Hypervolume Knowledge Gradient: A Lookahead Approach for Multi-Objective Bayesian Optimization with Partial Information

Samuel Daulton [1 2] **Maximilian Balandat** [1] **Eytan Bakshy** [1]

Abstract

Bayesian optimization is a popular method for sample efficient multi-objective optimization. However, existing Bayesian optimization techniques fail to effectively exploit common and often-neglected problem structure such as decoupled evaluations, where objectives can be queried independently from one another and each may consume different resources, or multi-fidelity evaluations, where lower fidelity-proxies of the objectives can be evaluated at lower cost. In this work, we propose a general one-step lookahead acquisition function based on the Knowledge Gradient that addresses the complex question of what to evaluate when and at which design points in a principled Bayesian decision-theoretic fashion. Hence, our approach naturally addresses decoupled, multi-fidelity, and standard multi-objective optimization settings in a unified Bayesian decision making framework. By construction, our method is the one-step Bayes-optimal policy for hypervolume maximization. Empirically, we demonstrate that our method improves sample efficiency in a wide variety of synthetic and real-world problems. Furthermore, we show that our method is general-purpose and yields competitive performance in standard (potentially noisy) multi-objective optimization.

1. Introduction

Black-box optimization is a ubiquitous problem in scientific and engineering applications. In many scenarios, there are multiple objective functions that a decision maker seeks to optimize simultaneously. Multi-objective Bayesian optimization (MOBO) is a powerful technique to achieve this with high sample efficiency (Hernandez-Lobato et al., 2016).

[1]Meta, USA [2]University of Oxford, Oxford, UK. Correspondence to: Samuel Daulton <sdaulton@meta.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Most MOBO algorithms assume that all objectives are evaluated jointly (i.e. the evaluations of the objectives are *coupled*). However, in practice there are many *partial information* settings in which this is not the case. For instance, we may have the ability to evaluate objectives individually (the *decoupled evaluation* setting, see Figure 1), or we may have lower-fidelity proxies available (the *multi-fidelity* setting) in order to save time and/or resources.

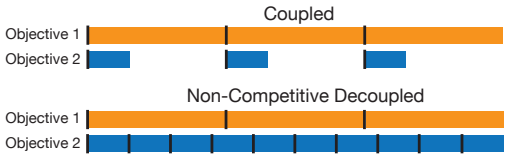

Figure 1: Decoupled evaluation allows for multiple objectives to be evaluated in non-blocking fashion (bar lengths correspond to evaluation time). With *non-competitive decoupling*, objectives have independent evaluation resources and do not compare for a shared resource.

Consider for example the problem of neural architecture search (NAS), in which we aim to identify optimal neural network architectures with respect to both model quality (e.g. accuracy) and hardware-specific metrics such as prediction latency measured on-device (Janapa Reddi et al., 2022). In general, neither metric can be computed analytically as a function of the network architecture. Measuring accuracy typically requires a substantial amount of computation time as the NN must be trained and evaluated. Measuring latency requires access to the specific hardware of interest (e.g. a particular mobile device type) and often few devices may be tested simultaneously (Ignatov et al., 2019). However, device-specific latency can be evaluated on untrained NNs with reasonable accuracy with a short benchmark, making evaluation less time-consuming. This setting is illustrated in Figure 2, where we consider the scenario where we have access to a number of compute nodes—each of which can be used to train and evaluated a model—and a small number of mobile devices that can be used for measuring latency. The time for training models and evaluating accuracy will typically be much longer than the time required to measure on-device latency, but can happen asynchronously.

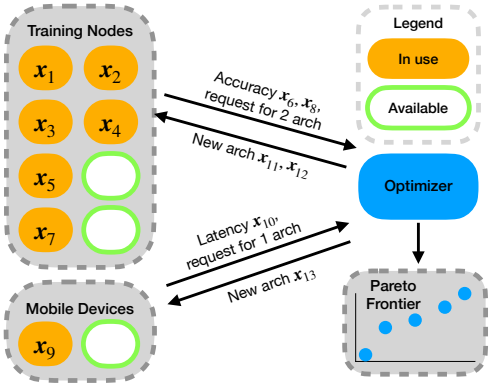

Figure 2: Multi-Objective Neural Architecture Search is one example problem with *decoupled* evaluations, where the objectives can be evaluated independently.

In this setting, a standard MOBO algorithm would simply generate architectures to be evaluated on all objectives and wait for evaluations to complete before generating new candidates (see, e.g., Guerrero-Viu et al. (2021); Eriksson et al. (2021)). This can be very inefficient, especially if evaluation time (more generally, cost) differs substantially between the objectives. Instead, one may asynchronously choose an architecture in a *decoupled* fashion to evaluate on a given objective whenever capacity for that objective becomes available. Figure 3 shows that even a simple policy that employs this strategy and selects architectures in a (quasi-)random fashion significantly outperforms standard MOBO methods.

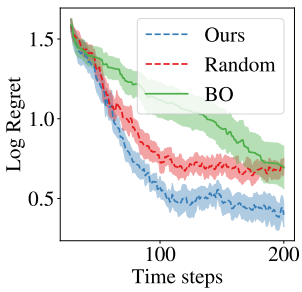

Figure 3: A random search algorithm that generates candidates for each objective in a decoupled asynchronous fashion outperforms a state-of-the-art MOBO method (qNEHVI) on a NAS problem. Our method significantly outperforms both.

Although some MOBO methods can exploit the problem's decoupled asynchronous structure (e.g., Hernandez-Lobato et al. (2016); Suzuki et al. (2020)), recent work noted that the performance improvements of existing decoupled methods relative to their non-decoupled counterparts are small (Tu et al., 2022). In contrast, we develop a method that significantly outperforms state-of-the-art MOBO algorithms for these settings.

The NAS problem described above is an instance of a more general class of problems that is ubiquitous in the physical science and engineering. For example, in material science high-throughput screening may be applied to discover candidate compounds, but computationally-expensive simulations and/or physical experiments may be necessary to characterize the final behavior of the compound (Mukadum et al., 2021). In the design of low-carbon-emission concrete, objectives of interest are evaluated at multiple timescales: carbon emissions can be measured within the first several hours of production, while properties such as compressive strength can take several weeks to evaluate (Barcelo et al., 2014). Low-fidelity proxies for compressive strength (such as strength after 3 days) can be evaluated in less time, but due to the destructive nature of testing, such measurements forgo the ability to evaluate the compressive strength at a target fidelity (e.g., after 60 days). As in the NAS example, there is limited capacity for testing certain properties (e.g., only so many rods of concrete can be cured or stored simultaneously).

The decoupled and multi-fidelity problems are instances in which a practitioner wishes to perform MOBO with *incomplete* information. By leveraging this partial information, one can substantially reduce the cost of optimization compared to naïve approaches.

Contributions

1. We formulate the MOBO problem by considering one-step look-ahead optimization of hypervolume.
2. We propose the Hypervolume Knowledge Gradient (HV-KG), a unifying acquisition strategy that allows for conditioning on incomplete information and generating candidates in a way that takes the evaluation structure into account.
3. We derive an unbiased gradient estimator and provide a computationally efficient technique for optimizing HV-KG using sample average approximation.
4. We demonstrate substantial gains in optimization performance of HV-KG over state-of-the-art MOBO methods on a variety synthetic and real-world multi-fidelity and decoupled problems.

2. Preliminaries

2.1. Multi-Objective Optimization (MOO)

In MOO, the goal is to optimize a vector valued function $\boldsymbol{f}(x) = (f^{(1)}(x), \ldots, f^{(M)}(x))$ over a compact hyperrectangular *search space* $\mathcal{X} \subset \mathbb{R}^d$. Typically there is no single best solution, and therefore the goal is to identify the set of designs with optimal objective trade-offs. We say a solution $\boldsymbol{f}(x)$ dominates another solution $\boldsymbol{f}(x')$, denoted by $\boldsymbol{f}(x) \succ \boldsymbol{f}(x')$, if $f^{(m)}(x) \geq f^{(m)}(x')$ for all m and there exists i such that $f^{(i)}(x) > f^{(i)}(x')$. An objective vector is Pareto optimal iff it is not dominated. The set

$\mathcal{P}^* = \{\boldsymbol{f}(x) \mid \nexists\, x' \in \mathcal{X}\ s.t.\ \boldsymbol{f}(x') \succ \boldsymbol{f}(x)\}$ of such vectors is called the Pareto frontier. The corresponding set of optimal *designs* is called the Pareto set $\mathcal{X}^*$ and is defined as

$$\mathcal{X}^* = \{x \in \mathcal{X} \mid \nexists\, x' \in \mathcal{X}\ s.t.\ \boldsymbol{f}(x') \succ \boldsymbol{f}(x)\}. \quad (1)$$

The image of $\mathcal{X}^*$ is $\mathcal{P}^*$. Given a Pareto frontier, a decision-maker can select a design with corresponding objectives that align with their preferences. The hypervolume indicator (HV) is a popular quality measure of a Pareto frontier.

Definition 2.1. The hypervolume indicator (HV) of a Pareto frontier $\mathcal{P}$ is the M-dimensional Lebesgue measure of the space $Z = \{z \in \mathbb{R}^M : \exists\, y \in \mathcal{P}\ s.t.\ y \succ z \succ r\}$ that is dominated by $\mathcal{P}$ and bounded from below by a reference point $r \in \mathbb{R}^M$: $\mathrm{HV}(\mathcal{P}, r) = \int_{\mathbb{R}^M} \mathbb{1}_Z(z)dz$, where $\mathbb{1}_Z(z)$ denotes characteristic function of Z.[1]

HV monotonically increases with Pareto dominance, which guarantees that it is maximized by the Pareto frontier (the image of the Pareto set) (Bader and Zitzler, 2011):

$$\mathrm{HV}\big[\{\boldsymbol{f}(x)\}_{x\in\mathcal{X}^*}\big] = \max_{\mathcal{X}'\subseteq\mathcal{X}} \mathrm{HV}[\{\boldsymbol{f}(x)\}_{x\in\mathcal{X}'}]. \quad (2)$$

Hence we can express the goal of MOO as finding the smallest[2] set of designs $\mathcal{X}^*$ that collectively maximize the HV:

$$\mathcal{X}^* = \arg\min\Big\{|\mathcal{X}''| : \mathcal{X}'' \in \arg\max_{\mathcal{X}'\subseteq\mathcal{X}} \mathrm{HV}[\{\boldsymbol{f}(x)\}_{x\in\mathcal{X}'}]\Big\}. \quad (3)$$

Maximizing HV is a commonly used optimization goal that has been shown to produce high-quality approximate Pareto frontiers (Emmerich et al., 2011).

2.2. Bayesian Optimization (BO)

BO is a sample-efficient optimization method that models the objectives using a probabilistic *surrogate*, typically a Gaussian process (GP). Leveraging this surrogate, BO employs an *acquisition function* (AF) that quantifies the value of evaluating a new design on the objective functions. One popular AF for MOBO is expected hypervolume improvement (EHVI) (Emmerich et al., 2011), which quantifies the improvement in HV of the observed data after evaluating x:

$$\alpha_{\mathrm{EHVI}}(x) = \mathbb{E}\big[\mathrm{HV}\big(\mathcal{Y} \cup \{\boldsymbol{f}(x)\}\big) - \mathrm{HV}\big(\mathcal{Y}\big) \mid \mathcal{D}\big],$$

where the expectation is over the model posterior $P(\boldsymbol{f}|\mathcal{D})$, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ are the designs evaluated so far and their corresponding observations, and $\mathcal{Y} := \{y_i\}_{i=1}^n$.

A BO policy selects one or more designs by finding the maximizer of the AF with respect to a single design x when evaluation is done sequentially, or a *batch* of designs $\boldsymbol{x} = \{x_1, \dots x_q\}$ when performing BO in parallel.[3] The designs are then evaluated on the objective functions, and the surrogate model is updated with the new observations $\boldsymbol{y}(\boldsymbol{x}) = \{y(x_1), \dots y(x_q)\}$. BO proceeds until an evaluation budget is depleted.

2.3. BO with Partial Information

We briefly review terminology and common approaches to multi-fidelity (MF) BO and BO with decoupled evaluation.

Multi-fidelity BO. In multi-fidelity (MF) optimization, designs can be evaluated at different qualities within a fidelity space $\mathcal{S} \subset \mathbb{R}^K$. Examples of fidelity parameters may include the number of datapoints used to train a machine learning model or the resolution of a simulator. Lower fidelity observations are assumed to incur lower cost (e.g., compute or physical resources, time), but may differ from the value of the target objective $f^{(i)}(\cdot, s_\diamond)$, where $s_\diamond$ is a known target fidelity. MF-BO policies select designs and fidelities to query $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{s})$ with the aid of a surrogate model that borrows strength across different fidelities. This can lead to significant improvements in performance within a cost budget (Poloczek et al., 2017; Takeno et al., 2020; Wu et al., 2020a; Irshad et al., 2021). Typically, designs and fidelities are selected in a cost-aware fashion to maximize the acquisition value per unit cost (Snoek et al., 2012; Lee et al., 2020; Wu et al., 2020a). Specifically, the acquisition value of evaluating a set of designs at corresponding fidelities is weighted by the inverse of a cost function $\lambda_{\mathrm{MF}}(\boldsymbol{x}, \boldsymbol{s}) : \mathcal{X}^q \times \mathcal{S}^q \to \mathbb{R}_{>0}$, where $\boldsymbol{s} = \{s_1, \dots, s_q\}$.

BO with Decoupled Evaluations. In *decoupled* problems, objectives can be evaluated independently at potentially different costs. As a result, any given evaluation of a design x may not contain a full vector of outputs $y \in \mathbb{R}^M$, but rather some subset of outcomes (typically, a single objective). We further distinguish between *competitive decoupling* (CD) and *non-competitive decoupling* (NCD) (Hernández-Lobato et al., 2016). With CD, evaluation resources are shared between objectives, whereas with NCD, they are not. Decoupled BO policies select designs to be evaluated on particular objectives. Similar to the MF setting, this is typically achieved by maximizing the acquisition value per unit cost. Here, the cost function $\lambda_{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{m}) : \mathcal{X}^q \times \mathcal{M}^q \to \mathbb{R}_{>0}$ characterizes the cost of evaluating a set of q designs, $\boldsymbol{x}$, with respect to $\boldsymbol{m} = \{m_1, \dots, m_q\} \in \mathcal{M}^q$ objectives, where $\mathcal{M} = \{m\}_{m=1}^M$ is the set of objective indices. Similar to the MF setting, exploiting decoupling can substantially improve optimization performance within a given budget.

[1] Henceforth, we omit r from HV for brevity.

[2] We are interested in the smallest hypervolume-maximizing set because for any hypervolume-maximizing set $\mathcal{X}' \subseteq \mathcal{X}$ and design $x \in \mathcal{X}$, $\mathrm{HV}(\mathcal{X}') = \mathrm{HV}(\mathcal{X}' \cup \{\boldsymbol{f}(x)\})$.

[3] For the sake of generality and notational simplicity, we will assume that acquisition functions are maximized with respect a set of designs (i.e., the joint value of $\boldsymbol{x}$) throughout this work.

3. Related Work

Many recent works have focused on multi-objective BO. Numerous techniques exist, the three most prominent families of methods are hypervolume-based approaches (Lukovic et al., 2020; Daulton et al., 2021; 2022b), information theoretic methods (Hernandez-Lobato et al., 2016; Belakaria et al., 2019; Suzuki et al., 2020; Tu et al., 2022; Garrido-Merchán et al., 2023), and scalarization-based techniques (Knowles, 2006; Golovin and Zhang, 2020; Daulton et al., 2022a). However, the setting with incomplete information is much less studied.

The only methods to consider MOO with decoupled evaluations are the entropy-based Predictive Entropy Search (PESMO) (Hernandez-Lobato et al., 2016) and Pareto Frontier Entropy Search (PFES) (Suzuki et al., 2020). Recent work on multi-objective Joint Entropy Search (JES) (Tu et al., 2022) noted that the improvements in sample efficiency appeared marginal at best in those works and therefore abstained from implementing and evaluating JES in the decoupled setting. In contrast to this finding, we observe that exploiting decoupled evaluations with HV-KG (and even random search) can greatly improve sample efficiency.

In the MF setting, Belakaria et al. (2020) proposed MF-OSEMO, a multi-objective extension of Multi-Fidelity Max-Value Entropy Search (Takeno et al., 2020). However, this method is only applicable in discrete fidelity settings, assumes that the objectives monotonically increase with the fidelity parameter, and, similar to Multi-Objective Max-Value Entropy Search (Belakaria et al., 2019), suffers from significant approximation error (see Tu et al. (2022) for details). MoFiBay (Chen et al., 2022) outperforms MF-OSEMO, but is also limited to discrete fidelities. Irshad et al. (2021) introduced a MF method called MOMF, which uses the fidelity parameter as an additional "trust" objective and employs an inverse cost-weighted EHVI over all objectives. Although this approach performs quite well empirically, it does not a employ a principled procedure for selecting the fidelity parameter, and it does not specifically aim to learn the Pareto frontier over the M objectives at the target fidelity, but rather to learn the Pareto frontier over the M objectives and the trust objective. He et al. (2022) also consider a MF EHVI variant, but it is limited to the bi-fidelity setting. Guerrero-Viu et al. (2021) extend the MF BO methods BANANAS (White et al., 2021) and BOHB (Falkner et al., 2018) to the multi-objective setting, but find that full-fidelity EHVI outperforms both methods.

While our contributions build upon previous work on the Knowledge Gradient (Frazier et al., 2008; Scott et al., 2011) and its MF extensions (Poloczek et al., 2017; Wu et al., 2020a), none of these works consider the MOO setting. Q. Yahyaa et al. (2014) consider KG in the multi-objective bandit setting leveraging linear and Chebyshev scalarizations, but they do not consider the BO setting and their evaluations are quite limited.

4. Pareto Set Selection

In MOBO, a decision maker must infer the Pareto optimal designs after receiving a finite number of observations. In the setting where observations of all objectives are available for all designs and are free of noise, a common approach is to restrict the Pareto set selection in (3) to only consider dominance with respect to $X_\mathcal{D} := \{x : (x, \cdot) \in \mathcal{D}\}$, the set of previously evaluated designs:

$$\hat{\mathcal{X}}^* = \left\{x \in \mathcal{X} \mid \nexists\, x' \in X_\mathcal{D} \; s.t. \; \boldsymbol{f}(x') \succ \boldsymbol{f}(x)\right\}.$$

or, equivalently, $\hat{\mathcal{X}}^* = \arg\max_{\mathcal{X}' \subseteq X} \text{HV}[\{\boldsymbol{f}(x)\}_{x \in \mathcal{X}'}]$. However, observations may be noisy $y \sim \mathcal{N}(\boldsymbol{f}(x), \sigma^2_{\text{noise}})$ and in which case the actual objective function values may not be directly observed (which can cause issues with traditional MOBO methods (Daulton et al., 2021)). Similarly, in the setting where not all objectives are evaluated for all designs or not all objectives are evaluated at the target fidelity, the set of designs that have been evaluated on all objectives can be small or empty. In such scenarios, it is common for a practitioner to identify the designs that are optimal with respect to their expected values under the surrogate model (Hernandez-Lobato et al., 2016; Belakaria et al., 2019; Suzuki et al., 2020; Tu et al., 2022) and to select the optimal designs over the entire search space. Concretely, under a Bayesian decision-theoretic framework, the optimal set of designs is selected as the set of designs $\mathcal{X}^*$ whose expected values under the posterior distribution of $\boldsymbol{f}$ conditional on the observed data $\mathcal{D}$ are Pareto optimal. In the standard sequential scenario,

$$\hat{\mathcal{X}}^* = \left\{x \in \mathcal{X} \mid \nexists\, x' \in \mathcal{X} \; s.t. \; \mathbb{E}_\mathcal{D}[\boldsymbol{f}(x')] \succ \mathbb{E}_\mathcal{D}[\boldsymbol{f}(x)]\right\},$$

where $\mathbb{E}_\mathcal{D}$ the expectation over the posterior of $\boldsymbol{f}$ conditional on $\mathcal{D}$. An equivalent problem is to find the set of designs that maximize the HV of the expected values:

$$\hat{\mathcal{X}}^* = \underset{\mathcal{X}' \subseteq \mathcal{X}}{\arg\max}\, \text{HV}\Big[\big\{\mathbb{E}_\mathcal{D}\big[\boldsymbol{f}(x)\big]\big\}_{x \in \mathcal{X}'}\Big]. \tag{4}$$

Since $\hat{\mathcal{X}}^*$ can be an infinite set, it is typically hard to identify exactly. A common approach is to identify a finite-cardinality approximate Pareto set $\hat{X}^*$ containing N_p designs, typically by running an evolutionary algorithm such as NSGA-II on $\mathbb{E}_\mathcal{D}[\boldsymbol{f}(x)]$ (Hernandez-Lobato et al., 2016; Belakaria et al., 2019; Suzuki et al., 2020; Tu et al., 2022). Often, the HV of the resulting Pareto frontiers is used for comparing their quality. We can directly express this optimization goal by restricting the HV maximization problem in Equation (4) to finite cardinality sets $|\mathcal{X}'| \leq N_p$:

$$\hat{X}^* = \underset{X \subseteq \mathcal{X},\; |X| \leq N_p}{\arg\max}\, \text{HV}\Big[\big\{\mathbb{E}_\mathcal{D}\big[\boldsymbol{f}(x)\big]\big\}_{x \in \mathcal{X}'}\Big].$$

Henceforth, we assume $|X| \leq N_p$. In the following, we will write $\boldsymbol{\mu}(X \mid \mathcal{D}) := \{\mathbb{E}_{\mathcal{D}}[\boldsymbol{f}(x)]\}_{x \in X}$.

