\mathcal{U}}(x)\big)^\top S^{-1} \big(y_{\mathcal{R}} - B^\top A^{-1} y_{\mathcal{U}}\big). \end{aligned}$$

Define

$$r_{\mathcal{U}} = y_{\mathcal{R}} - B^\top A^{-1} y_{\mathcal{U}}, \qquad c_{\mathcal{U}}(x) = k_{\mathcal{R}}(x) - B^\top A^{-1} k_{\mathcal{U}}(x).$$

Since $\mu_{\mathcal{U}}(x) = k_{\mathcal{U}}(x)^\top A^{-1} y_{\mathcal{U}}$, we obtain

$$\mu_{\mathcal{D}}(x) - \mu_{\mathcal{U}}(x) = c_{\mathcal{U}}(x)^\top S^{-1} r_{\mathcal{U}}.$$

Next, we derive the posterior variance gap. Expanding the quadratic form inside $\sigma^2_{\mathcal{D}}(x)$,

$$\begin{aligned} & k_{\mathcal{D}}(x)^\top \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}^{-1} k_{\mathcal{D}}(x) \\ &\quad = k_{\mathcal{U}}(x)^\top A^{-1} k_{\mathcal{U}}(x) + \big(k_{\mathcal{R}}(x) - B^\top A^{-1} k_{\mathcal{U}}(x)\big)^\top S^{-1} \big(k_{\mathcal{R}}(x) - B^\top A^{-1} k_{\mathcal{U}}(x)\big) \\ &\quad = k_{\mathcal{U}}(x)^\top A^{-1} k_{\mathcal{U}}(x) + c_{\mathcal{U}}(x)^\top S^{-1} c_{\mathcal{U}}(x). \end{aligned}$$

Therefore,

$$\sigma^2_{\mathcal{D}}(x) = k(x,x) - k_{\mathcal{U}}(x)^\top A^{-1} k_{\mathcal{U}}(x) - c_{\mathcal{U}}(x)^\top S^{-1} c_{\mathcal{U}}(x),$$

and hence

$$\sigma^2_{\mathcal{U}}(x) - \sigma^2_{\mathcal{D}}(x) = c_{\mathcal{U}}(x)^\top S^{-1} c_{\mathcal{U}}(x).$$

Since $S \succ 0$, the right-hand side is nonnegative, so

$$\sigma^2_{\mathcal{U}}(x) \geq \sigma^2_{\mathcal{D}}(x) \geq 0.$$

We next justify the coarse bound on $\|S^{-1}\|$. Observe that

$$S = \sigma_n^2 I + \Big(K_{\mathcal{R}\mathcal{R}} - K_{\mathcal{R}\mathcal{U}}\big(K_{\mathcal{U}\mathcal{U}} + \sigma_n^2 I\big)^{-1} K_{\mathcal{U}\mathcal{R}}\Big).$$

The term in parentheses is the conditional covariance of the latent function values on $\mathcal{R}$ given noisy observations on $\mathcal{U}$, hence it is positive semidefinite. Therefore,

$$S \succeq \sigma_n^2 I \qquad \Longrightarrow \qquad \|S^{-1}\| \leq \sigma_n^{-2}.$$

Finally, using

$$|u^\top M v| \leq \|u\| \, \|M\| \, \|v\|,$$

we obtain

$$|\mu_{\mathcal{D}}(x) - \mu_{\mathcal{U}}(x)| \leq \|c_{\mathcal{U}}(x)\| \, \|S^{-1}\| \, \|r_{\mathcal{U}}\|,$$

and

$$0 \leq \sigma^2_{\mathcal{U}}(x) - \sigma^2_{\mathcal{D}}(x) \leq \|c_{\mathcal{U}}(x)\|^2 \, \|S^{-1}\|.$$

Substituting $\|S^{-1}\| \leq \sigma_n^{-2}$ yields the coarse bounds stated in the theorem. □

A.3 Proof of Proposition 5.3

Let $z_{\mathcal{D}}$ denote the response vector centered by the chosen GP mean function. Under the zero-mean notation of Theorem 5.1, $z_{\mathcal{D}} = y_{\mathcal{D}}$. Define

$$\lambda := \big(K_{\mathcal{D}\mathcal{D}} + \sigma_n^2 I\big)^{-1} z_{\mathcal{D}} = \begin{bmatrix} \lambda_{\mathcal{U}} \\ \lambda_{\mathcal{R}} \end{bmatrix}.$$