5. A Knowledge Gradient Approach

Given the Bayesian decision-theoretic goal above, we derive a novel AF to explicitly target our end goal: inferring a hypervolume-maximizing finite Pareto set. Consider the scenario where one can obtain additional observations $(\boldsymbol{x}, \boldsymbol{y})$ before identifying the Pareto optimal designs conditional on $\mathcal{D}_{\boldsymbol{x}} := \mathcal{D} \cup \{(\boldsymbol{x}, \boldsymbol{y})\}$. Then, the one-step Bayes-optimal acquisition function, denoted as the Hypervolume Knowledge Gradient (HV-KG), is:

$$\alpha_{\text{HV-KG}}(\boldsymbol{x}) = \mathbb{E}_{\mathcal{D}}\left[\max_{X \subseteq \mathcal{X}} \text{HV}\left[\boldsymbol{\mu}(X \mid \mathcal{D}_{\boldsymbol{x}})\right] - \psi^*\right], \quad (5)$$

where $\psi^* := \max_{X \subseteq \mathcal{X}} \text{HV}\left[\boldsymbol{\mu}(X \mid \mathcal{D})\right]$. Conceptually, HV-KG quantifies the increase in hypervolume of the Pareto frontier across the expected values of the objectives. The outer expectation is necessary because $\boldsymbol{y}$ is a random variable and $\boldsymbol{\mu}$ depends on $\boldsymbol{y}$. Since ψ^* is constant conditional on $\mathcal{D}$, the maximizer of $\alpha_{\text{HV-KG}}(\boldsymbol{x})$ does not change if ψ^* is omitted.

Asynchronous Candidate Generation While (5) is the formulation for *parallel* (i.e. batch) candidate generation, it is straightforwardly extended to the setting of *asynchronous* generation, in which the result of some *pending points* $\tilde{\boldsymbol{x}}$ have yet to be observed. In this case the acquisition function is evaluated on $\boldsymbol{x} \cup \tilde{\boldsymbol{x}}$ but optimized only over $\boldsymbol{x}$.

6. Conditioning on Partial Information

Although HV-KG is applicable to standard MOBO problems where observations of all objectives are received for all designs, a key benefit of HV-KG is that it enables conditioning on incomplete information. In contrast, other popular HV-based methods (e.g. Emmerich and Fonseca (2011); Lukovic et al. (2020); Daulton et al. (2021)) cannot condition on incomplete information because they rely on utility functions that measure improvement with respect to an in-sample Pareto set and assume that observations of all objectives will be received for the selected candidate design. In contrast, HV-KG can leverage incomplete information (such as decoupled and multi-fidelity evaluations) simply by changing the new data $\mathcal{D}_{\boldsymbol{x}}$ that the model is conditioned on.

Decoupled Evaluations In the decoupled setting the objectives can be evaluated independently. The decoupled HV-KG acquisition function is[4]

$$\alpha_{\text{D-HV-KG}}(\boldsymbol{x}, \boldsymbol{m}) = \frac{\alpha_{\text{HV-KG}}(\boldsymbol{x})}{\lambda_D(\boldsymbol{x}, \boldsymbol{m})},$$

where now $\mathcal{D}_{\boldsymbol{x}} = \{(x_i, y_i^{(m_i)})\}_{i=1}^q$ and the cost λ_D is defined in Section 2.3. In CD, the evaluation budget is in terms of total cost and all objectives compete for shared resources. As such, we consider the case of $q = 1$, without loss of generality. The BO policy chooses the objective m and design x jointly in a cost-aware fashion. In NCD, the evaluation budget is in terms of time and all available evaluation capacity should be exploited.

Let $c \in \mathbb{N}^M$ denote the available evaluation capacity for each objective. The policy generates $q = \sum_{m=1}^M c^{(m)}$ candidates $\boldsymbol{x}$ jointly to exploit all available capacity. Each candidate is assigned to be evaluated on an objective specified by $\boldsymbol{m} \in \mathcal{M}^q$ such that $c^{(m)} = \sum_{i=1}^q \mathbb{1}(m_i = m)$ for all $m = 1, ..., M$.

Multi-Fidelity Let $\boldsymbol{\mu}_\diamond(X, \mathcal{D}) := \{\mathbb{E}_{\mathcal{D}}[\boldsymbol{f}(x, \boldsymbol{s}_\diamond)]\}_{x \in X}$. The multi-fidelity HV-KG AF is given by

$$\alpha_{\text{MF-HV-KG}}(\boldsymbol{x}, \boldsymbol{s}) = \frac{1}{\lambda_{\text{MF}}(\boldsymbol{x}, \boldsymbol{s})} \mathbb{E}_{\mathcal{D}}\left[\max_{X \subseteq \mathcal{X}} \text{HV}\left[\boldsymbol{\mu}_\diamond\big(X \mid \mathcal{D}_{(\boldsymbol{x}, \boldsymbol{s})}\big)\right] - \psi_\diamond^*\right],$$

where $\psi_\diamond^* := \max_{X \subseteq \mathcal{X}} \text{HV}\left[\boldsymbol{\mu}_\diamond(X \mid \mathcal{D})\right]$ and $\mathcal{D}_{(\boldsymbol{x}, \boldsymbol{s})} := \mathcal{D} \cup \{(\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{y})\}$. We note that in this general MF-HV-KG formulation, each objective has a (potentially empty) set of fidelity parameters, which can contain (i) fidelity parameters that are unique to that objective, (ii) fidelity parameters that are shared amongst multiple objectives, or (iii) a combination of (i) and (ii).

7. Computing and Optimizing HV-KG

7.1. Hypervolume Computation

To enable efficient optimization, one would like to compute the hypervolume in a differentiable fashion. The joint HV of N_p points can be computed exactly using the inclusion exclusion principle (IEP) (Lopez et al., 2015) and this approach is differentiable with respect to $\mathcal{X}_i'$ (Daulton et al., 2020). The IEP scales exponentially with N_p and therefore is only be feasible for small N_p, but a small N_p tends to work empirically here and for information theoretic approaches. Following Tu et al. (2022), we select $N_p = 10$ (and find that HV-KG is robust to choice of N_p in Appendix D.2).

[4] We clamp difference in HV inside the expectation in (5) to ensure the numerator remains non-negative in the cost-weighted variants. See Appendix C for discussion.

7.2. Unbiased Estimation via Nested Optimization

Although HV-KG cannot be computed analytically, we obtain an unbiased estimator by approximating the outer expectation via Monte Carlo:

$$\hat{\alpha}_{\text{HV-KG}}(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N}\left(\max_{X_i \subseteq \mathcal{X}} \text{HV}\left[\boldsymbol{\mu}(X_i \,|\, \mathcal{D}^i_{\boldsymbol{x}})\right]\right) - \psi^*, \tag{6}$$

where $\mathcal{D}^i_{\boldsymbol{x}} = \mathcal{D} \cup \{\boldsymbol{x}, \boldsymbol{y}^i\}$ with each $\boldsymbol{y}^i$ a realization or "fantasy" sample of the random variable $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{D})$. For each fantasy $\boldsymbol{y}^i$, the updated posterior mean can be computed analytically (Frazier et al., 2008). The inner maximization involves a numerical optimization over a $(N_p \cdot d)$-dimensional space conditional upon the selected $\boldsymbol{x}$. A common approach for solving the nested optimization problems in KG methods is to leverage the envelope theorem to obtain an unbiased gradient estimator (Wu et al., 2017; 2020a) and solve the inner optimization to completion whenever $\boldsymbol{x}$ changes. We derive a gradient of HV-KG in Theorem B.1 in Appendix B.2, which can be estimated without bias via Monte Carlo and optimized via stochastic gradient ascent.

However, solving the inner optimization problem to completion after each outer optimization step is computationally intensive and impractically slow (see Figure 4).

7.3. Deterministic Estimation and Optimization

Instead, we opt for using sample-average (SAA) approximation as Balandat et al. (2020). Using a fixed set of the standard normal *base samples* $\boldsymbol{\epsilon} := \{\epsilon^i\}_{i=1}^N$, $\epsilon^i \in \mathbb{R}^M$ for the fantasized observations $y^{i,(m)} = \mu^{(m)}_{\mathcal{D}}(\boldsymbol{x}) + L^{(m)}_{\mathcal{D}}(\boldsymbol{x})\epsilon^{i,(m)}$, where $L^{(m)}_{\mathcal{D}}$ is the Cholesky factor of the posterior covariance matrix, the fantasies y^i and updated posterior mean functions $\boldsymbol{\mu}(\cdot\,|\mathcal{D}^i_{\boldsymbol{x}})$ are deterministic (see Appendix B.1.2). Given fixed base samples, we can interchange maximization and summation in (6) to obtain

$$\hat{\alpha}_{\text{HV-KG}}(\boldsymbol{x}) = \max_{X_1,\ldots,X_N \subseteq \mathcal{X}} \frac{1}{N}\sum_{i=1}^{N} \text{HV}\left[\boldsymbol{\mu}(X_i \,|\, \mathcal{D}^i_{\boldsymbol{x}})\right] - \psi^*_t. \tag{7}$$

The SAA estimator in (7) can be maximized efficiently by optimizing over $\{\boldsymbol{x}, X_1, \ldots, X_N\}$ simultaneously in "one shot" (Balandat et al., 2020). Although such an approach requires optimizing over a $((N_p \cdot N + 1) \cdot d)$-dimensional space, HV-KG is differentiable with respect to $\boldsymbol{x}, X_1, \ldots, X_N$ and sample-path gradients can be computed via auto-differentiation. Since the SAA estimator is deterministic, (quasi-) second-order gradient-based optimizers can be employed. We can show that the maximizer $\boldsymbol{x}^*_N$ of our SAA estimator converges with probability one to an element of $\mathcal{X}^*_{\text{HV-KG}} = \arg\max_{\boldsymbol{x}\in\mathcal{X}} \alpha_{\text{HV-KG}}(\boldsymbol{x})$, the set of optimizers of the true $\alpha_{\text{HV-KG}}$, and that convergence occurs exponentially fast in the number of MC samples N.

Theorem 7.1. *Suppose that $\mathcal{X}$ is compact and that $\boldsymbol{f} \sim GP(\mu_0(\cdot), K_0(\cdot,\cdot))$ is a sample from a multi-output Gaussian process prior with continuously differentiable mean $\mu_0(\cdot)$ and covariance $K_0(\cdot,\cdot)$ functions. Let $\{\epsilon_i\}_{i=1}^N$ be i.i.d. base samples from $\mathcal{N}(0, I_M)$, let $\boldsymbol{x}^*_N \in \arg\max_{\boldsymbol{x}\in\mathcal{X}} \hat{\alpha}^N_{\text{HV-KG}}(\boldsymbol{x})$, and let $\alpha^*_{\text{HV-KG}} = \max_{\boldsymbol{x}\in\mathcal{X}} \alpha_{\text{HV-KG}}(\boldsymbol{x})$, then*

(i) $\hat{\alpha}_{\text{HV-KG}}(\boldsymbol{x}^*_N) \to \alpha^*_{\text{HV-KG}}$ *a.s.*

(ii) $\inf_{\boldsymbol{x}^*\in\mathcal{X}^*_{\text{HV-KG}}} ||\boldsymbol{x}^*_N - \boldsymbol{x}^*|| \to 0$ *a.s.*

(iii) $\forall\, \delta > 0, \exists\, K < \infty, \alpha > 0$ *such that*

$$p\left(\inf_{\boldsymbol{x}^*\in\mathcal{X}^*_{\text{HV-KG}}} ||\boldsymbol{x}^*_N - \boldsymbol{x}^*|| \geq \delta\right) \leq Ke^{-\alpha N}.$$

We find that optimizing Equation (7) using L-BFGS-B yields strong performance — in both optimization quality and wall time — using the initialization technique described in Appendix A.2. Figure 4 compares acquisition values and wall times for optimizing HV-KG using the stochastic, unbiased gradient estimator and using our deterministic SAA approach. We find that using SAA with determinsitic one-shot optimization finds better candidates than stochastic nested optimization and does so in a fraction of the wall time. See Appendix D.5 for details on the experiment setup.

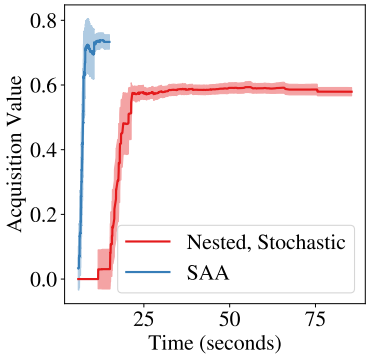

Figure 4: Acquisition optimization using (i) sample average approximation with deterministic one-shot optimization and (ii) nested optimization with stochastic unbiased gradients. For evaluation at each step of gradient-based optimization, we compute the HV-KG at the current design $\boldsymbol{x}$ by solving the inner optimization problem using L-BFGS-B using 32 (stochastic) fantasy samples. We report the mean and two standard errors of the mean across 20 replications.

8. Experiments

We evaluate HV-KG on synthetic and real-world problems including multi-fidelity problems and problems with decoupled evaluations.[5] For all HV-KG variants, we use $N = 32$

[5] Code is available in open source at https://github.com/pytorch/botorch.

fantasies and $N_p = 10$. Because HV-KG is the only acquisition function that handles all cases, we consider differing methods for different types of partial information. As a baseline, we include qNEHVI due to its consistent performance in all our tests, and scrambled Sobol sequences (Owen, 1998) as a quasi-random baseline. In the multi-fidelity case, we include a comparison with MOMF. For decoupled sampling, we compare with two information-based AFs, JES and PFES. JES has been shown to work at least as well as all other ES-based methods (Tu et al., 2022) and can straightforwardly be generalized to the decoupled setting (which is described, but not evaluated in Tu et al. (2022, Appendix M)). PFES has a decoupled variant that had not been evaluated outside of (Suzuki et al., 2020). All AFs are implemented in BoTorch and utilize GPs with a standard Matérn 5/2 kernel over the design space (see Appendix A.1 for additional details).

To compare methods, we first solve (4) by optimizing the posterior means using NSGA-II (Deb et al., 2002) to find the model-estimated Pareto set. Then, we compute the true objective values for the designs in the model-estimated Pareto set and compute the resulting hypervolume dominated by the true Pareto frontier of the model-selected Pareto set. This procedure is common in many works (see e.g. Hernandez-Lobato et al. (2016); Belakaria et al. (2019); Suzuki et al. (2020); Tu et al. (2022)). We report means and ± 2 standard errors of the mean across 20 replications of the log hypervolume regret: the difference in hypervolume between the image of Pareto set identified by the method and the true Pareto frontier.

While the focus of this work is on MOBO with partial information, we also include an evaluation of all applicable methods for the standard noiseless and noisy case with complete information in Appendix D.1. We find that HV-KG performs at least as well as other methods in all test problems considered. Additional details about test problems in the remainder of this section can be found in Appendix A.3.

8.1. Multi-Fidelity

We consider the performance of HV-KG relative to MOMF, a MF MOBO method, as well as the other non-MF baselines with respect to four MF test problems.

Synthetic Problems *(1)* Park ($d = 4$ inputs, $M = 2$ objectives) *(2)* MF Branin Currin ($d = 2, M = 2$) where the cost function is $\lambda(s) = \exp(4.8s)$ (Irshad et al., 2021).

Real-World Problems We consider two problems to highlight the importance of exploiting multi-fidelity information sources: (1) **Laser-plasma acceleration**, ($d = 4, M = 3$) from Irshad et al. (2023a), where a continuous fidelity parameter governs the simulation accuracy and simulation time. (2) **Recommender system ranking policy optimization** ($d = 15, M = 2$) from Liu et al. (2023) simulates a ranking policy which controls the number of items retrieved from different content sources in a recommender system. The target objectives are long-term engagement with the product and content serving cost, and the fidelity parameter is the experiment duration. This problem is designed to mimic setups common to Bayesian optimization of ranking policies with "A/B tests" (Letham and Bakshy, 2019), where selection bias and transient effects bias objectives in the short term (Bakshy et al., 2014).

8.2. Decoupled Evaluation

In the decoupled setting, we compare against three decoupled methods: decoupled PFES (Suzuki et al., 2020), the decoupled extension of JES-LB2 proposed in Tu et al. (2022, Appendix M), and a decoupled variant of Sobol. For Sobol, evaluated objectives are selected uniformly at random, and designs are sampled via scrambled Sobol sequences. We consider both types of coupling: CD, where evaluations occur sequentially, and NCD, where evaluations occur asynchronously.

Synthetic Problems We evaluate performance on the classic ZDT2 ($d = 6, M = 2$) and (Zitzler et al., 2000) and DTLZ2 ($d = 6, M = 2$) (Deb et al., 2002) test problems, and evaluate the objectives in a decoupled fashion. For CD, ZDT1 and DLTZ2 use a cost ratio of 1:3, and for NCD, the objectives have a evaluation time ratio of 1:3, each has a capacity of 1 and equal cost (here we are only concerned with time for NCD).

Real-World Problems We consider two real-world problems: (1) **NAS** ($d = 6, M = 2$) is the neural architecture search problem we use to motivate non-competitive decoupling in the Section 1. The goal is to maximize accuracy and minimize on-device latency for an ImageNet model. Here, we use data from NASBench201 (Dong and Yang, 2020) and HW-NAS-Bench (Li et al., 2021) for the first and second objective, respectively. The NCD version of the problem depicted in Figure 2. The training and latency objectives have an evaluation time ratio of 1:4 and capacities of 2 and 8 respectively and equal cost. For CD, latency and accuracy have costs 1 and 2, respectively. (2) **Vehicle Design** ($d = 5, M = 3$) poses an automotive design problem, where the goal is optimize the design a vehicle to maximize fuel economy, minimize vehicle damage in an off-frontal collision, and minimize passenger trauma in a full frontal crash (Liao et al., 2008). We leverage the surrogate from Tanabe and Ishibuchi (2020) for this problem. For CD, the three objectives have a cost ratio of 1:3:8, and for NCD, the objectives have an evaluation time ratio of 1:3:8 and each objective has an evaluation capacity of 1 and equal cost.

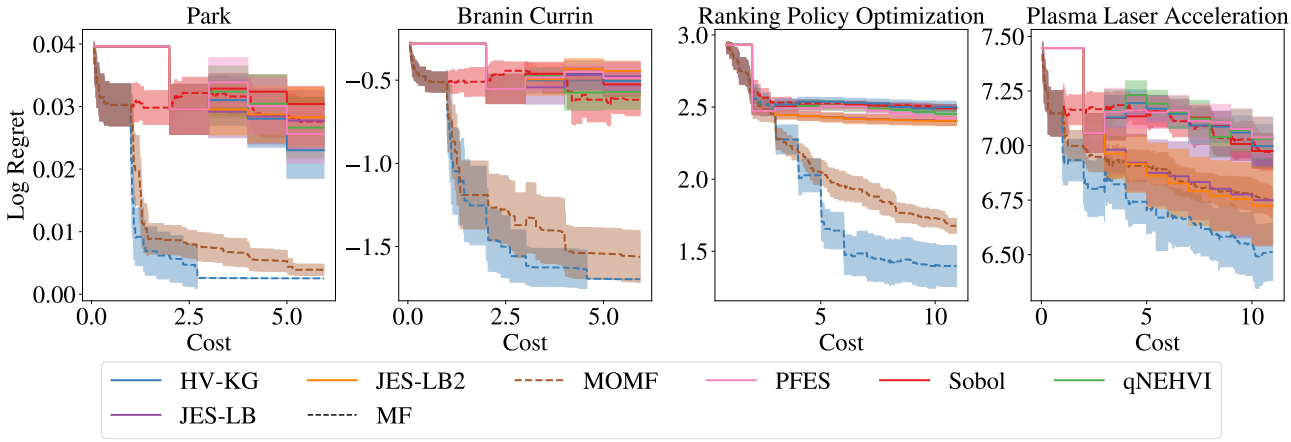

Figure 5: Multi-fidelity optimization performance.

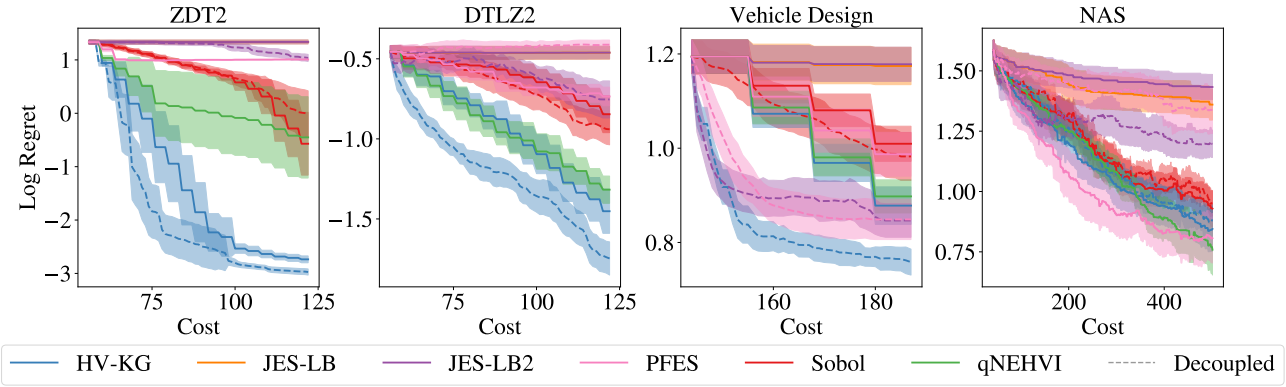

Figure 6: Optimization performance with *competitive decoupling*.

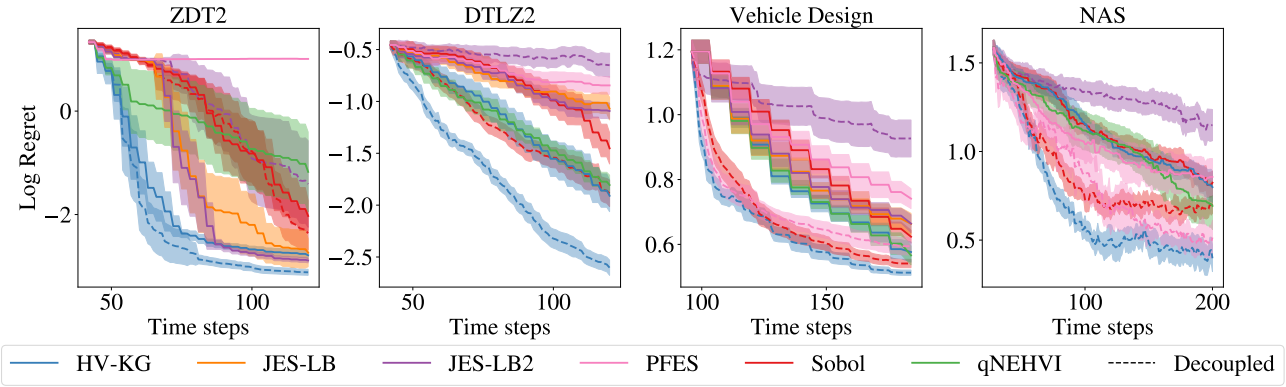

Figure 7: Optimization performance with *non-competitive decoupling*.

8.3. Results

We find that MF-HV-KG and decoupled HV-KG variants significantly improve sample efficiency and optimization performance compared to alternatives. Notably, Sobol baselines that exploit problem structure perform remarkably well on many problems. As shown in Figure 5, MOMF performs well on many MF problems, but is never better than MF-HV-KG. Furthermore, in Figure 17 in Appendix D.6, we show that MOMF consumes the budget much faster with high-fidelity queries due to a misaligned acquisition function where the fidelity is treated as an objective (Irshad et al., 2021). In the decoupled setting, the entropy-based decou-

pled methods struggle on many tasks and we find them to be sensitive to the cost function, whereas HV-KG is more robust across problems and costs (see Appendix D). With CD, HV-KG is again the top performer on 3/4 problems with the exception of NAS as shown in Figure 6. The poor performance of all methods on the NAS CD problem is likely due to poor surrogate model fits; selecting which objective to query in cost-aware fashion depends on having a well-specified and well-calibrated model. On the other hand, in NCD, decoupled methods do not need to rely on the model to select which objective to evaluate and can simply utilize all evaluation capacity. Unsurprisingly, non-decoupled methods perform poorly because they can only generate candidates once all metrics have been evaluated and are limited to the lowest evaluation capacity across all outcomes. Finally, we find (in Appendix D.4) that HV-KG generates candidates faster than entropy-based methods in the decoupled setting.

9. Discussion

HV-KG provides a principled approach to multi-objective Bayesian optimization with incomplete information, including situations in which objective values may be queried separately or at multiple fidelities. To the best of our knowledge, this is the first paper to consider a KG-based approach in the decoupled setting and the multi-objective setting with partial information, and we show that we are able to obtain state-of-the-art performance with respect to standard, decoupled, and multi-fidelity MOBO.

Our work opens the door to exploiting other problems with incomplete observations. Although we exploit MF evaluations, it is possible to leverage more sophisticated models that consider the evolution of objectives over time by leveraging trace observations such as learning curves in AutoML or reinforcement learning (Wu et al., 2020a; Nguyen et al., 2020). The HV-KG approach also lends itself to other instances where incomplete data is available. For example, in contextual BO, we may wish to identify the best configuration or set of best configurations across all contexts, and can transfer knowledge from one context to another. For instance, in the context of on-device AI one may target multiple possible devices and when developing low-carbon-emission concrete, one may wish to develop mixes that are efficient across a variety of environments (e.g., temperature conditions). In other cases, experiments may involve multiple dependent stages, such as cascade and function networks (Astudillo and Frazier, 2021; Kusakawa et al., 2022) common in manufacturing pipelines. HV-KG could be extended to target learning at various stages. In addition, optimiziation improvements might be obtained by exploring alternative approaches handling discrete parameters (e.g. Moss et al. (2020); Jain et al. (2022)). Finally, while the performance of HV-KG is competitive with other methods, the speed of these algorithms might be further improved via exploiting alternative approaches for computing HV (Shang et al., 2022).

Acknowledgements

We would like to thank Peter Frazier, Daniel Jiang, and Michael Osborne for their thoughtful discussions.

References

Robert J. Adler. An introduction to continuity, extrema, and related topics for general gaussian processes. *Lecture Notes-Monograph Series*, 12:i–155, 1990. ISSN 07492170.

Raul Astudillo and Peter Frazier. Bayesian optimization of function networks. *Advances in Neural Information Processing Systems*, 34:14463–14475, 2021.

Johannes Bader and Eckart Zitzler. Hype: An algorithm for fast hypervolume-based many-objective optimization. *Evolutionary Computation*, 19(1):45–76, 2011. doi: 10.1162/EVCO_a_00009.

Eytan Bakshy, Dean Eckles, and Michael S. Bernstein. Designing and deploying online field experiments. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, page 283–292. Association for Computing Machinery, 2014.

Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Advances in Neural Information Processing Systems 33*, 2020.

Laurent Barcelo, John Kline, Gunther Walenta, and Ellis Gartner. Cement and carbon emissions. *Materials and structures*, 47(6):1055–1065, 2014.

R. Bartle. The elements of integration and lebesgue measure. 1995.

Syrine Belakaria, Aryan Deshwal, and Janardhan Rao Doppa. Max-value entropy search for multi-objective Bayesian optimization. In *Advances in Neural Information Processing Systems 32*, 2019.

Syrine Belakaria, Aryan Deshwal, and Janardhan Rao Doppa. Multi-fidelity multi-objective bayesian optimization: An output space entropy search approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34 (06):10035–10043, Apr. 2020. doi: 10.1609/aaai.v34i06.6560.

Karl Bringmann and Tobias Friedrich. Approximation quality of the hypervolume indicator. *Artificial Intelligence*, 195:265–290, 2013. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2012.09.005.

Zefeng Chen, Yuren Zhou, Zhengxin Huang, and Xiaoyun Xia. Towards efficient multiobjective hyperparameter optimization: A multiobjective multi-fidelity bayesian optimization and hyperband algorithm. In *Parallel Problem Solving from Nature – PPSN XVII: 17th International Conference*, page 160–174, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-14713-5. doi: 10.1007/978-3-031-14714-2_12.

Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Differentiable expected hypervolume improvement for parallel multi-objective bayesian optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9851–9864. Curran Associates, Inc., 2020.

Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Parallel Bayesian optimization of multiple noisy objectives with expected hypervolume improvement. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2187–2200. Curran Associates, Inc., 2021.

Samuel Daulton, Sait Cakmak, Maximilian Balandat, Michael A. Osborne, Enlu Zhou, and Eytan Bakshy. Robust multi-objective Bayesian optimization under input noise. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 4831–4866. PMLR, 17–23 Jul 2022a.

Samuel Daulton, David Eriksson, Maximilian Balandat, and Eytan Bakshy. Multi-objective bayesian optimization over high-dimensional search spaces. In James Cussens and Kun Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 507–517. PMLR, 01–05 Aug 2022b.

Samuel Daulton, Xingchen Wan, David Eriksson, Maximilian Balandat, Michael A Osborne, and Eytan Bakshy. Bayesian optimization over discrete and mixed spaces via probabilistic reparameterization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 12760–12774. Curran Associates, Inc., 2022c.

K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002.

Kalyan Deb, L. Thiele, Marco Laumanns, and Eckart Zitzler. Scalable multi-objective optimization test problems. volume 1, 2002.

Xuanyi Dong and Yi Yang. NAS-Bench-201: Extending the scope of reproducible neural architecture search. In *International Conference on Learning Representations (ICLR)*, 2020.

Philippe Duchon, Philippe Flajolet, Guy Louchard, and Gilles Schaeffer. Boltzmann samplers for the random generation of combinatorial structures. *Comb. Probab. Comput.*, 13(4–5):577–625, jul 2004. ISSN 0963-5483. doi: 10.1017/S0963548304006315.

M. T. M. Emmerich, A. H. Deutz, and J. W. Klinkenberg. Hypervolume-based expected improvement: Monotonicity properties and exact computation. In *2011 IEEE Congress of Evolutionary Computation (CEC)*, 2011.

Michael T. M. Emmerich and Carlos M. Fonseca. Computing hypervolume contributions in low dimensions: Asymptotically optimal algorithm and complexity results. In *Evolutionary Multi-Criterion Optimization*, Berlin, Heidelberg, 2011.

David Eriksson, Pierce I-Jen Chuang, Samuel Daulton, Peng Xia, Akshat Shrivastava, Arun Babu, Shicong Zhao, Ahmed Aly, Ganesh Venkatesh, and Maximilian Balandat. Latency-aware neural architecture search with multi-objective bayesian optimization. In *ICML Workshop on AutoML*, 2021.

Stefan Falkner, Aaron Klein, and Frank Hutter. BOHB: Robust and efficient hyperparameter optimization at scale. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1436–1445, 2018.

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

Eduardo C. Garrido-Merchán, Daniel Fernández-Sánchez, and Daniel Hernández-Lobato. Parallel predictive entropy search for multi-objective bayesian optimization with constraints applied to the tuning of machine learning algorithms. *Expert Systems with Applications*, 215: 119328, 2023. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2022.119328.

Daniel Golovin and Qiuyi Zhang. Random hypervolume scalarizations for provable multi-objective black box optimization, 2020.

Julia Guerrero-Viu, Sven Hauns, Sergio Izquierdo, Guilherme Miotto, Simon Schrodi, Andre Biedenkapp, Thomas Elsken, Difan Deng, Marius Lindauer, and Frank Hutter. Bag of baselines for multi-objective joint neural architecture search and hyperparameter optimization. In *ICML 2021 Workshop on AutoML*, 2021.

Youwei He, Jinju Sun, Peng Song, and Xuesong Wang. Variable-fidelity hypervolume-based expected improvement criteria for multi-objective efficient global optimization of expensive functions. *Eng. with Comput.*, 38(4):3663–3689, aug 2022. ISSN 0177-0667. doi: 10.1007/s00366-021-01404-9.

Daniel Hernandez-Lobato, Jose Hernandez-Lobato, Amar Shah, and Ryan Adams. Predictive entropy search for multi-objective Bayesian optimization. In *Proceedings of The 33rd International Conference on Machine Learning*, 2016.

José Miguel Hernández-Lobato, Michael A. Gelbart, Ryan P. Adams, Matthew W. Hoffman, and Zoubin Ghahramani. A general framework for constrained bayesian optimization using information-based search. *Journal of Machine Learning Research*, 17(160):1–53, 2016.

Tito Homem-de-Mello. On rates of convergence for stochastic optimization problems under non-independent and identically distributed sampling. *SIAM Journal on Optimization*, 19(2):524–551, 2008.

Christopher A. Hone, Nicholas Holmes, Geoffrey R. Akien, Richard A. Bourne, and Frans L. Muller. Rapid multistep kinetic model generation from transient flow data. *React. Chem. Eng.*, 2:103–108, 2017. doi: 10.1039/C6RE00109B. URL http://dx.doi.org/10.1039/C6RE00109B.

Andrey Ignatov, Radu Timofte, Andrei Kulik, Seungsoo Yang, Ke Wang, Felix Baum, Max Wu, Lirong Xu, and Luc Van Gool. Ai benchmark: All about deep learning on smartphones in 2019. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3617–3635. IEEE, 2019.

F. Irshad, S. Karsch, and A. Döpp. Multi-objective and multi-fidelity bayesian optimization of laser-plasma acceleration. *Phys. Rev. Res.*, 5:013063, Jan 2023a. doi: 10.1103/PhysRevResearch.5.013063.

Faran Irshad, Stefan Karsch, and Andreas Döpp. Expected hypervolume improvement for simultaneous multi-objective and multi-fidelity optimization, 2021.

Faran Irshad, Stefan Karsch, and Andreas Doepp. Reference dataset of multi-objective and multi- fidelity optimization in laser-plasma acceleration, January 2023b.

Moksh Jain, Sharath Chandra Raparthy, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Yoshua Bengio, Santiago Miret, and Emmanuel Bengio. Multi-objective gflownets, 2022.

Vijay Janapa Reddi, David Kanter, Peter Mattson, Jared Duke, Thai Nguyen, Ramesh Chukka, Ken Shiring, Koan-Sin Tan, Mark Charlebois, William Chou, et al. Mlperf mobile inference benchmark: An industry-standard open-source machine learning benchmark for on-device ai. *Proceedings of Machine Learning and Systems*, 4:352–369, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

J. Knowles. Parego: a hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation*, 10(1):50–66, 2006.

Shunya Kusakawa, Shion Takeno, Yu Inatsu, Kentaro Kutsukake, Shogo Iwazaki, Takashi Nakano, Toru Ujihara, Masayuki Karasuyama, and Ichiro Takeuchi. Bayesian Optimization for Cascade-Type Multistage Processes. *Neural Computation*, 34(12):2408–2431, 11 2022. ISSN 0899-7667. doi: 10.1162/neco_a_01550.

Eric Hans Lee, Valerio Perrone, Cedric Archambeau, and Matthias Seeger. Cost-aware Bayesian Optimization. *arXiv e-prints*, page arXiv:2003.10870, March 2020.

Benjamin Letham and Eytan Bakshy. Bayesian optimization for policy search via online-offline experimentation. *Journal of Machine Learning Research*, 20(145):1–30, 2019.

Chaojian Li, Zhongzhi Yu, Yonggan Fu, Yongan Zhang, Yang Zhao, Haoran You, Qixuan Yu, Yue Wang, Cong Hao, and Yingyan Lin. {HW}-{nas}-bench: Hardware-aware neural architecture search benchmark. In *International Conference on Learning Representations*, 2021.

Qiaohao Liang and Lipeng Lai. Scalable bayesian optimization accelerates process optimization of penicillin production. In *NeurIPS 2021 AI for Science Workshop*, 2021.

Xingtao Liao, Qing Li, Xujing Yang, Weigang Zhang, and Wei Li. Multiobjective optimization for crash safety design of vehicles using stepwise regression model. *Structural and Multidisciplinary Optimization*, 35, 2008.

Sulin Liu, Qing Feng, David Eriksson, Benjamin Letham, and Eytan Bakshy. Sparse bayesian optimization. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, AISTATS, page *In press*, 2023.

Edgar Manoatl Lopez, Luis Miguel Antonio, and Carlos A. Coello Coello. A gpu-based algorithm for a faster hypervolume contribution computation. In António Gaspar-Cunha, Carlos Henggeler Antunes, and Carlos Coello Coello, editors, *Evolutionary Multi-Criterion Optimization*, pages 80–94. Springer International Publishing, 2015.

Mina Konakovic Lukovic, Yunsheng Tian, and Wojciech Matusik. Diversity-Guided Multi-Objective Bayesian Optimization With Batch Evaluations. In *Advances in Neural Information Processing Systems 33*, 2020.

Richard M Meyer. *Essential mathematics for applied fields*. Springer Science & Business Media, 2012.

Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

Henry Moss, David Leslie, Daniel Beck, Javier González, and Paul Rayson. Boss: Bayesian optimization over string spaces. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15476–15486. Curran Associates, Inc., 2020.

Fatemah Mukadum, Quan Nguyen, Daniel M. Adrion, Gabriel Appleby, Rui Chen, Haley Dang, Remco Chang, Roman Garnett, and Steven A. Lopez. Efficient discovery of visible light-activated azoarene photoswitches with long half-lives using active search. *Journal of Chemical Information and Modeling*, 61(11):5524–5534, 2021. doi: 10.1021/acs.jcim.1c00954.

Vu Nguyen, Sebastian Schulze, and Michael Osborne. Bayesian optimization for iterative learning. *Advances in Neural Information Processing Systems*, 33:9361–9371, 2020.

Art B Owen. Scrambling sobol'and niederreiter–xing points. *Journal of complexity*, 14(4):466–489, 1998.

Ji Won Park, Samuel Don Stanton, Saeed Saremi, Andrew Martin Watkins, Henri Dwyer, Vladimir Gligorijevic, Richard Bonneau, Stephen Ra, and Kyunghyun Cho. PropertyDAG: Multi-objective bayesian optimization of partially ordered, mixed-variable properties for biological sequence design. In *NeurIPS 2022 AI for Science: Progress and Promises*, 2022.

Michael Parsons and Randall Scott. Formulation of multicriterion design optimization problems for solution with scalar numerical optimization methods. *Journal of Ship Research*, 48:61–76, 03 2004. doi: 10.5957/jsr.2004.48.1.61.

Matthias Poloczek, Jialei Wang, and Peter Frazier. Multi-information source optimization. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Saba Q. Yahyaa, Madalina M. Drugan, and Bernard Manderick. Knowledge gradient for multi-objective multi-armed bandit algorithms. In *Proceedings of the 6th International Conference on Agents and Artificial Intelligence - Volume 1*, ICAART 2014, page 74–83, Setubal, PRT, 2014. SCITEPRESS - Science and Technology Publications, Lda. ISBN 9789897580154. doi: 10.5220/0004796600740083.

Carl Edward Rasmussen. *Gaussian Processes in Machine Learning*, pages 63–71. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

Warren Scott, Peter Frazier, and Warren Powell. The correlated knowledge gradient for simulation optimization of continuous parameters using gaussian process regression. *SIAM Journal on Optimization*, 21:996–1026, 07 2011. doi: 10.1137/100801275.

Ke Shang, Weiyu Chen, Weiduo Liao, and Hisao Ishibuchi. Hv-net: Hypervolume approximation based on deepsets. *IEEE Transactions on Evolutionary Computation*, pages 1–1, 2022. doi: 10.1109/TEVC.2022.3181306.

Benjamin J Shields, Jason Stevens, Jun Li, Marvin Parasram, Farhan Damani, Jesus I Martinez Alvarado, Jacob M Janey, Ryan P Adams, and Abigail G Doyle. Bayesian reaction optimization as a tool for chemical synthesis. *Nature*, 590(7844):89–96, 2021.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

Shinya Suzuki, Shion Takeno, Tomoyuki Tamura, Kazuki Shitara, and Masayuki Karasuyama. Multi-objective Bayesian optimization using pareto-frontier entropy. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, 2020.

Shion Takeno, Hitoshi Fukuoka, Yuhki Tsukada, Toshiyuki Koyama, Motoki Shiga, Ichiro Takeuchi, and Masayuki Karasuyama. Multi-fidelity Bayesian optimization with max-value entropy search and its parallelization. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9334–9345. PMLR, 13–18 Jul 2020.

Ryoji Tanabe and Hisao Ishibuchi. An easy-to-use real-world multi-objective optimization problem suite. *Applied Soft Computing*, 89, 2020.

Jose Antonio Garrido Torres, Sii Hong Lau, Pranay Anchuri, Jason M. Stevens, Jose E. Tabora, Jun Li, Alina Borovika, Ryan P. Adams, and Abigail G. Doyle. A multi-objective active learning platform and web app for reaction optimization. *Journal of the American Chemical Society*, 144(43):19999–20007, 2022. doi: 10.1021/jacs.2c08592. PMID: 36260788.

Ben Tu, Axel Gandy, Nikolas Kantas, and Behrang Shafei. Joint entropy search for multi-objective bayesian optimization. In *Advances in Neural Information Processing Systems 35*, 2022.