The block linear system gives

$$A\lambda_{\mathcal{U}} + B\lambda_{\mathcal{R}} = z_{\mathcal{U}}, \qquad B^\top \lambda_{\mathcal{U}} + C\lambda_{\mathcal{R}} = z_{\mathcal{R}}.$$

From the first equation,

$$\lambda_{\mathcal{U}} = A^{-1}(z_{\mathcal{U}} - B\lambda_{\mathcal{R}}).$$

Substituting this into the second equation yields

$$\big(C - B^\top A^{-1} B\big)\lambda_{\mathcal{R}} = z_{\mathcal{R}} - B^\top A^{-1} z_{\mathcal{U}}.$$

Using $S = C - B^\top A^{-1} B$ and the definition of $r_{\mathcal{U}}$, this is

$$S\lambda_{\mathcal{R}} = r_{\mathcal{U}}, \qquad \lambda_{\mathcal{R}} = S^{-1} r_{\mathcal{U}}.$$

The full-data scalar response-gradient score vector is $s = -\lambda$, and therefore

$$s_{\mathcal{R}} = -S^{-1} r_{\mathcal{U}}, \qquad r_{\mathcal{U}} = -S s_{\mathcal{R}}.$$

Substituting $S^{-1} r_{\mathcal{U}} = -s_{\mathcal{R}}$ into Theorem 5.1 gives

$$\mu_{\mathcal{D}}(x) - \mu_{\mathcal{U}}(x) = c_{\mathcal{U}}(x)^\top S^{-1} r_{\mathcal{U}} = -c_{\mathcal{U}}(x)^\top s_{\mathcal{R}}.$$

The pointwise and uniform norm bounds follow from the Cauchy-Schwarz inequality and the definition of $\rho_{\mathcal{U}}$. □

A.4 Uniform posterior distortion bound

Corollary A.1 (Uniform posterior distortion bound for a fixed retained subset)**.** *Assume $\mathcal{X}$ is compact and the kernel is continuous, so that*

$$\rho_{\mathcal{U}} := \sup_{x \in \mathcal{X}} \|c_{\mathcal{U}}(x)\|$$

is finite. Define

$$\xi_{\mathcal{U}} := \|r_{\mathcal{U}}\|.$$

Then

$$\sup_{x \in \mathcal{X}} |\mu_{\mathcal{D}}(x) - \mu_{\mathcal{U}}(x)| \le \rho_{\mathcal{U}}\, \xi_{\mathcal{U}}\, \|S^{-1}\|,$$

and

$$\sup_{x \in \mathcal{X}} \big(\sigma^2_{\mathcal{U}}(x) - \sigma^2_{\mathcal{D}}(x)\big) \le \rho^2_{\mathcal{U}}\, \|S^{-1}\|.$$

In particular,

$$\sup_{x \in \mathcal{X}} |\mu_{\mathcal{D}}(x) - \mu_{\mathcal{U}}(x)| \le \frac{\rho_{\mathcal{U}}\, \xi_{\mathcal{U}}}{\sigma_n^2}, \qquad \sup_{x \in \mathcal{X}} \big(\sigma^2_{\mathcal{U}}(x) - \sigma^2_{\mathcal{D}}(x)\big) \le \frac{\rho^2_{\mathcal{U}}}{\sigma_n^2}.$$

A.5 A standard-deviation gap lemma

Lemma A.2 (From variance gap to standard-deviation gap)**.** *For every test point x,*

$$|\sigma_{\mathcal{U}}(x) - \sigma_{\mathcal{D}}(x)| \le \sqrt{\sigma^2_{\mathcal{U}}(x) - \sigma^2_{\mathcal{D}}(x)}.$$

Consequently,

$$|\sigma_{\mathcal{U}}(x) - \sigma_{\mathcal{D}}(x)| \le \|c_{\mathcal{U}}(x)\| \sqrt{\|S^{-1}\|} \le \frac{\|c_{\mathcal{U}}(x)\|}{\sigma_n}.$$