Colin White, Willie Neiswanger, and Yash Savani. Bananas: Bayesian optimization with neural architectures for neural architecture search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10293–10301, May 2021. doi: 10.1609/aaai.v35i12.17233.

Jian Wu and Peter Frazier. The parallel knowledge gradient method for batch bayesian optimization. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Jian Wu, Matthias Poloczek, Andrew G Wilson, and Peter Frazier. Bayesian optimization with gradients. In *Advances in Neural Information Processing Systems*, pages 5267–5278, 2017.

Jian Wu, Saul Toscano-Palmerin, Peter I. Frazier, and Andrew Gordon Wilson. Practical multi-fidelity bayesian optimization for hyperparameter tuning. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 788–798. PMLR, 22–25 Jul 2020a.

Tony C. Wu, Daniel Flam-Shepherd, and Alán Aspuru-Guzik. Bayesian variational optimization for combinatorial spaces, 2020b.

Eckart Zitzler, Kalyanmoy Deb, and Lothar Thiele. Comparison of multiobjective evolutionary algorithms: Empirical results. *Evol. Comput.*, 8(2):173–195, jun 2000. ISSN 1063-6560. doi: 10.1162/106365600568202.

A. Experiment Details

A.1. Implementation of Acquisition Functions and Models

We use the BoTorch implementations of qNEHVI (Daulton et al., 2020), MOMF (Irshad et al., 2021), JES-LB and JES-LB2 (Tu et al., 2022) developed by the original authors of these works.[5] To determine the Pareto frontier, we use Tu et al. (2022)'s NSGA-II-based implementation.[6] For PFES (Suzuki et al., 2020), we utilize Tu et al. (2022)'s open source implementation in BoTorch, which includes the lower-bound batch variant (for $q > 1$). For decoupled sampling, we modified the existing BoTorch implementation of Suzuki et al. (2020) to include the decoupled approach from the original paper, and we implemented the extension of JES-LB2 to the decoupled setting, proposed in Tu et al. (2022, Appendix M).

For all MC acquisition functions, we use quasi-random (QMC) base samples and sample average approximation (Balandat et al., 2020). All methods are optimized using L-BFGS-B from 20 starting points using the default initialization heuristic in BoTorch (Balandat et al., 2020)—except for HV-KG, which we optimize from a single starting point to limit computational overhead. All methods use independent GPs with ARD Matérn 5/2 kernels. We use Gamma(2,2) priors over the lengthscales (with allow learning large lengthscales for irrelevant parameters) and Gamma(2, 0.15) priors on the outputscales. We assume that the noise level is known.

For the NAS and chemistry problems, we one-hot encode each categorical x with C categories as $\boldsymbol{x}' = [x'_1, ..., x'_C] \in [0, 1]^C$, apply exact discretization functions (i.e., $x = \text{ONE-HOT}(\arg\max_{c \in C} x'_c)$) before evaluating the GP, and use straight through-gradient gradient estimators (Daulton et al., 2022c).

A.2. Initialization of HV-KG

The optimization of HV-KG can be significantly sped up by choosing good initial conditions for the design point and the fantasy optimizers. If we assume that the additional observation $(\boldsymbol{x}, \boldsymbol{y})$ does not drastically change the location of the Pareto set in input space, then solving the optimization problem

$$\max_{X \subseteq \mathcal{X}} \text{HV}\Big[\boldsymbol{\mu}(X \mid \mathcal{D})\Big] \tag{8}$$

under the current posterior (having observed data $\mathcal{D}$) will yield an optimizer that will likely be quite close to the optimal $\{X_i\}_{i=1}^N$ after fantasizing about the unknown function values $\boldsymbol{y}$. Equation (8) can be solved efficiently using gradient-based optimization. Using different starting points, we can identify N solution sets X to use as initial values for the HV-KG optimization problem. Lastly, we can find starting points for $\boldsymbol{x}$ conditional on $(X_1, \dots, X_N)$ using standard BO initialization heuristics such as Boltzmann sampling on the HV-KG values (Balandat et al., 2020). We use the resulting starting point $(\boldsymbol{x}, X_1, \dots, X_N)$ to optimize HV-KG via quasi-second order methods (L-BFFS-B) using SAA.[7]

A.3. Problem Details

All noisy variants use additive zero-mean Gaussian noise, where the noise standard deviations (denoted by σ) are set as a percentage of the range of each objective as indicated in parentheses. These noise levels come from in previous works (Daulton et al., 2021; Tu et al., 2022).

We use the multi-fidelity versions of Park ($M = 2$ objectives, $d = 4$ inputs, $K = 1$ fidelity parameters) and Branin-Currin ($M = 2, d = 2, K = 1$) from Irshad et al. (2021), the Penicillin manufacturing problem ($M = 3, d = 7, \sigma = 1\%$) from Liang and Lai (2021),[8]. When considering the standard test problems DTLZ2 ($M = 2, d = 6, \sigma = 10\%$) (Deb et al., 2002), ZDT2 ($M = 2, d = 6, \sigma = 10\%$) (Zitzler et al., 2000), and Vehicle Design ($M = 3, d = 5, \sigma = 1\%$) (Tanabe and Ishibuchi, 2020) all of which are implemented in BoTorch.[5]

We use implementations of the **Marine** design problem ($M = 4, d = 6, \sigma = 3\%$) (Parsons and Scott, 2004; Tanabe and

[6] Code is available in open source at https://github.com/benmltu/JES.

[7] In the case that the objectives are not modeled directly and do not have analytic expressions in terms of the model outputs, the inner expectation could also be approximated with Monte Carlo samples. Such cases would arise in the case of constrained optimization where the goal is to optimize feasibility-weighted objectives and unweighted objectives and constraint slacks are both modeled and subsequently combined.

[8] We modify the search space slightly to make the simulations less bimodal (as identified in Park et al. (2022)) by reducing the number of designs that lead to a zero fermentation time objective. The modified search space sets the lower bounds of the 4^{th} and 5^{th} parameters to be 4 and $\frac{1}{4}$, respectively.

Table 1: Reference points for all benchmark problems (assuming minimization of all objectives). In our benchmarks, we maximize all objectives by multiplying objectives and reference points by -1.

PROBLEM	REFERENCE POINT
ZDT2	(11, 11)
DTLZ2	(1.1,1.1)
VEHICLE DESIGN	(1698.55, 11.21, 0.29)
NAS	(-7.319, 30.847)
PARK	(0,0)
BRANIN-CURRIN	(0,0)
RANKING POLICY OPTIMIZATION	(5.353, -44.39)
PLASMA LASER ACCELERATION	(280.864, -50.613, -36.412)
PENICILLIN	(-5.657, 64.1, 340.0)
MARINE	(-250, $2 \cdot 10^4$, $2.5 \cdot 10^4$, 15)
SNAR	(-5.5, 5)
CHEMISTRY	(32.669, -0.107)

Ishibuchi, 2020) and **SnAr** ($M = 2, d = 4\sigma = 3\%$, a chemical reaction optimization problem) (Hone et al., 2017) from Tu et al. (2022).[6]

Chemistry problem aims to tune experimental conditions to maximize chemical reaction yield while minimizing cost ($M = 2, d = 5$). We adopt this problem from Daulton et al. (2022c) (*Direct Arylation Chemical Synthesis*). A GP surrogate is fit to chemical reaction data from Shields et al. (2021),[9] and corresponding reaction cost data[10] from Torres et al. (2022)

NAS problem ($M = 2, d = 6$) uses accuracy data from NASBench201[11] (Dong and Yang, 2020), augmented with edge GPU latency estimates from HW-NAS-Bench[12] (Li et al., 2021).

Vehicle Design problem ($d = 5, M = 3$) poses a hypothetical automotive problem. We leverage the surrogate from Tanabe and Ishibuchi (2020) and formulate the problem the problem with respect to the surrogate in the following way: we minimize mass (a proxy for maximizing fuel economy), minimize length of toe-box intrusion in case of a crash (a proxy for vehicle damage), and minimize acceleration (a proxy for passenger trauma), vehicle damage in an off-frontal collision (measure in a by toe-box intrusion distance), and minimize acceleration (a proxy for passenger trauma in a full frontal crash) (Liao et al., 2008). This problem can most naturally be thought of as a NCD problem, since the evaluation of the last two objectives as destructive, so that each objective requires a different type of collision. The fuel economy objective is less costly to evaluate, as it does not require manufacturing and crashing a car. For CD, the three objectives have a cost ratio of 1:3:8, and for NCD, the objectives have an evaluation time ratio of 1:3:8 and each objective has an evaluation capacity of 1 and equal cost.

Recommendation System Ranking Policy Optimization We use variant of the ranking policy ($M = 2, d = 15, K = 1$) optimization problem system from (Liu et al., 2023). To create a multi-fidelity variant of this problem, we add bias term to emulate the "novelty effect", an ephemeral boost in engagement, that commonly affects engagement metrics when new ranking policies conducted via "A/B tests" (Bakshy et al., 2014). Running longer experiments (high fidelity experiments), will reduce the novelty effect and provide more accurate estimates of the long term effect. We use the same search space as in Liu et al. (2023), but restricted to 15 dimensions.

Plasma Laser Acceleration The plasma laser acceleration problem comes from the recent work by Irshad et al. (2023a). We fit GP surrogate models to the data (Irshad et al., 2023b) collected by the original authors via simulations.[13]

A.4. Initial Point Selection for Multi-Fidelity Experiments

The cost budget for selecting initial design points is set equal to the cost of 2 full-fidelity evaluations $2\lambda(s)$. Full fidelity methods sample 2 designs from a scrambled Sobol sequence. Multi-fidelity methods sample the design parameters uniformly at random and the fidelity parameter is sampled (via the inverse transform) from the probability distribution with pdf

[9]Data is available at https://github.com/b-shields/edbo.
[10]Data is available at https://github.com/doyle-lab-ucla/edboplus.
[11]Code is available at https://github.com/D-X-Y/NAS-Bench-201.
[12]Code is available at https://github.com/GATECH-EIC/HW-NAS-Bench/.
[13]Data is available at https://doi.org/10.5281/zenodo.7565882.

$p(s) \propto \frac{1}{\lambda(s)}$. Designs are added until the next sampled point exceeds the cost budget. Hence, multi-fidelity methods use an initialization with cost $\leq 2\lambda(s)$.

B. Theoretical Results

B.1. Preliminaries

B.1.1. Hypervolume Computation

For a set of N_p points $\mathcal{Y} = \{y_j\}_{j=1}^{N_p}$, the HV w.r.t to a reference point r can be computed in a differentiable fashion (Daulton et al., 2020) as

$$\text{HV}(\mathcal{Y}, r) = \sum_{j=1}^{N_p} \sum_{Y_j \in \mathcal{Y}_j} (-1)^{j+1} \prod_{m=1}^{M} \left[z_{Y_j}^{(m)} - r^{(m)}\right]_+, \tag{9}$$

where $\mathcal{Y}_j := \{Y_j \subseteq \mathcal{Y} : |Y_j| = j\}$ is the set of all subsets of $\mathcal{Y}$ of size j and $z_{Y_j}^{(m)} := \min\left[y_{i_1}^{(m)}, \dots, y_{i_j}^{(m)}\right]$ for $Y_j = \{y_{i_1}, \dots, y_{i_j}\}$.

B.1.2. Gaussian Processes

In this work, we place independent Gaussian process priors on the different objectives. In this section we therefore restrict ourselves to modeling a single objective $f \sim GP(\mu_0, K_0)$, where $\mu_0 : \mathcal{X} \to \mathbb{R}$ is the prior function (assumed to be constant) and $K_0 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the the prior covariance function. We assume that observations of the objectives are subject to iid zero-mean Gaussian noise with variance σ^2. Then after conditioning on $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ observations, the mean and covariance functions conditioned on $\mathcal{D}$ at a set of points $\boldsymbol{x}$ are given by (Rasmussen, 2004)

$$\mu_{\mathcal{D}}(\boldsymbol{x}) = \mu_0(\boldsymbol{x}) + K_0(\boldsymbol{x}, \boldsymbol{x}_{1:n})[K_0^\sigma(\boldsymbol{x}_{1:n}, \boldsymbol{x}_{1:n})]^{-1}[y_{1:n} - \mu_0(\boldsymbol{x}_{1:n})]$$

$$K_{\mathcal{D}}(\boldsymbol{x}, \boldsymbol{x}') = K_0(\boldsymbol{x}, \boldsymbol{x}') - K_0(\boldsymbol{x}, \boldsymbol{x}_{1:n})[K_0^\sigma(\boldsymbol{x}_{1:n}, \boldsymbol{x}_{1:n})]^{-1} K_0(\boldsymbol{x}_{1:n}, \boldsymbol{x}'),$$

where $\boldsymbol{x}_{1:n} := \{x_1, \dots, x_n\}$, $K_{\mathcal{D}}^\sigma(\boldsymbol{x}_{1:n}, \boldsymbol{x}_{1:n})$ denotes $K_{\mathcal{D}}(\boldsymbol{x}_{1:n}, \boldsymbol{x}_{1:n}) + \text{diag}(\sigma^2(x_1), ..., \sigma^2(x_n))$.

In this work, we are often interested in fantasization; i.e. fantasizing about the observations $\boldsymbol{y}$ that we would receive if we were to evaluate $\boldsymbol{x}$. In this case, $\boldsymbol{y}$ is a random vector, which according to our beliefs is $y \sim \mathcal{N}(\mu_{\mathcal{D}}(\boldsymbol{x}), K_{\mathcal{D}}(\boldsymbol{x}, \boldsymbol{x}) + \sigma^2(\boldsymbol{x}))$. Then conditioned on evaluating $\boldsymbol{x}$ and observing $\boldsymbol{y}$, the updated posterior mean function would be

$$\mu_{\mathcal{D}_{\boldsymbol{x}}}(\boldsymbol{x}') = \mu_{\mathcal{D}}(\boldsymbol{x}') + K_{\mathcal{D}}(\boldsymbol{x}', \boldsymbol{x})[K_{\mathcal{D}}^\sigma(\boldsymbol{x}, \boldsymbol{x})]^{-1}[\boldsymbol{y} - \mu_{\mathcal{D}}(\boldsymbol{x})].$$

As in previous works, it is convenient to express the updated mean in terms of a standard normal random variable (Wu et al., 2020a; Wu and Frazier, 2016). We can rewrite $K_{\mathcal{D}}^\sigma(\boldsymbol{x}, \boldsymbol{x})$ in terms of its Cholesky factors $K_{\mathcal{D}}^\sigma(\boldsymbol{x}, \boldsymbol{x}) = L_{\mathcal{D}}(\boldsymbol{x}) L_{\mathcal{D}}(\boldsymbol{x})^T$. So $[K_{\mathcal{D}}^\sigma(\boldsymbol{x}, \boldsymbol{x})]^{-1} = (L_{\mathcal{D}}(\boldsymbol{x})^T)^{-1} L_{\mathcal{D}}(\boldsymbol{x})^{-1}$. Since $[\boldsymbol{y} - \mu_{\mathcal{D}}(\boldsymbol{x})] \sim \mathcal{N}(0, K_{\mathcal{D}}^\sigma(\boldsymbol{x}, \boldsymbol{x}))$, $L_{\mathcal{D}}(\boldsymbol{x})^{-1}[\boldsymbol{y} - \mu_{\mathcal{D}}(\boldsymbol{x})]$ is a standard normal random vector. Letting $\hat{\Sigma}_{\mathcal{D}}(\boldsymbol{x}', \boldsymbol{x}) := K_{\mathcal{D}}(\boldsymbol{x}', \boldsymbol{x})(L_{\mathcal{D}}(\boldsymbol{x})^T)^{-1}$, we can express the update posterior mean as

$$\mu_{\mathcal{D}_{\boldsymbol{x}}}(\boldsymbol{x}') = \mu_{\mathcal{D}}(\boldsymbol{x}') + \hat{\Sigma}_{\mathcal{D}}(\boldsymbol{x}', \boldsymbol{x})\epsilon,$$

where ϵ is a standard normal random vector.

B.2. Proofs

Without loss of generality, we consider case with a batch size $q = 1$ (i.e. $\boldsymbol{x} = \{x\}$). Since $\boldsymbol{x}$ only affects the new data $\mathcal{D}_{\boldsymbol{x}}$ that the model is conditioned on, partial derivatives can be computed for all $q \cdot d$ elements of $\boldsymbol{x}$ and extending the results that follow is straightforward.[14] Moreover, for brevity we only consider (iid) Monte Carlo sampling in this section. Balandat et al. (2020) also prove basic results for SAA using (randomized) quasi-Monte Carlo (RQMC) sampling; leveraging those results the proofs in this section can be extended to the RQMC setting in a straightforward fashion.

At a high level, we derive a gradient estimator and prove that it is unbiased (Theorem B.1) by building upon the work of Wu et al. (2020b) and leveraging our proof that value function in HV-KG is Lipschitz continuous (Lemma B.1). Then, we prove

[14] Note that there is a minor complication if $\boldsymbol{x}$ contains duplicate points as the posterior mean will not be differentiable at such points. However, the set of such points is of measure zero and so does not affect the derivations and the results below.

our main result (Theorem 7.1), which proves three convergence properties of our SAA estimator, building upon work from Balandat et al. (2020) and leveraging Lemma B.2.

Lemma B.1. *For a fixed X, let $A(\boldsymbol{x}, \boldsymbol{\epsilon}) := \mathrm{HV}\big[\boldsymbol{\mu}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X)\big]$, where $\boldsymbol{\mu}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X) := \big[\mu^{(1)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X), \dots, \mu^{(M)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X)\big]$, $\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X) := \mu^{(m)}_{\mathcal{D}}(X) + \hat{\Sigma}^{(m)}_{\mathcal{D}}(X, \boldsymbol{x})\epsilon^{(m)}$ for $m = 1, \dots M$, and $\boldsymbol{\epsilon} := [\epsilon^{(1)} ... , \epsilon^{(M)}]$. Then, $A(\boldsymbol{x}, \boldsymbol{\epsilon})$ is Lipschitz continuous with respect to $\boldsymbol{x}$ for any given $\boldsymbol{\epsilon}$.*

Proof. Note that

$$A(\boldsymbol{x}, \boldsymbol{\epsilon}) = \mathrm{HV}\big[\boldsymbol{\mu}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X)\big] = \sum_{j=1}^{N_p} \sum_{X_j \in \mathbb{X}_j} (-1)^{j+1} \prod_{m=1}^{M} \big[\min[\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_1}), \dots, \mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_j})] - r^{(m)}\big]_+,$$

where $\mathbb{X}_j := \{X_j \subseteq X : |X_j| = j\}$ is the set of all subsets of X of size j.

We wish to show that there exists a function $l : \mathbb{R}^M \to \mathbb{R}$ such that $|A(\boldsymbol{x}, \boldsymbol{\epsilon}) - A(\boldsymbol{y}, \boldsymbol{\epsilon})| \leq l(\boldsymbol{\epsilon})||\boldsymbol{x} - \boldsymbol{y}||$.

Let $\tilde{a}_{m,j,X_j}(\boldsymbol{x}, \boldsymbol{\epsilon}) = \big[\min[\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_1}), ..., \mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_j})] - r^{(m)}\big]_+$ and let $\tilde{A}_{j,X_j}(\boldsymbol{x}, \boldsymbol{\epsilon}) = \prod_{m=1}^{M} \tilde{a}_{m,j,X_j}(\boldsymbol{x}, \boldsymbol{\epsilon})$. Since $A(\boldsymbol{x}, \boldsymbol{\epsilon}) = \sum_{j=1}^{N_p} \sum_{X_j \in \mathcal{X}_j} (-1)^{j+1} \tilde{A}_{j,X_j}(\boldsymbol{x}, \boldsymbol{\epsilon})$, it suffices to show that there exists a function $l : \mathbb{R}^M \to \mathbb{R}$ such that $|\tilde{A}_{j,X_j}(\boldsymbol{x}, \boldsymbol{\epsilon}) - \tilde{A}_{j,X_j}(\boldsymbol{y}, \boldsymbol{\epsilon})| \leq l(\boldsymbol{\epsilon})||\boldsymbol{x} - \boldsymbol{y}||$.

We have that

$$\begin{aligned}
\tilde{a}_{m,j,X_j}(\boldsymbol{x}, \boldsymbol{\epsilon}) &= \big[\min[\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_1}), ..., \mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_j})] - r^{(m)}\big]_+ \\
&\leq |r^{(m)}| + \big|\min[\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_1}), ..., \mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_j})]\big| \\
&\leq |r^{(m)}| + \sum_{k=1}^{j} \big|\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_k})\big|.
\end{aligned}$$

Note that for a given $\boldsymbol{\epsilon}$, $\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X)$ is continuously differentiable with respect to $\boldsymbol{x}$ for any fixed X and continuously differentiable w.r.t to X for any $\boldsymbol{x}$ because $\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X) = \mu^{(m)}_{\mathcal{D}}(X) + \hat{\Sigma}^{(m)}_{\mathcal{D}}(X, \boldsymbol{x})\epsilon^{(m)}$ and $\mu^{(m)}_{\mathcal{D}}(X)$ and $\hat{\Sigma}^{(m)}_{\mathcal{D}}(X, \boldsymbol{x})$ are continuously differentiable with respect to $\boldsymbol{x}$ (Wu et al., 2020a).[15] Note that $|\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X)| \leq ||\mu^{(m)}_{\mathcal{D}}(X)|| + ||\hat{\Sigma}^{(m)}_{\mathcal{D}}(X, \boldsymbol{x})|| \cdot |\epsilon^{(m)}|$. Since $\mu^{(m)}_{\mathcal{D}}(X)$ and $\hat{\Sigma}^{(m)}_{\mathcal{D}}(X, \boldsymbol{x})$ are uniformly bounded for each $m = 1, \dots, M$, there exist $C^{(m)}_1, C^{(m)}_2 \in \mathbb{R}$ such that $|\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X)| \leq C^{(m)}_1 + C^{(m)}_2|\epsilon^{(m)}|$ for each $m = 1, \dots, M$. Hence, $|\tilde{a}_{m,j,X_j}(\boldsymbol{x}, \boldsymbol{\epsilon})| \leq |r^{(m)}| + j(C^{(m)}_1 + C^{(m)}_2|\epsilon^{(m)}|)$.

Omitting the subscripts j, X_j for brevity, and considering $M = 2$ for now, we have that

$$\begin{aligned}
|\tilde{A}_{j,X_j}(\boldsymbol{x}, \boldsymbol{\epsilon}) - \tilde{A}_{j,X_j}(\boldsymbol{y}, \boldsymbol{\epsilon})| &= \big|\tilde{a}_1(\boldsymbol{x}, \boldsymbol{\epsilon})\tilde{a}_2(\boldsymbol{x}, \boldsymbol{\epsilon}) - \tilde{a}_1(\boldsymbol{y}, \boldsymbol{\epsilon})\tilde{a}_2(\boldsymbol{y}, \boldsymbol{\epsilon})\big| &(10)\\
&= \big|\tilde{a}_1(\boldsymbol{x}, \boldsymbol{\epsilon})((\tilde{a}_2(\boldsymbol{x}, \boldsymbol{\epsilon}) - \tilde{a}_2(\boldsymbol{y}, \boldsymbol{\epsilon})) + \tilde{a}_2(\boldsymbol{y}, \boldsymbol{\epsilon})(\tilde{a}_1(\boldsymbol{x}, \boldsymbol{\epsilon}) - \tilde{a}_1(\boldsymbol{y}, \boldsymbol{\epsilon}))\big| &(11)\\
&\leq \big|\tilde{a}_1(\boldsymbol{x}, \boldsymbol{\epsilon})\big|\big|\tilde{a}_2(\boldsymbol{x}, \boldsymbol{\epsilon}) - \tilde{a}_2(\boldsymbol{y}, \boldsymbol{\epsilon})\big| + \big|\tilde{a}_2(\boldsymbol{y}, \boldsymbol{\epsilon})\big|\big|\tilde{a}_1(\boldsymbol{x}, \boldsymbol{\epsilon}) - \tilde{a}_1(\boldsymbol{y}, \boldsymbol{\epsilon})\big|. &(12)
\end{aligned}$$

Note that

$$\begin{aligned}
&|a_{m,j,X_j}(\boldsymbol{x}, \boldsymbol{\epsilon}) - a_{m,j,X_j}(\boldsymbol{y}, \boldsymbol{\epsilon})| \\
&\quad = \Big|\big[\min[\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(\boldsymbol{x}_{i_1}), ..., \mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(\boldsymbol{x}_{i_j})] - r^{(m)}\big]_+ - \big[\min[\mu^{(m)}_{\boldsymbol{y},\boldsymbol{\epsilon}}(\boldsymbol{x}_{i_1}), ..., \mu^{(m)}_{\boldsymbol{y},\boldsymbol{\epsilon}}(\boldsymbol{x}_{i_j})] - r^{(m)}\big]_+\Big|.
\end{aligned}$$

For brevity, we assume without loss of generality that $r = 0$ (otherwise this is just a constant shift in the means μ).

Case 1: If both terms are zero, then $|a_{m,j,X_j}(\boldsymbol{x}, \boldsymbol{\epsilon}) - a_{m,j,X_j}(\boldsymbol{y}, \boldsymbol{\epsilon})| = 0$.

[15] Technically, this is only true if the noise terms $\{\sigma^2(X_i)\}_{i=1}^n$ are strictly positive; otherwise $\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X)$ is not differentiable if $K^{\sigma}_0(\boldsymbol{x}_{1:n}, \boldsymbol{x}_{1:n})$ is singular. However, even in this case that happens only on a set of measure zero, and thus our arguments remain valid in the almost everywhere sense.

Case 2: Suppose that one of the terms inside of $[\cdot]_+$ is greater than 0 and one term is less than zero. Without loss of generality suppose that $\min[\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_1}), ..., \mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_j})] \leq 0$ and $\min[\mu^{(m)}_{\boldsymbol{y},\boldsymbol{\epsilon}}(X_{i_1}), ..., \mu^{(m)}_{\boldsymbol{y},\boldsymbol{\epsilon}}(X_{i_j})] \geq 0$. Let $k = \arg\min_{k=1,...,j} \mu^{(m)}_{\boldsymbol{y},\boldsymbol{\epsilon}}(X_{i_k})$. Recall that $\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X) = \mu^{(m)}_{\mathcal{D}}(X) + \hat{\Sigma}^{(m)}_{\mathcal{D}}(X, \boldsymbol{x})\epsilon^{(m)}$ and $\hat{\Sigma}^{(m)}_{\mathcal{D}}(X, \boldsymbol{x})$ is continuously differentiable with respect to $\boldsymbol{x}$ (Wu et al., 2020a), so they are Lipschitz with respect to $\boldsymbol{x}$. Hence,

$$\begin{aligned}
|\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_k}) - \mu^{(m)}_{\boldsymbol{y},\boldsymbol{\epsilon}}(X_{i_k})| &= |\mu^{(m)}_t(X) + \hat{\Sigma}^{(m)}_{\mathcal{D}}(X, \boldsymbol{x})\epsilon^{(m)} - \mu^{(m)}_t(X) - \hat{\Sigma}^{(m)}_{\mathcal{D}}(X, \boldsymbol{y})\epsilon^{(m)}| \\
&= |\hat{\Sigma}^{(m)}_{\mathcal{D}}(X, \boldsymbol{x})\epsilon^{(m)} - \hat{\Sigma}^{(m)}_{\mathcal{D}}(X, \boldsymbol{y})\epsilon^{(m)}| \\
&\leq C^{(m)}_3 |\epsilon^{(m)}| \cdot ||\boldsymbol{x} - \boldsymbol{y}||,
\end{aligned}$$

where $C^{(m)}_3 \in \mathbb{R}$, for all $m = 1, ..., M$. Note that $\mu_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_k}) \leq 0 \leq \mu_{\boldsymbol{y},\boldsymbol{\epsilon}}(X_{i_k})$. So $|a_{m,j,X_j}(\boldsymbol{x}, \boldsymbol{\epsilon}) - a_{m,j,X_j}(\boldsymbol{y}, \boldsymbol{\epsilon})| = |0 - \mu_{\boldsymbol{y},\boldsymbol{\epsilon}}(X_{i_k})| \leq |\mu_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_k}) - \mu_{\boldsymbol{y},\boldsymbol{\epsilon}}(X_{i_k})| \leq C^{(m)}_3 |\boldsymbol{\epsilon}| \cdot ||\boldsymbol{x} - \boldsymbol{y}||$.

Case 3: Suppose both terms are not zero, i.e., $\min[\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_1}), ..., \mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_j})] \geq 0$ and $\min[\mu^{(m)}_{\boldsymbol{y},\boldsymbol{\epsilon}}(X_{i_1}), ..., \mu^{(m)}_{\boldsymbol{y},\boldsymbol{\epsilon}}(X_{i_j})] \geq 0$. Let $k = \arg\min_{k=1,...,j} \mu^{(m)}_{\boldsymbol{y},\boldsymbol{\epsilon}}(X_{i_k})$. Let $q = \arg\min_{k=1,...,j} \mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_k})$.

Suppose $k = q$. Then, $|a_{m,j,X_j}(\boldsymbol{x}, \boldsymbol{\epsilon}) - a_{m,j,X_j}(\boldsymbol{y}, \boldsymbol{\epsilon})| = |\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_k}) - \mu^{(m)}_{\boldsymbol{y},\boldsymbol{\epsilon}}(X_{i_k})| \leq C^{(m)}_3 |\epsilon^{(m)}| \cdot ||\boldsymbol{x} - \boldsymbol{y}||$.

Suppose $k \neq q$.

Suppose $\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_q}) \leq \mu^{(m)}_{\boldsymbol{y},\boldsymbol{\epsilon}}(X_{i_k})$. Since $\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_q}) \leq \mu^{(m)}_{\boldsymbol{y},\boldsymbol{\epsilon}}(X_{i_k}) \leq \mu^{(m)}_{\boldsymbol{y},\boldsymbol{\epsilon}}(X_{i_q})$, we have that $|\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_q}) - \mu^{(m)}_{\boldsymbol{y},\boldsymbol{\epsilon}}(X_{i_k})| \leq |\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_k}) - \mu^{(m)}_{\boldsymbol{y},\boldsymbol{\epsilon}}(X_{i_k})| \leq C^{(m)}_3 |\epsilon^{(m)}| \cdot ||\boldsymbol{x} - \boldsymbol{y}||$ because $\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(\cdot), \mu^{(m)}_{\boldsymbol{y},\boldsymbol{\epsilon}}(\cdot)$ are Lipschitz w.r.t. $\boldsymbol{x}$, $\boldsymbol{y}$ respectively, as noted above.

Suppose $\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_q}) > \mu^{(m)}_{\boldsymbol{y},\boldsymbol{\epsilon}}(X_{i_k})$. Similarly, since $\mu^{(m)}_{\boldsymbol{y},\boldsymbol{\epsilon}}(X_{i_k}) < \mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_q}) \leq \mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_k})$, we have that $|\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_q}) - \mu^{(m)}_{\boldsymbol{y},\boldsymbol{\epsilon}}(X_{i_k})| \leq |\mu^{(m)}_{\boldsymbol{x},\boldsymbol{\epsilon}}(X_{i_k}) - \mu^{(m)}_{\boldsymbol{y},\boldsymbol{\epsilon}}(X_{i_k})| \leq C^{(m)}_3 |\epsilon^{(m)}| \cdot ||\boldsymbol{x} - \boldsymbol{y}||$.

So, $|a_{m,j,X_j}(\boldsymbol{x}, \boldsymbol{\epsilon}) - a_{m,j,X_j}(\boldsymbol{y}, \boldsymbol{\epsilon})| \leq C^{(m)}_3 |\epsilon^{(m)}| \cdot ||\boldsymbol{x} - \boldsymbol{y}||$.

Hence, in all cases, $|a_{m,j,X_j}(\boldsymbol{x}, \boldsymbol{\epsilon}) - a_{m,j,X_j}(\boldsymbol{y}, \boldsymbol{\epsilon})| \leq C^{(m)}_3 |\epsilon^{(m)}| \cdot ||\boldsymbol{x} - \boldsymbol{y}||$.

Plugging into (12), we have

$$|\tilde{A}_{j,X_j}(\boldsymbol{x}, \boldsymbol{\epsilon}) - \tilde{A}_{j,X_j}(\boldsymbol{y}, \boldsymbol{\epsilon})| \leq l(\boldsymbol{\epsilon}) ||\boldsymbol{x} - \boldsymbol{y}||.$$

with

$$l(\boldsymbol{\epsilon}) = \big(|r^{(1)}| + j(C^{(1)}_1 + C^{(1)}_2 |\epsilon^{(1)}|)\big) \cdot C^{(2)}_3 |\epsilon^{(2)}| + (|r^{(2)}| + j\big(C^{(2)}_1 + C^{(2)}_2 |\epsilon^{(2)}|)\big) \cdot C^{(1)}_3 |\epsilon^{(1)}|).$$

This result can be generalized for any M by telescoping the expressions in (10). Hence $\tilde{A}_{j,X_j}(\boldsymbol{x}, \boldsymbol{\epsilon})$ is $l(\boldsymbol{\epsilon})$-Lipschitz continuous and thus $A(\boldsymbol{x}, \boldsymbol{\epsilon})$ is Lipschitz continuous. □

Theorem B.1. *Let the search space $\mathcal{X}$ be compact, the prior mean function μ_0 be constant, and the prior covariance function K_0 be continuously differentiable. Let $X^* \in \arg\max_{X \subseteq \mathcal{X}} \mathrm{HV}\big[\boldsymbol{\mu}(X \mid \mathcal{D}_{\boldsymbol{x}})\big]$. Then*

$$\nabla_{\boldsymbol{x}} \mathbb{E}_{\mathcal{D}} \left[\max_{X \subseteq \mathcal{X}} \mathrm{HV}\Big[\boldsymbol{\mu}(X \mid \mathcal{D}_{\boldsymbol{x}})\Big] \right] = \mathbb{E}_{\mathcal{D}} \left[\nabla_{\boldsymbol{x}} \mathrm{HV}\Big[\boldsymbol{\mu}(X^* \mid \mathcal{D}_{\boldsymbol{x}})\Big] \right].$$

Proof. The proof follows that of (Wu et al., 2020a, Theorem 1). We wish to show that

$$\begin{aligned}
\nabla_{\boldsymbol{x}} \mathbb{E}_{\mathcal{D}} \left[\max_{X \subseteq \mathcal{X}} \mathrm{HV}\Big[\boldsymbol{\mu}(X \mid \mathcal{D}_{\boldsymbol{x}})\Big] \right] &= \mathbb{E}_{\mathcal{D}} \left[\nabla_{\boldsymbol{x}} \max_{X \subseteq \mathcal{X}} \mathrm{HV}\Big[\boldsymbol{\mu}(X \mid \mathcal{D}_{\boldsymbol{x}})\Big] \right] && (13) \\
&= \mathbb{E}_{\mathcal{D}} \left[\nabla_{\boldsymbol{x}} \mathrm{HV}\Big[\boldsymbol{\mu}(X^* \mid \mathcal{D}_{\boldsymbol{x}})\Big] \right]. && (14)
\end{aligned}$$

To justify (14), we begin by expressing the posterior mean $\boldsymbol{\mu}(X \mid \mathcal{D}_{\boldsymbol{x}}) = \big[\mu^{(1)}(X \mid \mathcal{D}_{\boldsymbol{x}}), \ldots, \mu^{(M)}(X \mid \mathcal{D}_{\boldsymbol{x}})\big]$ for each outcome in terms of a standard normal random vector: $\mu^{(m)}(X) = \mu^{(m)}(X \mid \mathcal{D}) + \hat{\Sigma}^{(m)}_{\mathcal{D}}(X, \boldsymbol{x} \mid \mathcal{D})\epsilon^{(m)}$ for $m = 1, ..., M$.

Note that for a fixed $\boldsymbol{x}$, $\mu^{(m)}(X \mid \mathcal{D})$ and $\Sigma^{(m)}(X, \boldsymbol{x} \mid \mathcal{D})$ are continuously differentiable in X a.e.,[16] and for a fixed X, $\Sigma^{(m)}(X, \boldsymbol{x} \mid \mathcal{D})$ is continuously differentiable in $\boldsymbol{x}$ (Wu et al., 2020a, Lemma 1).[16] Hence, $\boldsymbol{\mu}(X \mid \mathcal{D}_{\boldsymbol{x}})$ is continuously differentiable w.r.t to $\boldsymbol{x}$ for fixed X and continuously differentiable w.r.t to X for fixed $\boldsymbol{x}$.[16]

From Equation (9), it is easily to see that $\mathrm{HV}(Y, r)$ is continuous over $Y \in \mathbb{R}^M$. Moreover, the partial derivatives of the input Y exist almost everywhere, since HV is an incarnation of hypervolume improvement with no incumbent Pareto frontier and hypervolume improvement is differentiable almost everywhere w.r.t Y (Daulton et al., 2020).[17] Hence, $\mathrm{HV}\big[\boldsymbol{\mu}(X \mid \mathcal{D}_{\boldsymbol{x}})\big]$ is differentiable w.r.t $\boldsymbol{x}$ almost everywhere for a fixed X and $\boldsymbol{\epsilon}$ and is differentiable w.r.t X almost everywhere for a fixed $\boldsymbol{x}$ and $\boldsymbol{\epsilon}$.

To employ the envelope theorem (Milgrom and Segal, 2002), we need to show that the following conditions of Milgrom and Segal (2002, Theorem 2) hold:

1. $\mathrm{HV}\big[\boldsymbol{\mu}(X \mid \mathcal{D}_{\boldsymbol{x}})\big]$ is absolutely continuous w.r.t. $\boldsymbol{x}$ for a fixed $\boldsymbol{\epsilon}$ and a fixed X.
2. There exists an integrable function $b : \mathcal{X} \to \mathbb{R}$ such that $||\nabla_{\boldsymbol{x}} \mathrm{HV}\big[\boldsymbol{\mu}(X \mid \mathcal{D}_{\boldsymbol{x}})\big]|| \leq b(\boldsymbol{x})$ for almost all $\boldsymbol{x} \in \mathcal{X}$ and for all X.

From Lemma B.1, $\mathrm{HV}\big[\boldsymbol{\mu}(X \mid \mathcal{D}_{\boldsymbol{x}})\big]$ is Lipschitz continuous in $\boldsymbol{x}$ for a fixed $\boldsymbol{\epsilon}, X$, so it is absolutely continuous. Furthermore, since it is Lipschitz continuous, its gradient is bounded almost everywhere. Hence, we have

$$\nabla_{\boldsymbol{x}} \max_{X \subseteq \mathcal{X}} \mathrm{HV}\big[\boldsymbol{\mu}(X \mid \mathcal{D}_{\boldsymbol{x}})\big] = \nabla_{\boldsymbol{x}} \mathrm{HV}\Big[\max_{X \subseteq \mathcal{X}} \boldsymbol{\mu}(X \mid \mathcal{D}_{\boldsymbol{x}})\Big] = \nabla_{\boldsymbol{x}} \mathrm{HV}\big[\boldsymbol{\mu}(X^* \mid \mathcal{D}_{\boldsymbol{x}})\big],$$

showing equality between (13) and (14). To show (13), we note that since $\mathcal{X}$ is compact, $\mathrm{HV}\big[\boldsymbol{\mu}(X \mid \mathcal{D}_{\boldsymbol{x}})\big]$ is bounded, which satisfies the conditions of Bartle (1995, Corollary 5.8). Given the result in Bartle (1995, Corollary 5.8) and noting again that the partial derivatives of $\mathrm{HV}\big[\boldsymbol{\mu}(X \mid \mathcal{D}_{\boldsymbol{x}})\big]$ are bounded almost everywhere, we can interchange expectation and gradient (Bartle, 1995, Corollary 5.9), which justifies (13) and completes the proof. □

Corollary B.1. *Let* $\mu_i^{(m)}(X) := \mu^{(m)}(X \mid \mathcal{D}) + \hat{\Sigma}^{(m)}(X, \boldsymbol{x} \mid \mathcal{D})\epsilon_i^{(m)}$ *for* $m = 1, \ldots, M$, $\boldsymbol{\mu}_i(X) := \big[\mu_i^{(1)}(X), \ldots, \mu_i^{(M)}(X)\big]$, *and* $\epsilon_i \sim \mathcal{N}(0, I_M)$ *iid. Let* $X_i^* \in \arg\max_{X \subseteq \mathcal{X}} \mathrm{HV}\big[\boldsymbol{\mu}(X \mid \mathcal{D}_{\boldsymbol{x}}^i)\big]$, *where* $\mathcal{D}_{\boldsymbol{x}}^i = \mathcal{D} \cup \{(\boldsymbol{x}, \boldsymbol{y}^i)\}$ *with* $\boldsymbol{y}^i \sim p(\boldsymbol{y} \mid \mathcal{D}, \boldsymbol{x})$. *Then an unbiased estimator of the gradient* $\nabla_{\boldsymbol{x}} \alpha_{\text{HV-KG}}(\boldsymbol{x})$ *is given by the average of the sample-level gradients*

$$\nabla_{\boldsymbol{x}} \alpha_{\text{HV-KG}}(\boldsymbol{x}) \approx \frac{1}{N} \sum_{i=1}^{N} \nabla_{\boldsymbol{x}} \mathrm{HV}\Big[\boldsymbol{\mu}(X_i^* \mid \mathcal{D}_{\boldsymbol{x}}^i)\Big].$$

The result follow directly from Theorem B.1 and approximating the expectation via Monte Carlo (using independence of the ϵ_i). Computing the gradient estimator in Corollary B.1 requires solving the inner maximization problem to obtain X_i^* and computing the sample-level gradient for each of the N samples.