Proof. By Theorem 5.1, $\sigma^2_{\mathcal{U}}(x) \ge \sigma^2_{\mathcal{D}}(x) \ge 0$. Hence

$$\big(\sigma_{\mathcal{U}}(x) - \sigma_{\mathcal{D}}(x)\big)^2 = \sigma^2_{\mathcal{U}}(x) + \sigma^2_{\mathcal{D}}(x) - 2\sigma_{\mathcal{U}}(x)\sigma_{\mathcal{D}}(x) \le \sigma^2_{\mathcal{U}}(x) - \sigma^2_{\mathcal{D}}(x),$$

because

$$\sigma^2_{\mathcal{U}}(x) - \sigma^2_{\mathcal{D}}(x) - \big(\sigma_{\mathcal{U}}(x) - \sigma_{\mathcal{D}}(x)\big)^2 = 2\sigma_{\mathcal{D}}(x)\big(\sigma_{\mathcal{U}}(x) - \sigma_{\mathcal{D}}(x)\big) \ge 0.$$

Taking square roots yields the first inequality. The remaining bounds follow from Theorem 5.1. □

A.6 Proof of Corollary 5.4

For any $x \in \mathcal{X}$, the triangle inequality gives

$$|\alpha_{\mathcal{D}}^{\mathrm{UCB}}(x) - \alpha_{\mathcal{U}}^{\mathrm{UCB}}(x)| \leq |\mu_{\mathcal{D}}(x) - \mu_{\mathcal{U}}(x)| + \sqrt{\beta_t}|\sigma_{\mathcal{D}}(x) - \sigma_{\mathcal{U}}(x)|.$$

Proposition 5.3 bounds the posterior-mean gap by

$$|\mu_{\mathcal{D}}(x) - \mu_{\mathcal{U}}(x)| \leq \|c_{\mathcal{U}}(x)\| \, \|s_{\mathcal{R}}\|.$$

Lemma A.2 bounds the standard-deviation gap by

$$|\sigma_{\mathcal{D}}(x) - \sigma_{\mathcal{U}}(x)| \leq \|c_{\mathcal{U}}(x)\| \sqrt{\|S^{-1}\|}.$$

Combining these two inequalities proves the pointwise bound. Taking the supremum over $x \in \mathcal{X}$ proves the uniform bound. □

A.7 Scope and limitations

The results in this appendix should be read with the following scope in mind.

- The posterior-gap theorem is a statement about subset-of-data fitting under shared hyper-parameters and shared observation-noise variance. It isolates the effect of data pruning itself.
- The theorem does not account for additional error introduced by re-estimating hyper-parameters on the retained subset.
- The posterior-gap theorem applies to any retained subset and therefore serves as a diagnostic result for the subset selected by GSSBO; it is not, by itself, an unconditional proof that the GSSBO selection rule always yields a small posterior distortion.
- The residual bridge explains the posterior-mean term for a fixed retained subset. It does not prove that the selection rule minimizes $\|s_{\mathcal{R}}\|$, and it does not control the variance or UCB exploration term by itself.
- The UCB acquisition-stability corollary is not a cumulative-regret bound and does not prove that the selection rule always produces small acquisition distortion.

B Appendix Experiments

B.1 Additional Implementation Details

Greedy cosine-diversity selection. For each selection pool $\mathcal{P}_t$, we compute the pool-specific sensitivity embeddings $\mathbf{g}_i^{(\mathcal{P}_t)}$ and normalize them as $\mathbf{v}_i = \mathbf{g}_i^{(\mathcal{P}_t)}/\|\mathbf{g}_i^{(\mathcal{P}_t)}\|_2$. The latest observation is inserted first. We then greedily choose the remaining $M-1$ samples by selecting candidates with small cumulative cosine similarity to the already retained set. After the pairwise pool scores $C_{ij} = \mathbf{v}_i^\top \mathbf{v}_j$ are available, we maintain $h_i = \sum_{j \in \mathcal{S}} C_{ij}$ for each unselected candidate i, where $\mathcal{S}$ is the current retained set. When candidate a is added, the scores are updated as $h_i \leftarrow h_i + C_{ia}$. Thus the greedy stage costs $\mathcal{O}(q_t M)$ after the pool-level cosine scores have been constructed or cached; embedding construction and cosine-score construction are counted separately in the overhead terms described in Section 4.3.