Lemma B.2. *Suppose that* $\mathcal{X}$ *is compact and that* $\boldsymbol{f}(\boldsymbol{x}) \sim GP(\boldsymbol{0}, K_0(\boldsymbol{x}, \boldsymbol{x}))$ *is a zero-mean multi-output Gaussian Process prior with* M *outputs. Suppose that* $||\boldsymbol{f}(\boldsymbol{x})|| < \infty$ *almost surely for all* $\boldsymbol{x} \in \mathcal{X}$. *Let* $X \subseteq \mathcal{X}$ *such that* $|X| \leq N_p$, *and let* $r \in \mathbb{R}^M$. *Then, the moment generating function*

$$\mathbb{E}\left[\exp\Big(t \cdot \sup_{X \subseteq \mathcal{X}} \mathrm{HV}\big[\boldsymbol{f}(X)\big]\Big)\right]$$

of $\sup_{X \subseteq \mathcal{X}} \mathrm{HV}[\boldsymbol{f}(X)]$, *where* $t \in \mathbb{R}$, *is finite for all* t.

Proof. Let use denote the components of $\boldsymbol{f}(\boldsymbol{x})$ by $f^{(1)}(\boldsymbol{x}), ..., f^{(M)}(\boldsymbol{x})$. Since $||\boldsymbol{f}(\boldsymbol{x})|| < \infty$ for all $\boldsymbol{x} \in \mathcal{X}$ a.s., we have that $|f^{(i)}(\boldsymbol{x})| < \infty$ for all $\boldsymbol{x} \in \mathcal{X}$ and $m = 1, \ldots, M$ a.s.. Therefore, $\mathbb{E}\big[\sup_{\boldsymbol{x} \in \mathcal{X}} f^{(m)}(\boldsymbol{x})\big] < \infty$ for $m = 1, \ldots, M$ (Adler,

[16] If there are repeated points in X and noise variance is not positive at all points X, then the gradient does not exist.

[17] The partial derivative of HV with respect to $Y_{i,j}$, the element in the i^{th} row and j^{th} column of Y, is not defined when there exists another $k! = i$ such that $Y_{i,j} = Y_{k,j}$ or when $Y_{i,j} = r_j$, where r_j is the reference point value for objective j. The set of points defined by the union of these settings has zero measure under any GP posterior (Daulton et al., 2020). Furthermore, the gradient is only computed at the optimal $\mathcal{X}'^*$, and typically, $N_p \leq |\mathcal{X}^*|$, so columns of $\boldsymbol{\mu}_{t+1}(\mathcal{X}'^*)$ will contain unique values if $\boldsymbol{\mu}_{t+1}$ is representative of the underlying objectives.

1990, Theorem 2.1). Let $\boldsymbol{f}^*$ denote the component-wise supremum of $\boldsymbol{f}$: i.e. $\boldsymbol{f}^* = [\sup_{\boldsymbol{x}\in\mathcal{X}} f^{(1)}(\boldsymbol{x}), ..., \sup_{\boldsymbol{x}\in\mathcal{X}} f^{(M)}\boldsymbol{x})]$. By definition $\boldsymbol{f}^* \succeq \boldsymbol{f}(\boldsymbol{x})$ for all $\boldsymbol{x} \in X$. Hence, $\text{HV}[\{\boldsymbol{f}^*\}] \geq \sup_{X\subseteq\mathcal{X}} \text{HV}[\boldsymbol{f}(X)]$. Since HV is non-negative, it is sufficient to consider $t \geq 0$. From Equation (9), we have that $\text{HV}[\{\boldsymbol{f}^*\}] = \prod_{m=1}^{M} \max(\sup_{\boldsymbol{x}\in\mathcal{X}} f^{(m)}(\boldsymbol{x}) - r^{(m)}, 0)$. Without loss of generality, we may assume $r = 0$ (otherwise this corresponds to a simple shift of $\boldsymbol{f}$). Then,

$$\begin{aligned}
\text{HV}(\boldsymbol{f}^*) &= \prod_{m=1}^{M} \max\Big(\sup_{\boldsymbol{x}\in\mathcal{X}} f^{(m)}(\boldsymbol{x}), 0\Big) \\
&\leq \bigg(\max_{m=1...,M}\Big[\max\Big(\sup_{\boldsymbol{x}\in\mathcal{X}} f^{(m)}(\boldsymbol{x}), 0\Big)\Big]\bigg)^M \\
&\leq \max_{m=1...,M}\Big|\sup_{\boldsymbol{x}\in\mathcal{X}} f^{(m)}(\boldsymbol{x})\Big|^M \\
&\leq \max_{m=1...,M}\Big(\sup_{\boldsymbol{x}\in\mathcal{X}} |f^{(m)}(\boldsymbol{x})|\Big)^M.
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{E}\Big[\exp\big(t\cdot\text{HV}(\boldsymbol{f}^*)\big)\Big] &\leq \mathbb{E}\bigg[\exp\bigg(t\cdot\max_{m=1...,M}\Big(\sup_{\boldsymbol{x}\in\mathcal{X}} |f^{(m)}(\boldsymbol{x})|\Big)^M\bigg)\bigg] &(15)\\
&= \mathbb{E}\bigg[\max_{m=1...,M}\exp\bigg(t\cdot\Big(\sup_{\boldsymbol{x}\in\mathcal{X}} |f^{(m)}(\boldsymbol{x})|\Big)^M\bigg)\bigg] &(16)\\
&\leq \sum_{m=1}^{M}\mathbb{E}\bigg[\exp\bigg(t\cdot\big(\sup_{\boldsymbol{x}\in\mathcal{X}} |f^{(m)}(\boldsymbol{x})|\big)^M\bigg)\bigg], &(17)
\end{aligned}$$

where the final inequality comes from noting that all terms in the max are positive. Since all moments of $\sup_{\boldsymbol{x}\in\mathcal{X}} |f^{(m)}(\boldsymbol{x})|$ are finite (Balandat et al., 2020, Lemma 4), $\mathbb{E}\big[\sup_{\boldsymbol{x}\in\mathcal{X}} |f^{(m)}(\boldsymbol{x})|^M\big] \leq \infty$ for all $m = 1, ..., M$. From here, our proof follows that of Balandat et al. (2020, Lemma 4). Consider the m^{th} term in (17) and let $Z^{(m)} := \big(\sup_{\boldsymbol{x}\in\mathcal{X}} |f^{(m)}(\boldsymbol{x})|\big)^M$:

$$\begin{aligned}
\mathbb{E}[\exp(t\cdot Z^{(m)})] &= \int_0^\infty p\big(\exp(t\cdot Z^{(m)}) > u\big)du \\
&\leq 1 + \int_1^\infty p\big(\exp(t\cdot Z^{(m)}) > u\big)du \\
&= 1 + \int_1^\infty p\big(Z^{(m)} > \frac{\log u}{t}\big)du \\
&= 1 + \int_1^\infty p\big(Z^{(m)} - \mathbb{E}[Z^{(m)}] > \frac{\log u}{t} - \mathbb{E}[Z^{(m)}]\big)du
\end{aligned}$$

Using a change of variables where $v = \frac{\log u}{t} - \mathbb{E}[Z^{(m)}]$, we have that $dv = \frac{du}{ut}$ and $ut = te^{tv}e^{t\mathbb{E}[Z^{(m)}]}$. Hence via substitution,

$$\begin{aligned}
\mathbb{E}[\exp(t\cdot Z^{(m)})] &\leq 1 + \int_1^\infty p\big(Z^{(m)} - \mathbb{E}[Z^{(m)}] > \frac{\log u}{t} - \mathbb{E}[Z^{(m)}]\big)du \\
&= 1 + te^{\mathbb{E}[Z^{(m)}]}\int_{-\mathbb{E}[Z^{(m)}]}^\infty p\big(Z^{(m)} - \mathbb{E}[Z^{(m)}] > v\big)e^{tv}dv \\
&= 1 + te^{\mathbb{E}[Z^{(m)}]}\bigg[\int_{\min(-\mathbb{E}[Z^{(m)}],0)}^0 p\big(Z^{(m)} - \mathbb{E}[Z^{(m)}] > v\big)e^{tv}dv \\
&\qquad + \int_0^\infty p\big(Z^{(m)} - \mathbb{E}[Z^{(m)}] > v\big)e^{tv}dv\bigg] \\
&\leq 1 + te^{\mathbb{E}[Z^{(m)}]}\bigg[|\mathbb{E}[Z^{(m)}]| + \int_0^\infty p\big(Z^{(m)} > v + \mathbb{E}[Z^{(m)}]\big)e^{tv}dv\bigg]
\end{aligned}$$

Note that since $Z^{(m)} = \big(\sup_{\boldsymbol{x}\in\mathcal{X}} |f^{(m)}(\boldsymbol{x})|\big)^M$,

$$p\big(Z^{(m)} > v + \mathbb{E}[Z^{(m)}]\big) = p\Big(\sup_{\boldsymbol{x}\in\mathcal{X}} |f^{(m)}(\boldsymbol{x})| > (v + \mathbb{E}[Z^{(m)}])^{\frac{1}{M}}\Big).$$

Let $\sigma^2_{\mathcal{X}} = \sup_{\boldsymbol{x}\in\mathcal{X}} \big[(f^{(m)}(\boldsymbol{x}))^2\big]$. Then, the tail probability $p\big(\sup_{\boldsymbol{x}\in\mathcal{X}} f^{(m)}(\boldsymbol{x}) > \alpha\big)$ can be bounded as

$$p\Big(\sup_{\boldsymbol{x}\in\mathcal{X}} f^{(m)}(\boldsymbol{x}) > \alpha\Big) \leq e^{-\alpha^2/(2\sigma^2_{\mathcal{X}})}$$

by Borell's inequality (Adler, 1990, Section 2.1). Hence,

$$p\Big(\sup_{\boldsymbol{x}\in\mathcal{X}} |f^{(m)}(\boldsymbol{x})| > \alpha\Big) \leq 2e^{-\alpha^2/(2\sigma^2_{\mathcal{X}})}.$$

Letting $\alpha = v + \mathbb{E}[Z^{(m)}]$, we have that

$$p\big(Z^{(m)} > v + \mathbb{E}[Z^{(m)}]\big) \leq 2e^{-(v+\mathbb{E}[Z^{(m)}])^2/(2\sigma^2_{\mathcal{X}})}.$$

Hence we obtain

$$\begin{aligned}
\mathbb{E}[\exp(t \cdot Z^{(m)})] &\leq 1 + te^{\mathbb{E}[Z^{(m)}]}\left[|\mathbb{E}[Z^{(m)}]| + \int_0^\infty p\big(Z^{(m)} > v + \mathbb{E}[Z^{(m)}]\big)e^{tv}dv\right] \\
&\leq 1 + te^{\mathbb{E}[Z^{(m)}]}|\mathbb{E}[Z^{(m)}]| + te^{\mathbb{E}[Z^{(m)}]}\int_0^\infty 2e^{tv-(v+\mathbb{E}[Z^{(m)}])^2/(2\sigma^2_{\mathcal{X}})}dv \\
&< \infty.
\end{aligned}$$

So,

$$\mathbb{E}\left[\exp\Big(t \cdot \sup_{X\subseteq\mathcal{X}} \mathrm{HV}\big[\boldsymbol{f}(X)\big]\Big)\right] < \infty.$$

□

Theorem 7.1. *Suppose that $\mathcal{X}$ is compact and that $\boldsymbol{f} \sim GP(\mu_0(\cdot), K_0(\cdot,\cdot))$ is a sample from a multi-output Gaussian process prior with continuously differentiable mean $\mu_0(\cdot)$ and covariance $K_0(\cdot,\cdot)$ functions. Let $\{\epsilon_i\}_{i=1}^N$ be i.i.d. base samples from $\mathcal{N}(0, I_M)$, let $\boldsymbol{x}^*_N \in \arg\max_{\boldsymbol{x}\in\mathcal{X}} \hat{\alpha}^N_{\text{HV-KG}}(\boldsymbol{x})$, and let $\alpha^*_{\text{HV-KG}} = \max_{\boldsymbol{x}\in\mathcal{X}} \alpha_{\text{HV-KG}}(\boldsymbol{x})$, then*

(i) $\hat{\alpha}_{\text{HV-KG}}(\boldsymbol{x}^*_N) \to \alpha^*_{\text{HV-KG}}$ *a.s.*

(ii) $\inf_{\boldsymbol{x}^*\in\mathcal{X}^*_{\text{HV-KG}}} ||\boldsymbol{x}^*_N - \boldsymbol{x}^*|| \to 0$ *a.s.*

(iii) $\forall\, \delta > 0, \exists\, K < \infty, \alpha > 0$ *such that*

$$p\Big(\inf_{\boldsymbol{x}^*\in\mathcal{X}^*_{\text{HV-KG}}} ||\boldsymbol{x}^*_N - \boldsymbol{x}^*|| \geq \delta\Big) \leq Ke^{-\alpha N}.$$

Proof. Let us express the integrand in $\hat{\alpha}^N_{\text{HV-KG}}(\boldsymbol{x})$ as $G(\boldsymbol{x}, \boldsymbol{\epsilon}) = \max_{X\subseteq\mathcal{X}} \mathrm{HV}(\boldsymbol{\mu}_i(X))$, where $\boldsymbol{\mu}_i(\cdot)$ is defined in Corollary B.1. As in Balandat et al. (2020); Daulton et al. (2020), we leverage Homem-de-Mello (2008, Proposition 2.2) to obtain our (i) and (ii). Homem-de-Mello (2008, Proposition 2.2) requires that two conditions be met (Homem-de-Mello, 2008, Assumptions A1, A2):

(A1) $\forall \boldsymbol{x} \in \mathcal{X},\ \hat{\alpha}^N_{\text{HV-KG}}(\boldsymbol{x}) \to \alpha_{\text{HV-KG}}(\boldsymbol{x})$ a.s.

(A2) there exists an integrable function $L(\boldsymbol{\epsilon}) : \mathbb{R}^M \to \mathbb{R}$ such that for almost every $\boldsymbol{\epsilon}$ and $\forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$,

$$|G(\boldsymbol{x}, \boldsymbol{\epsilon}) - G(\boldsymbol{y}, \boldsymbol{\epsilon})| \leq L(\boldsymbol{\epsilon})||\boldsymbol{x} - \boldsymbol{y}||.$$

Note that for any $\boldsymbol{\epsilon}$, the restriction from $\boldsymbol{x} \to G(\boldsymbol{x}, \boldsymbol{\epsilon})$ to the k^{th} coordinate, where $\boldsymbol{x} = (x_1, \ldots, x_d)$ and $k \in \{1, ..., d\}$, is Lipschitz continuous by Theorem B.1. Therefore, the partial derivative $\frac{\partial G(\boldsymbol{x},\boldsymbol{\epsilon})}{\partial x_k}$ exists and is bounded almost everywhere. That is, there exists $c_k \in \mathbb{R}^M$ such that $||c_k|| < \infty$ and $|\frac{\partial G(\boldsymbol{x},\boldsymbol{\epsilon})}{\partial x_k}| \leq c_k^T|\boldsymbol{\epsilon}|$, where $|\cdot|$ denotes the component-wise absolute value.

Consider the difference $|G(\boldsymbol{x}, \boldsymbol{\epsilon}) - G(\boldsymbol{y}, \boldsymbol{\epsilon})|$. We can bound this difference by summing the component-wise differences and leveraging the bounded partial derivatives to obtain

$$|G(\boldsymbol{x}, \boldsymbol{\epsilon}) - G(\boldsymbol{y}, \boldsymbol{\epsilon})| \leq \sum_{k=1}^{d} c_k^T |\boldsymbol{\epsilon}| \cdot |x_k - y_k| \leq \max_{k \in \{1,\dots,d\}} c_k^T |\boldsymbol{\epsilon}| \cdot ||x - y||_1. \tag{18}$$

Let $L_1(\boldsymbol{\epsilon}) = \max_{k \in \{1,\dots,d\}} c_k^T |\boldsymbol{\epsilon}|$. We need only to verify that $L_1(\boldsymbol{\epsilon})$ is integrable. Since $\boldsymbol{\epsilon}$ is a vector of standard Normal random variables,

$$\mathbb{E}\big[|L_1(\boldsymbol{\epsilon})|\big] \leq \max_{k \in \{1,\dots,d\}} \sum_{m=1}^{M} c_k^{(m)} \mathbb{E}\big[|\epsilon^{(m)}|\big] = \sqrt{\frac{2}{\pi}} \max_{k \in \{1,\dots,d\}} ||c_k||_1.$$

So $L_1(\boldsymbol{\epsilon})$ is integrable, and assumption (A2) holds.

Note that $G(\boldsymbol{x}, \boldsymbol{\epsilon})$ is the maximum hypervolume where the objectives are GPs. From Lemma B.2, the moment generating function $\mathbb{E}[e^{tG(\boldsymbol{x},\boldsymbol{\epsilon})}]$ is finite for all t. Noting that $G(\boldsymbol{x}, \boldsymbol{\epsilon})$ is positive for all $\boldsymbol{x}, \boldsymbol{\epsilon}$, we have that $\mathbb{E}[e^{t|G(\boldsymbol{x},\boldsymbol{\epsilon})|}]$ is also finite for all t. Hence, all of their absolute moments (Meyer, 2012, Exercise 9.15) and $\mathbb{E}[|G(\boldsymbol{x}, \boldsymbol{\epsilon})|]$ are finite for all $\boldsymbol{x}$. Thus, by the strong law of large numbers $\hat{\alpha}^N_{\text{HV-KG}}(\boldsymbol{x}) \to \alpha_{\text{HV-KG}}(\boldsymbol{x})$ a.s. where $\{\boldsymbol{\epsilon}_i\}_{i=1}^N$ are i.i.d. Therefore assumption (A1) holds.

To obtain (iii), we additionally need to show that there exists an integrable function $L_2(\boldsymbol{\epsilon}) : \mathbb{R}^M \to \mathbb{R}$ such that $G(\boldsymbol{x}, \boldsymbol{\epsilon})$ is $L_2(\boldsymbol{\epsilon})$-Lipschitz and the moment generating function $\mathbb{E}[e^{tL_2(\boldsymbol{\epsilon})}]$ of $L_2(\boldsymbol{\epsilon})$ is finite in an open neighborhood of $t = 0$ (originally from Homem-de-Mello (2008) and written concisely in Balandat et al. (2020, Proposition 2)). Let us define

$$L_2(\boldsymbol{\epsilon}) := M||\boldsymbol{\epsilon}||_\infty \cdot ||c_k||_\infty \geq \max_{k \in \{1,\dots,d\}} c_k^T |\boldsymbol{\epsilon}|.$$

From (18) it follows that $G(\boldsymbol{x}, \boldsymbol{\epsilon})$ is $L_2(\boldsymbol{\epsilon})$-Lipschitz in $\boldsymbol{x}$. Furthermore, $||\boldsymbol{\epsilon}||_\infty \leq ||\boldsymbol{\epsilon}||_1$. So, $L_2(\boldsymbol{\epsilon}) \leq C_1||\boldsymbol{\epsilon}||_1$, where $C_1 := M \cdot ||\boldsymbol{c}_k||_\infty < \infty$. Moreover,

$$\mathbb{E}\left[e^{tL_2(\boldsymbol{\epsilon})}\right] \leq \mathbb{E}\left[e^{tC_1||\boldsymbol{\epsilon}||_1}\right] = \mathbb{E}\left[e^{tC_1 \sum_{m=1}^M |\epsilon^{(m)}|}\right] = \mathbb{E}\left[\prod_{m=1}^{M} e^{tC_1|\epsilon^{(m)}|}\right] = \prod_{m=1}^{M} \mathbb{E}\left[e^{tC_1|\epsilon^{(m)}|}\right],$$

where we arrive at the last equality since $\epsilon^{(1)}, ..., \epsilon^{(M)}$ are independent. Let $M(t) = \prod_{m=1}^M \mathbb{E}\big[e^{tC_1|\epsilon^{(m)}|}\big]$. Note that $M(t)$ is simply the moment generating function of a folded Normal variable with scale parameter C_1^2, which is finite for all t. Hence $\mathbb{E}[e^{tL_2(\boldsymbol{\epsilon})}] < \infty$ for all t, which completes the proof. □

C. Alternative Knowledge Gradient Acquisition Functions

HV-KG is strongly motivated by the Bayesian decision-theoretic best point selection described in Section 4. That is, given a model of the objectives a decision maker will typically wish to infer the Pareto set of optimal designs and select one design from the Pareto set based on their preferences and estimates of the objectives for each design. In many MOBO works that consider inference regret (Hernandez-Lobato et al., 2016; Suzuki et al., 2020; Tu et al., 2022), it is common practice to determine the Pareto set over the search space under the posterior mean. Hence, HV-KG is constructed to be the one-step Bayes optimal acquisition function maximizing the hypervolume of the Pareto set under the posterior mean $\text{HV}(\boldsymbol{\mu}(X))$.