Runtime-triggered buffer initialization. The parameter Z is used only to initialize the buffer size at the first switch. In the reported experiments, GSSBO and RSSBO use the same runtime-triggered rule, hardware setting, and evaluation protocol; after the switch, M is kept fixed. Since the trigger depends on measured iteration time, the exact switching iteration may vary across machines. For deterministic reruns, one can instead fix M directly or fix the switching iteration and then apply the same subset-maintenance rule. This affects only the practical initialization of M, not the cosine-diversity selection rule once M and $\mathcal{P}_t$ are fixed.

B.2 Supplementary experiments

Figure 5 and 6 compare the cumulative runtime over 1000 iterations on Eggholder2, Levy20, Rastrigin100, and the diabetes hyperparameter optimization task.

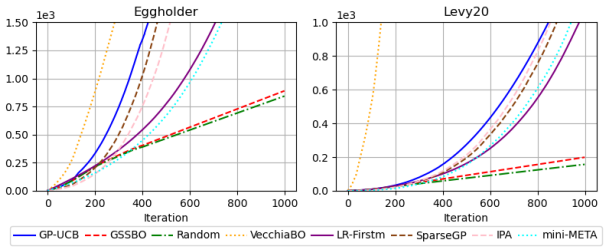

Figure 5: Cumulative runtime (seconds) on Eggholder2 and Levy20.

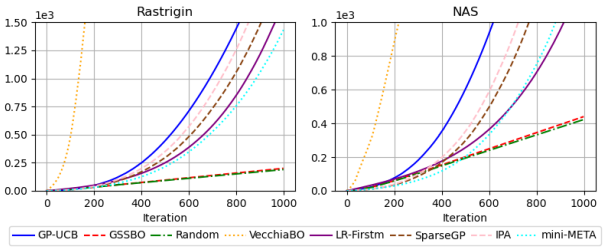

Figure 6: Cumulative runtime (seconds) on Rastrigin100 and the diabetes hyperparameter optimization task.

B.3 Subset Samples Distribution Study

Figure 7 illustrates the sample distribution of GSSBO and standard GP-UCB, on the first two dimensions of the Hartmann6 function. In this experiment, we recorded the first 200 sequential samples from a standard BO process and constrained the buffer size to 100. The objective is to identify the global minimum, and darker-colored samples correspond to values closer to the optimal. During the optimization process, only a small number of samples are located near the optimal value. In this run, GSSBO selection retains more samples near low-value regions while still covering diverse observed locations, as indicated by a higher number of darker-colored samples in the middle panel. In contrast, the random selection strategy appears to retain fewer samples near the optimal or suboptimal regions in this run, as represented by the retention of many lighter-colored samples in the right panel.

While BO algorithms are designed to balance exploitation and exploration, with limited budget in practice, they can concentrate evaluations around current best regions before shifting to exploration (Wang and Ng, 2020), which may affect optimization performance. With GSSBO selection, the relative density of retained samples near low-value regions increases in this visualization. This pattern may help avoid excessive concentration around the current best region and encourage broader coverage of the observed search space.

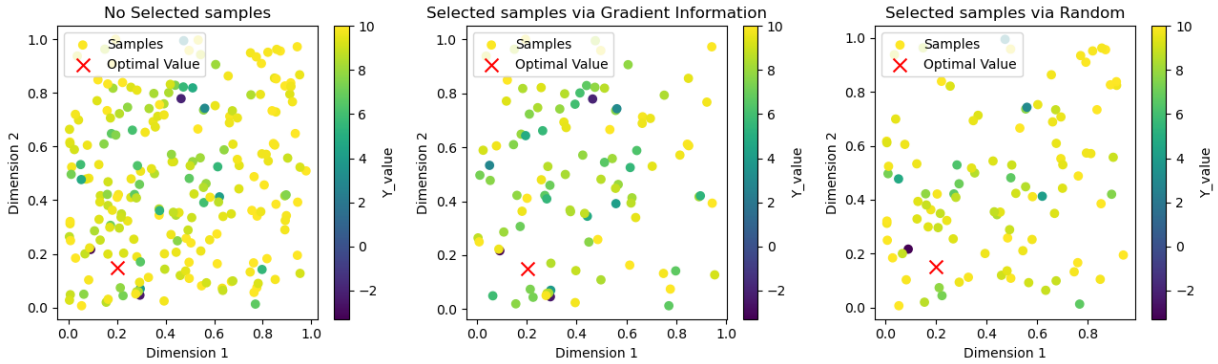

Figure 7: Sample distribution: GP-UCB (left), GSSBO (middle), and RSSBO (right).