An alternative formulation would be to consider hypervolume as a utility function and seek to maximize the expected utility $\mathbb{E}[\text{HV}(\boldsymbol{f}(X))]$. Although many BO acquisition functions including the the single objective knowledge gradient are formulated as expected utilities, expected hypervolume would be difficult to leverage in a Bayesian decision-theoretic framework because it quanitifies the expected utility of a set of points rather than an individual point. In the single objective setting with a utility function $g : \mathbb{R} \to \mathbb{R}$ the point[18] x^* that maximizes the expected utility is given by $x^* = \arg\max_{x \in \mathcal{X}} \mathbb{E}[g(f(x))]$. Hence, it is simple to determine the best point with maximum expected utility in this framework. In the multi-objective setting, the hypervolume indicator is a set function and quantifies the utility of a set of points. Although one could identify the optimal set of points[19] $X^* = \arg\max_{X \in \mathcal{X}} \mathbb{E}[\text{HV}(\boldsymbol{f}(X))]$, selecting a single point from X^* to implement according to one's preferences would be challenging. Although the set X^* would be optimal with

[18] Technically there could be set of maximizers, but here we consider only one for simplicity.

[19] Rather, one could identify an approximation of the optimal set of designs, as discussed in Section 4.

respect to the expected hypervolume utility, using the posterior mean to estimate the objectives for each point in X^* may yield confusing results. Namely, the points in X^* would not necessarily in the Pareto optimal under the posterior mean. Hence, the expected utility would be misaligned given the method for selecting the best point. In contrast, maximizing the hypervolume of the posterior mean would directly align with the best point selection method.

Nevertheless, we define and evaluate a KG acquisition function that arises when treating HV as an expected utility.

$$\alpha_{\text{E-HV-KG}}(\boldsymbol{x}) = \mathbb{E}_{\mathcal{D}}\left[\max_{X \subseteq \mathcal{X}} \mathbb{E}\Big[\text{HV}(\boldsymbol{f}(X)) \mid \mathcal{D}_{\boldsymbol{x}}\Big] - \phi_*\right], \tag{19}$$

where $\phi^* := \max_{X \subseteq \mathcal{X}} \mathbb{E}\Big[\text{HV}(\boldsymbol{f}(X)) \mid \mathcal{D}\Big]$. This acquisition function has the desirable property of non-negativity.

Theorem C.1. *$\alpha_{\text{E-HV-KG}}(\boldsymbol{x})$ is non-negative for all $\boldsymbol{x}$ in $\mathcal{X}^q$*

Proof. We have that

$$\begin{aligned}\alpha_{\text{E-HV-KG}}(\boldsymbol{x}) &= \mathbb{E}_{\mathcal{D}}\left[\max_{X \subseteq \mathcal{X}} \mathbb{E}\Big[\text{HV}(\boldsymbol{f}(X)) \mid \mathcal{D}_{\boldsymbol{x}}\Big] - \phi_*\right] \\ &= \mathbb{E}_{\mathcal{D}}\left[\max_{X \subseteq \mathcal{X}} \mathbb{E}\Big[\text{HV}(\boldsymbol{f}(X)) \mid \mathcal{D}_{\boldsymbol{x}}\Big]\right] - \phi_*.\end{aligned}$$

The proof is straightforward and follows from the fact that the $\max$ function is convex. From Jensen's inequality, we have that

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[\max_{X \subseteq \mathcal{X}} \mathbb{E}\Big[\text{HV}(\boldsymbol{f}(X)) \mid \mathcal{D}_{\boldsymbol{x}}\Big]\right] &\geq \max_{X \subseteq \mathcal{X}} \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}\Big[\text{HV}(\boldsymbol{f}(X)) \mid \mathcal{D}_{\boldsymbol{x}}\Big]\right] \\ &= \max_{X \subseteq \mathcal{X}} \mathbb{E}\Big[\text{HV}(\boldsymbol{f}(X)) \mid \mathcal{D})\Big]] \\ &= \phi_*\end{aligned}$$

Hence, $\alpha_{\text{E-HV-KG}}(\boldsymbol{x}) \geq 0$. □

We leave the analysis of the non-negativity of HV-KG with a multi-output Gaussian process prior to future work. We note that hypervolume is not convex, and that for non-Gaussian priors, simple examples show that it can be negative.

C.1. Empirical Evaluation

Computing the expected utility requires Monte Carlo integration, and we evaluate the performance below with 16 samples. The decoupled and multi-fidelity variants are straightforward extensions of $\alpha_{\text{E-HV-KG}}$ using the same conditioning on partial information as with HV-KG.

In Figure 11, we evaluate $\alpha_{\text{E-HV-KG}}$ on single fidelity benchmarks with coupled evaluations and find that HV-KG typically performs at least as well as $\alpha_{\text{E-HV-KG}}$, but is much faster to optimize. $\alpha_{\text{E-HV-KG}}$ is much more expensive to compute due to the nested Monte Carlo integration and is slow even on a GPU as shown in Table 7.

In Figures 8 and 9, we evaluate a decoupled variant of $\alpha_{\text{E-HV-KG}}$ and similar results with respect to optimization performance, wall times, as shown in Tables 5 and 6. We where unable to run $\alpha_{\text{E-HV-KG}}$ on the NAS problem with non-competitive decoupling due to memory issues on a CPU and excessive runtime on a CPU. In Figure 10, we evaluate a MF variant of $\alpha_{\text{E-HV-KG}}$ and find that it works quite well, but is quite slow and we were unable to run it on the plasma laser acceleration problem and the ranking problem due to memory issues on a GPU and excessive wall time on a CPU. Wall times are reported in Table 4.

D. Additional Experiments

D.1. MOBO Problems with Complete Information

D.1.1. Sequential MOBO with Complete Information

We evaluate optimization performance in the standard sequential (i.e. $q = 1$), complete information multi-objective setting (Figure 12). We find that HV-KG is a top performer on most problems. HV-KG is outperformed by qNEHVI for noiseless

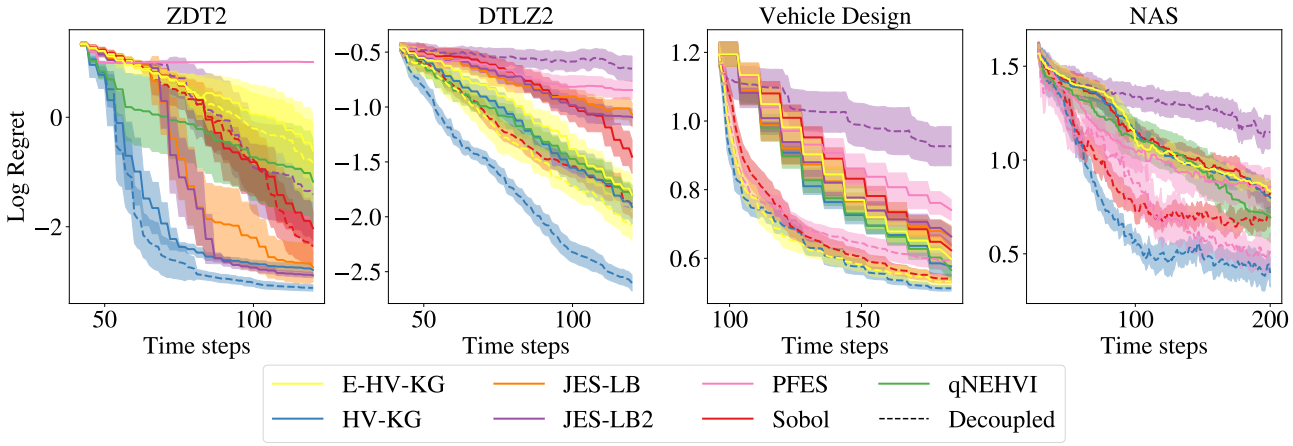

Figure 8: NCD benchmarks with $\alpha_{\text{E-HV-KG}}$.

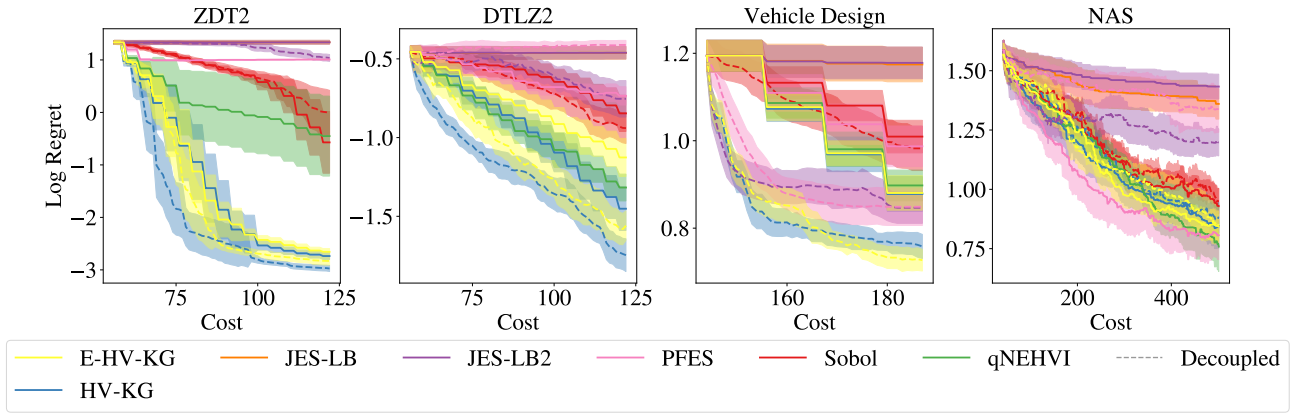

Figure 9: CD benchmarks with $\alpha_{\text{E-HV-KG}}$.

Penicilin, and performance is otherwise slightly better than, or not statistically significant from qNEHVI.

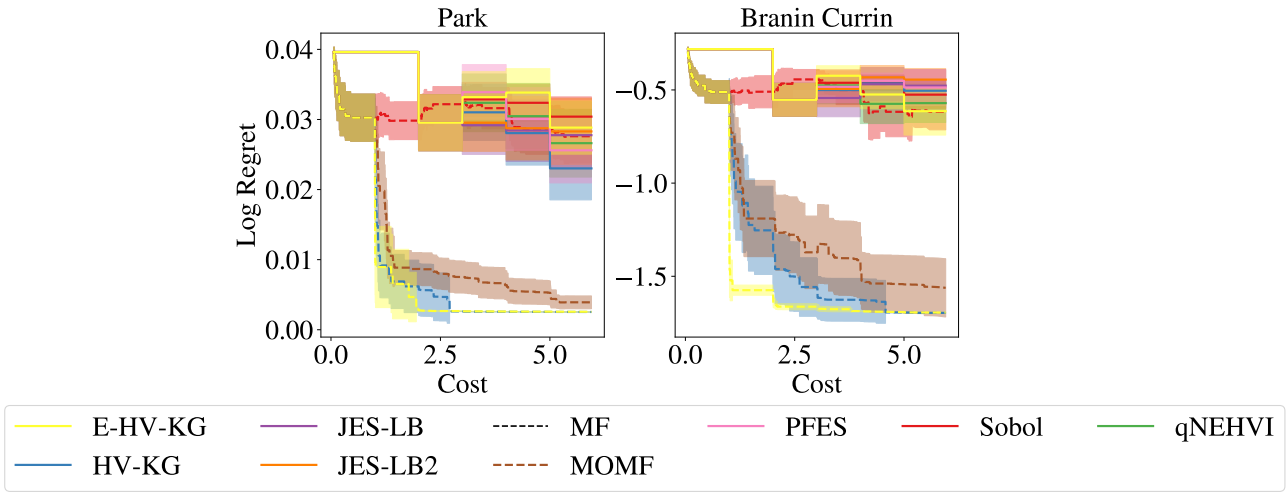

Figure 10: MF benchmarks with $\alpha_{\text{E-HV-KG}}$.

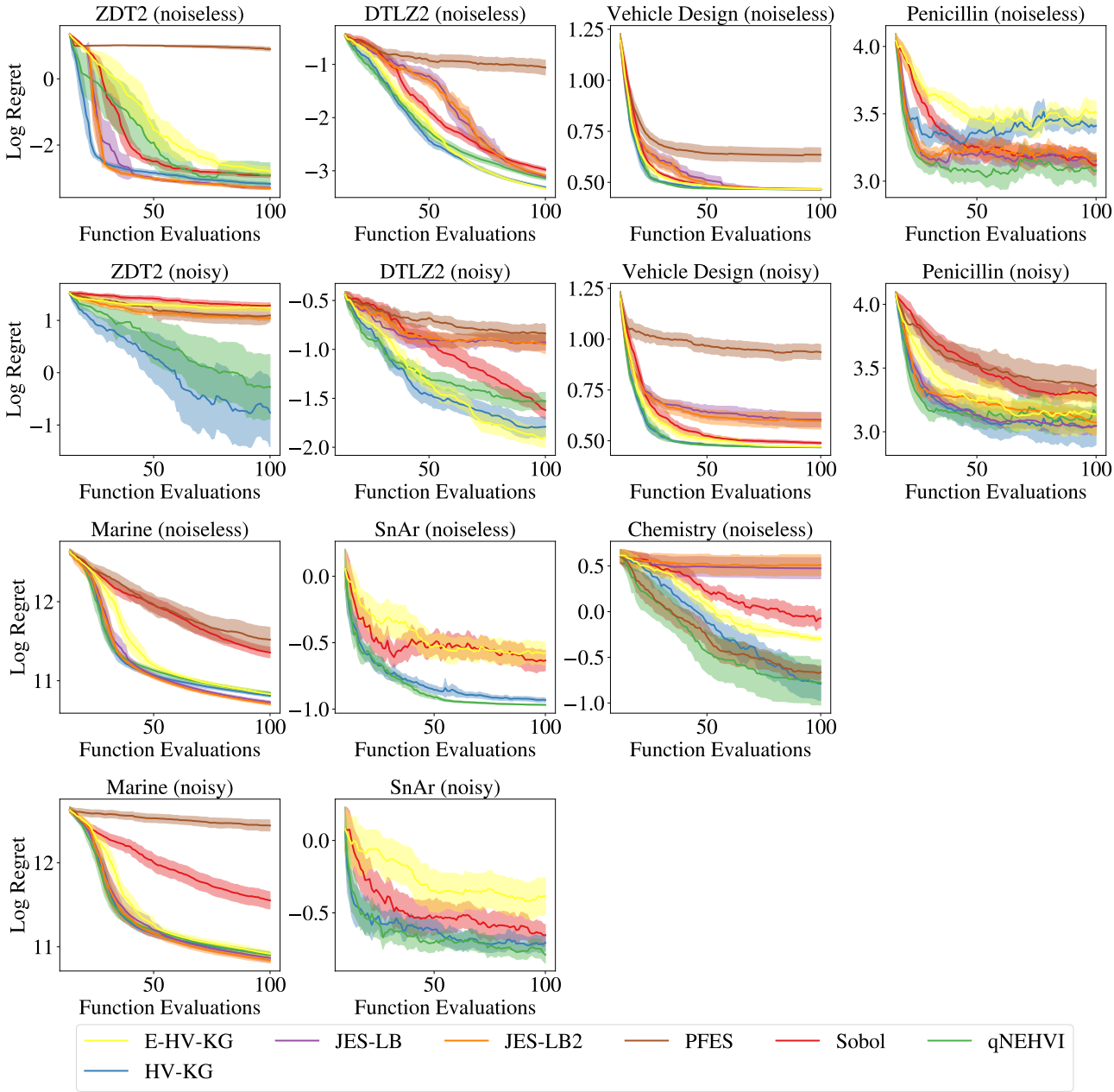

Figure 11: Single fidelity, coupled evaluation benchmarks with $\alpha_{\text{E-HV-KG}}$.

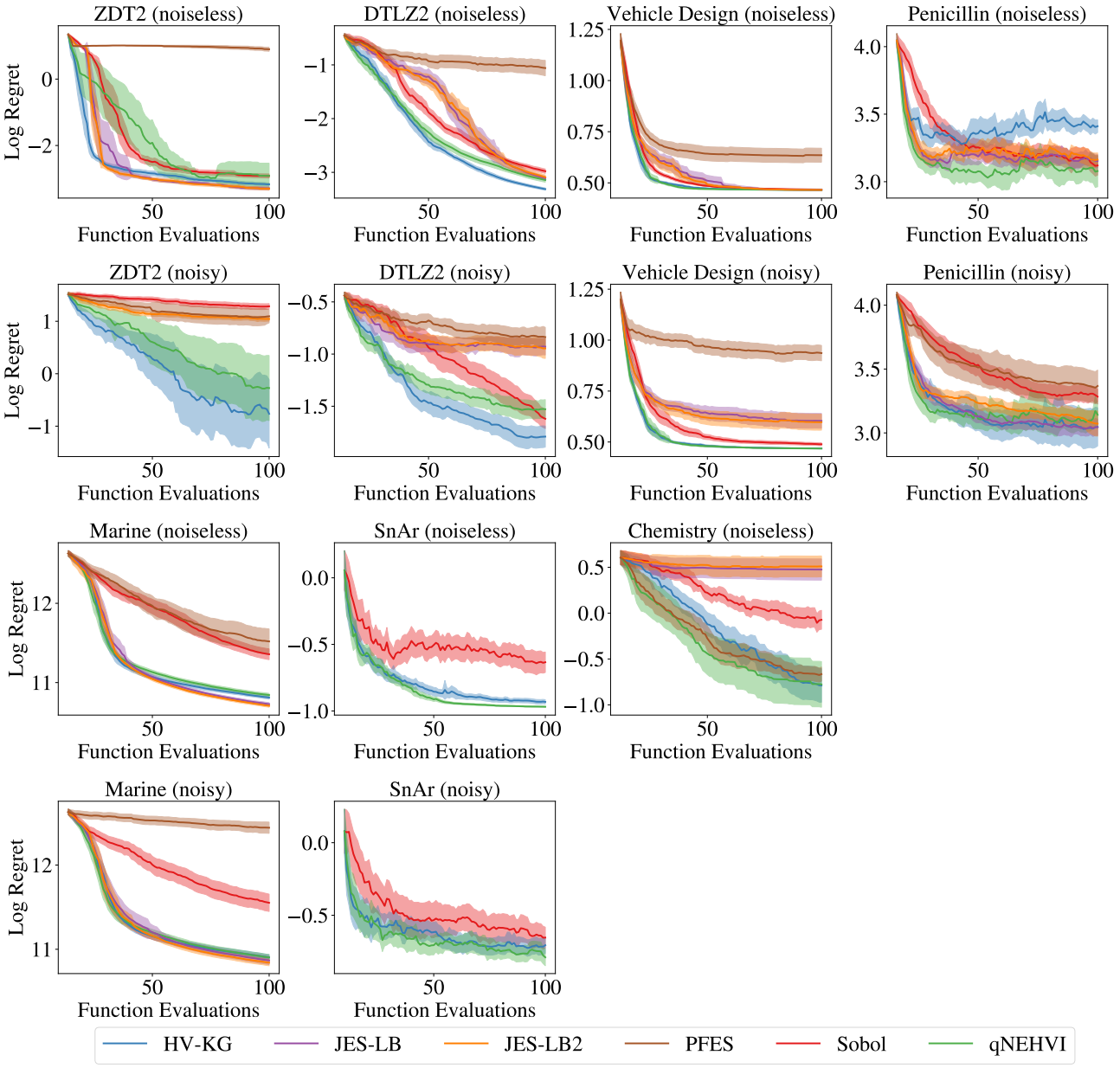

Figure 12: Sequential ($q = 1$) optimization performance on single fidelity problems.

D.1.2. Parallel MOBO with Complete Information

We evaluate optimization performance using a batch size of $q = 4$ (Figure 13). We find that HV-KG is a top performer on most problems. Like the sequential case, HV-KG is outperformed by qNEHVI for noiseless Penicilin, and performance is otherwise slightly better than, or not statistically significant from qNEHVI.