B.4 Experiments on K-means++ Selection

Oglic and Gärtner (2017) and Hayakawa et al. (2023) proposed to select a subset in RKHS, then employed them to construct the Nyström low-rank approximation. We included it as an additional baseline. In these

experiments, Fig. 8 shows that K-means++ has no clear cumulative-regret advantage over GSSBO and incurs higher runtime.

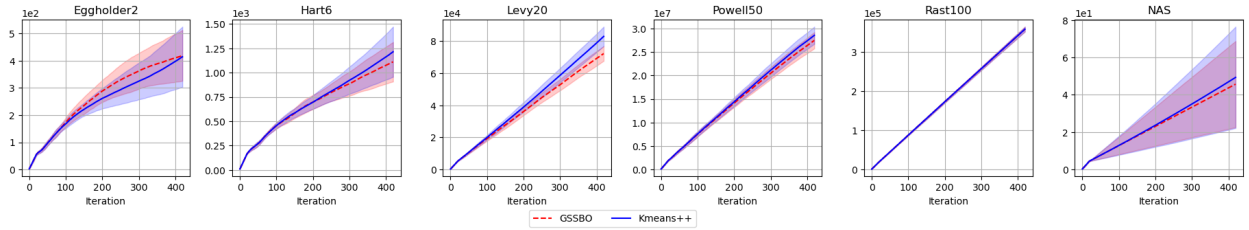

Figure 8: Cumulative regret of GSSBO and K-means++ on Eggholder2, Hart6, Levy20, Powell50, Rastrigin100, and the diabetes hyperparameter optimization task.

B.5 Experiments on high-dimensional BO methods

We noticed that several compared methods show degraded performance on high-dimensional tasks, so we included additional high-dimensional baselines: REMBO (Wang et al., 2016) and HeSBO (Nayebi et al., 2019). The results are shown in Fig. 9. The difference between the three methods is subtle on Levy20, while REMBO and HeSBO perform worse on Powell50 and Rastrigin100 in our results, possibly because the random embedding can discard task-relevant directions.

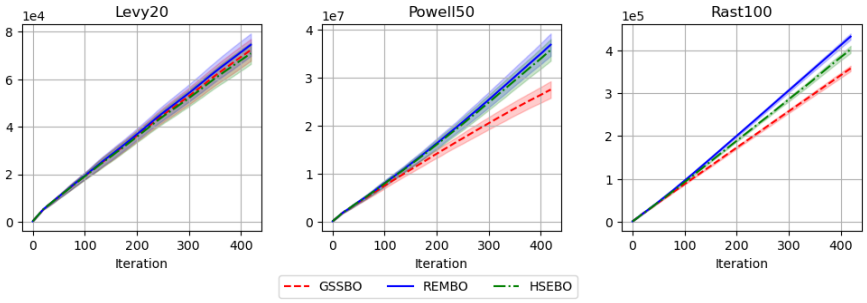

Figure 9: Cumulative regret of GSSBO, REMBO, and HeSBO on the Levy20, Powell50, and Rastrigin100 functions.

B.6 Experiments on other surrogates

Although our theoretical analysis focuses on Gaussian Process surrogates, we also explore whether the same subset-maintenance idea can be paired empirically with other surrogate models. For Bayesian neural networks (BNN) (Mullachery et al., 2018) and Deep Kernel (DK) methods (Wilson et al., 2016), the selection step uses surrogate-specific selection features rather than the GP-specific precision-column form in Eq. 3. The corresponding results are shown in Fig. 10.

B.7 Auxiliary RMSE Diagnostic for Subset GP Fits

As an auxiliary posterior-mean diagnostic, we compare the RMSE of three surrogate models on a test dataset: (i) the full-data GP used in GP-UCB, (ii) the GP trained on the GSSBO-selected subset, and (iii) the GP trained on a randomly chosen subset of the same size (RSSBO). Subset selection starts at the 60th iteration. At each iteration, we construct the corresponding GP surrogate and compute the root mean square error (RMSE) on a test set of size n_{test}, where $\hat{\mu}_t(\tilde{x}_j)$ denotes the posterior mean prediction of the surrogate at the

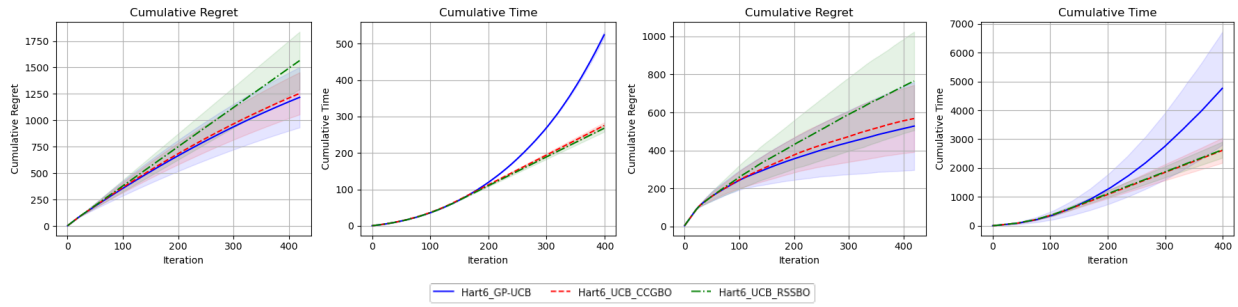

Figure 10: Surrogate-specific Hart6 experiments using Deep Kernel (left pair) and BNN (right pair) surrogate settings. Within each pair, the first panel reports cumulative regret and the second reports cumulative runtime.

j-th test input $\tilde{x}_j$:

$$\mathrm{RMSE}_t = \sqrt{\frac{1}{n_{\mathrm{test}}} \sum_{j=1}^{n_{\mathrm{test}}} \left(\hat{\mu}_t(\tilde{x}_j) - f(\tilde{x}_j)\right)^2}.$$

As shown in Figure 11, GSSBO often attains RMSE values close to the full-data GP in this experiment, while RSSBO shows higher error and variance in these runs. This suggests that the GSSBO-selected subset can preserve posterior-mean prediction accuracy better than a random subset of the same size in this diagnostic. This result should be interpreted as supplementary evidence only: RMSE measures posterior-mean prediction error and does not directly test posterior variance, UCB acquisition values, or the residual quantities appearing in the theoretical posterior-gap bounds.

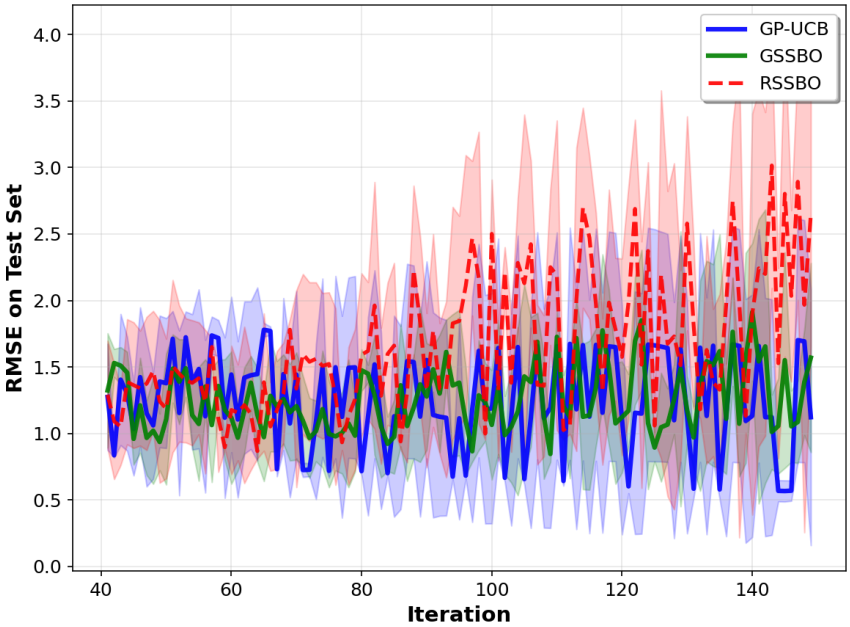

Figure 11: Auxiliary RMSE diagnostic over iterations for full-data, GSSBO subset, and RSSBO subset GP fits.