D.2. Sensitivity with Respect to Pareto Set Size and MC Samples

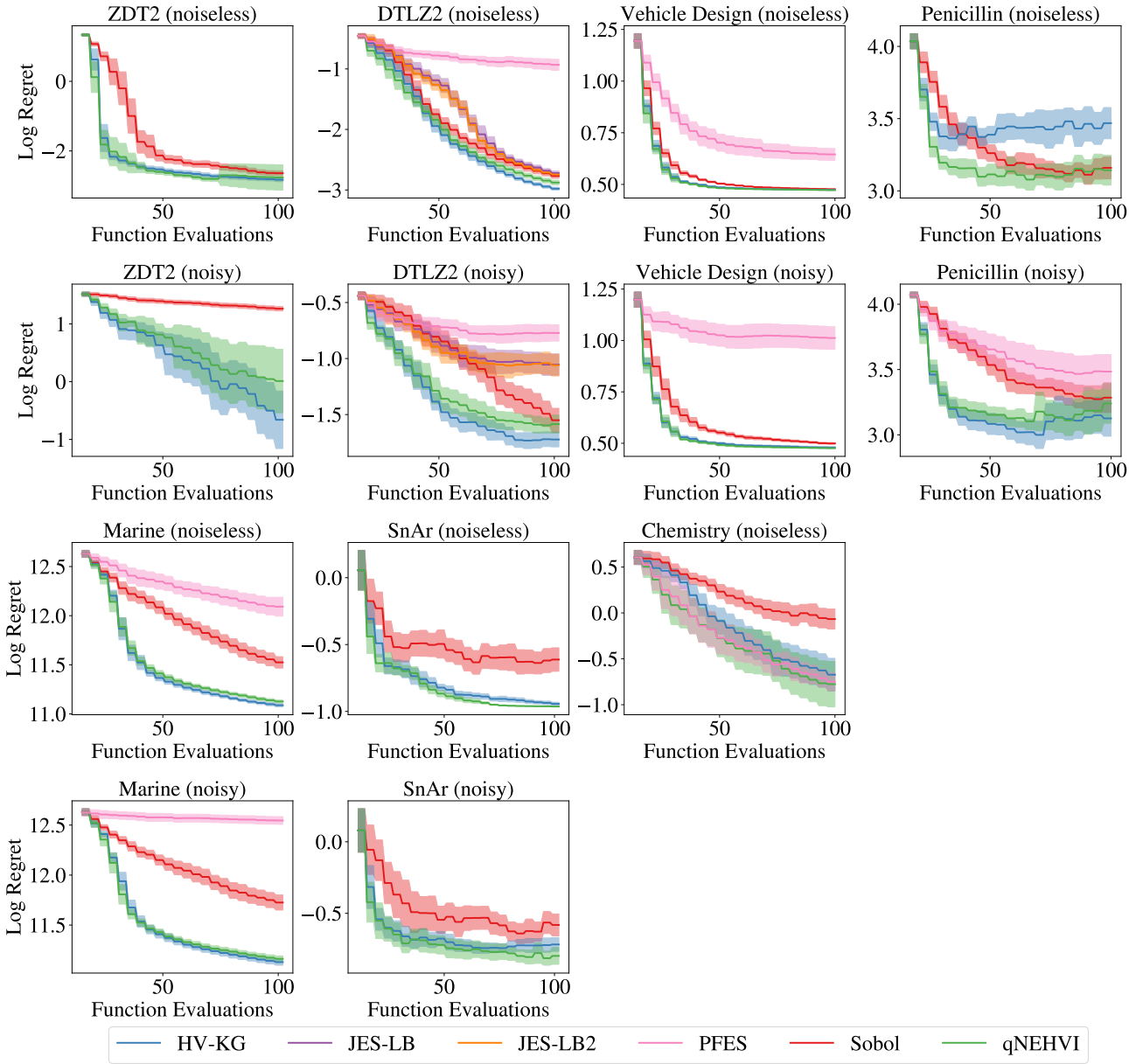

Figure 13: Parallel ($q = 4$) optimization performance on single fidelity problems. Many PFES and JES-LB(2) runs to failed with numerical errors, and so they are not reported in for some problems.

We evaluate the sensitivity of HV-KG to the Pareto set size N_p and the number of MC samples N and find that HV-KG is quite robust to both as show in Figures 14 and 15.

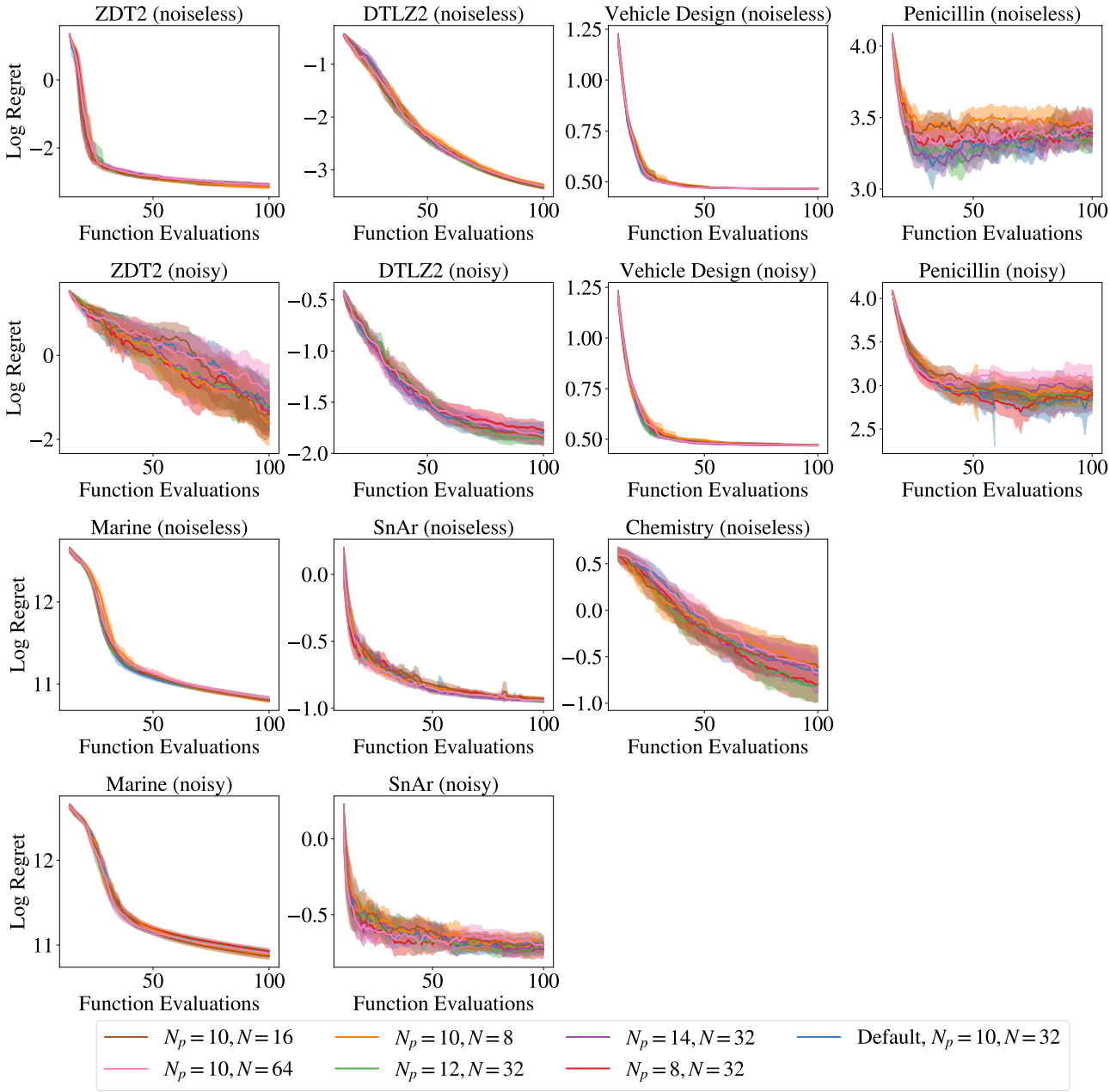

Figure 14: Sensitivity analysis on single fidelity problems. We do not observe any meaningful differences in performance across multiple problems and over a range of values for either the number of points in the finite Pareto Frontier approximation (N_p) or the number of fantasy samples (N).

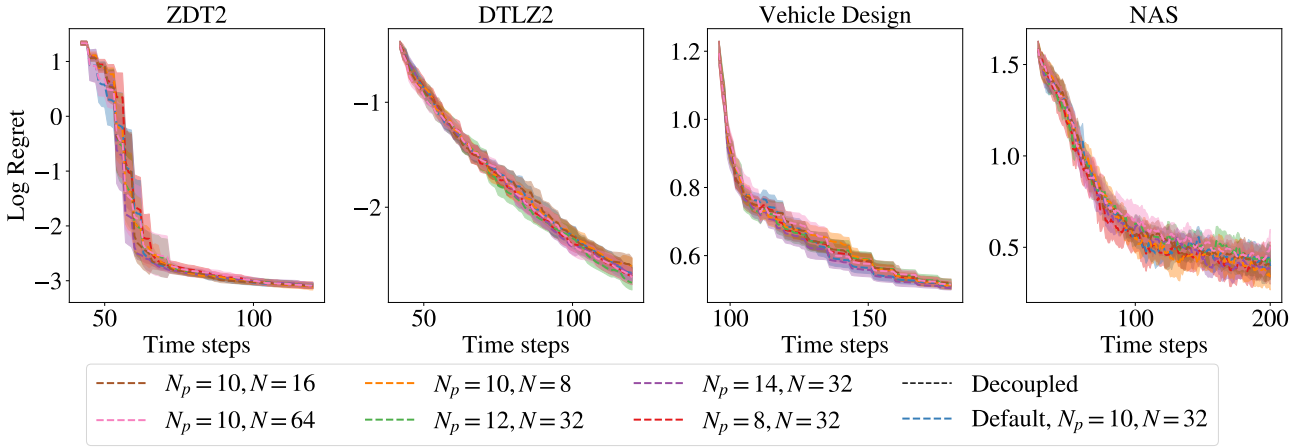

Figure 15: Sensitivity analysis on NCD problems. We do not observe any meaningful differences in performance across multiple problems and over a range of values for either the number of points in the finite Pareto Frontier approximation (N_p) or the number of fantasy samples (N).

D.3. Sensitivity to Costs in Competitive Decoupling

In this study, we examine the extent to which CD results are sensitive to the costs used for each objective, which can be particularly relevant when some objectives are more challenging to model than others. To do this, we swap the cost functions such that for ZDT2 and DTLZ2 the two objectives costs 3 and 1, respectively; for Vehicle Design, the objectives have costs 8, 3, and 1 respectively, and for NAS, the objectives have costs 2 and 1 respectively. We observe that decoupled entropy methods works significantly better on ZDT2 with the cost functions swapped. We note that the first objective is far simpler than the second objective, and in this case, the first objective is 3 times more expensive. Comparing Figure 6 in the main text and Figure 16 here, we find that HV-KG is robust with both cost configurations. In addition, we evaluate which objectives

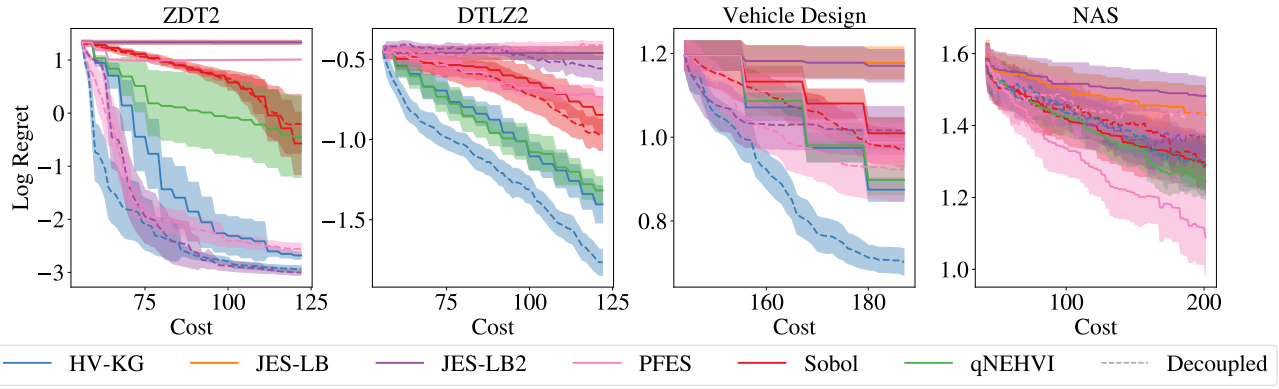

Figure 16: Competitive decoupling with swapped costs across objectives. Results are qualitatively similar to that of the main text, but the performance of decoupled JES improves for ZDT2, and deteriorates for Vehicle Design.

different decoupled algorithms choose to evaluate in the competitive decoupling setting. We observe that the behaviors of JES-LB2 and PFES are far more sensitive to the cost function and that those methods assign significantly more samples to lower cost objectives. When the costs are swapped, JES-LB2 and PFES again allocate significantly more evaluations to lower cost objectives, where as the change in HV-KG's behavior is less severe. We suspect the performance of JES-LB2 and PFES is quite sensitive to the choice of cost function.

	ZDT2 (1, 3)	DTLZ2 (1,3)	VEHICLE DESIGN (1, 3, 8)	NAS (1, 2)
HV-KG	[34, 29]	[45, 25]	[17, 16, 15]	[182, 159]
JES-LB2	[68, 18]	[78, 14]	[39, 15, 12]	[371, 65]
PFES	-	[66, 18]	[48, 12, 12]	[17, 242]

Table 2: Number of evaluations of each objective in the competitive decoupling setting.

	ZDT2 (3,1)	DTLZ2 (3,1)	VEHICLE DESIGN (8, 3, 1)	NAS (2,1)
HV-KG	[30, 31]	[27, 39]	[14, 18, 21]	[47, 107]
JES-LB2	[21, 57]	[14, 78]	[12, 12, 48]	[26, 148]
PFES	[20, 62]	[30, 32]	[13, 15, 31]	[14, 172]

Table 3: Number of evaluations of each objective in the competitive decoupling setting with swapped costs across objectives.

D.4. Wall Times

We find that candidate generation time with HV-KG is competitive with other methods in the decoupled setting as shown in Tables 5 and 6. Notably, HV-KG is significantly faster than the information theoretic alternatives on problems with decoupled evaluations. In the MF setting, MF-HV-KG and HV-KG are slower than alternatives as shown in Table 4, but MF-HV-KG is also the best performing method with respect to regret.

D.5. Details on Wall time Comparison of Nested Optimization via Unbiased Estimation

We use the default stochastic optimization routine in BoTorch (Balandat et al., 2020), which uses Adam (Kingma and Ba, 2014) with a constant learning rate of $\frac{1}{40}$ and an exponential moving average stopping strategy. We use L-BFGS-B to solve the inner optimization problem. To select starting points for gradient-based optimization, we sample 8 points from a scrambled Sobol sequence, evaluate HV-KG via solving the inner optimization problem, and use the standard Boltzmann sampling (Duchon et al., 2004) initialization procedure in BoTorch to select a single starting point. We limit the number of quasi-random points to 8 because HV-KG via solving the inner optimization problem is computationally intensive. For the SAA, we use initialization procedure described in Appendix A.2 and we use 1024 quasi-random points to select one starting point (i.e. the current design to select) because evaluating HV-KG in a one-shot fashion (i.e. not solving the inner optimization problem to completion for each $\boldsymbol{x}$) is fast. From the starting point, we use L-BFGS-B to HV-KG in a one-shot fashion. We report results on optimizing HV-KG under these two approaches in Figure 4 using a GP fit to 14 data points collected from the DTLZ2 $(d = 6, M = 2)$ problem (Deb et al., 2002). It worth noting that we limited the number of quasi-random points to 8 to run this comparison in a reasonable amount of time, but by the time gradient-based optimization starts for the nested stochastic approach, the SAA approach has achieved a higher HV-KG than the stochastic approach will ever reach (on average).

	PARK	BRANIN-CURRIN	RANKING POLICY OPTIMIZATION	PLASMA LASER ACCELERATION
E-HV-KG	336.6 (±45.7)	140.6 (±14.1)	-	-
E-MF-HV-KG	97.3 (±6.0)	74.5 (±2.5)	-	-
HV-KG	15.3 (±1.7)	14.1 (±4.7)	158.6 (±24.4)	55.3 (±9.2)
JES-LB	148.6 (±11.3)	54.8 (±3.1)	68.9 (±4.9)	154.4 (±6.1)
JES-LB2	133.6 (±9.6)	54.1 (±4.4)	70.0 (±3.5)	187.5 (±9.9)
MF-HV-KG	17.9 (±1.4)	16.0 (±0.8)	49.8 (±5.4)	42.7 (±3.1)
MF-SOBOL	0.3 (±0.0)	0.3 (±0.0)	0.3 (±0.0)	0.3 (±0.0)
MOMF	6.5 (±0.4)	5.2 (±0.2)	5.3 (±1.1)	8.6 (±0.6)
PFES	9.3 (±0.2)	11.8 (±1.8)	9.6 (±0.3)	28.0 (±4.0)
SOBOL	0.3 (±0.0)	0.3 (±0.0)	0.3 (±0.0)	0.3 (±0.0)
QNEHVI	2.2 (±0.2)	1.9 (±0.6)	5.0 (±0.3)	3.7 (±0.3)

Table 4: Acquisition function optimization wall time in seconds on a Tesla V100 SXM2 GPU (16GB RAM) for the multi-fidelity problems. The mean and two standard errors are reported.

	ZDT2	DTLZ2	VEHICLE DESIGN	NAS
HV-KG	16.2 (±1.3)	17.5 (±1.1)	30.6 (±3.8)	12.0 (±0.6)
HV-KG, DECOUPLED	28.7 (±2.3)	30.8 (±1.7)	55.4 (±6.5)	24.4 (±1.4)
E-HV-KG	203.3 (±9.6)	259.8 (±21.7)	265.4 (±25.1)	15.7 (±0.6)
E-HV-KG, DECOUPLED	290.0 (±12.7)	479.9 (±30.8)	584.9 (±35.9)	34.8 (±2.0)
JES-LB	100.6 (±8.1)	162.5 (±14.0)	119.9 (±8.9)	44.2 (±0.7)
JES-LB2	110.5 (±9.2)	163.7 (±11.9)	133.1 (±10.7)	43.6 (±0.6)
JES-LB2, DECOUPLED	172.1 (±13.5)	181.7 (±23.3)	305.4 (±26.4)	44.9 (±0.6)
PFES	15.5 (±1.3)	21.0 (±1.9)	40.5 (±9.0)	16.0 (±0.5)
PFES, DECOUPLED	-	20.7 (±1.8)	22.1 (±0.8)	17.3 (±0.5)
SOBOL	0.3 (±0.0)	0.3 (±0.0)	0.3 (±0.0)	4.2 (±0.3)
SOBOL, DECOUPLED	0.3 (±0.0)	0.3 (±0.0)	0.3 (±0.0)	4.4 (±0.3)
QNEHVI	3.8 (±0.2)	3.9 (±0.2)	4.1 (±0.2)	8.8 (±0.6)

Table 5: Acquisition function optimization wall time for problems with *competitive decoupling* in seconds on a Tesla V100 SXM2 GPU (16GB RAM). The mean and two standard errors are reported.

	ZDT2	DTLZ2	VEHICLE DESIGN	NAS
HV-KG	9.8 (±0.6)	12.9 (±1.7)	22.0 (±2.0)	13.7 (±0.6)
HV-KG, DECOUPLED	10.7 (±0.5)	12.9 (±0.7)	16.5 (±0.8)	132.4 (±5.4)
E-HV-KG	95.0 (±4.5)	185.4 (±10.1)	149.1 (±10.2)	17.5 (±0.7)
E-HV-KG, DECOUPLED	116.8 (±5.5)	262.1 (±11.3)	158.1 (±9.1)	-
JES-LB	62.0 (±4.1)	70.3 (±3.6)	68.6 (±3.4)	-
JES-LB2	80.1 (±5.9)	91.3 (±5.8)	91.2 (±6.7)	-
JES-LB2, DECOUPLED	104.6 (±5.6)	131.8 (±16.1)	177.2 (±11.3)	152.3 (±4.9)
PFES	13.9 (±0.8)	25.4 (±2.2)	38.9 (±3.8)	768.9 (±12.3)
PFES, DECOUPLED	-	-	349.6 (±8.4)	1421.1 (±36.1)
SOBOL	0.3 (±0.0)	0.3 (±0.0)	0.3 (±0.0)	4.1 (±0.3)
SOBOL, DECOUPLED	0.3 (±0.0)	0.3 (±0.0)	0.3 (±0.0)	4.1 (±0.2)
QNEHVI	4.2 (±0.4)	4.4 (±0.3)	4.9 (±0.5)	9.5 (±0.6)

Table 6: Acquisition function optimization wall time for problems with *non-competitive decoupling* in seconds on a Tesla V100 SXM2 GPU (16GB RAM). The mean and two standard errors are reported.

	DTLZ2 (NOISELESS)	DTLZ2 (NOISY)	ZDT2 (NOISELESS)	ZDT2 (NOISY)
E-HV-KG	135.3 (±6.8)	122.2 (±9.8)	69.7 (±3.0)	62.9 (±3.8)
HV-KG	11.3 (±0.2)	11.3 (±0.2)	11.9 (±1.0)	10.1 (±0.6)
JES-LB	98.2 (±9.0)	49.3 (±1.7)	-	44.2 (±1.7)
JES-LB2	133.2 (±3.9)	49.3 (±2.2)	130.8 (±5.9)	42.1 (±1.3)
PFES	33.2 (±1.9)	17.2 (±0.9)	16.0 (±0.7)	14.1 (±0.6)
SOBOL	2.8 (±0.0)	2.5 (±0.1)	3.3 (±0.0)	2.3 (±0.0)
QNEHVI	6.2 (±0.1)	6.9 (±0.3)	5.8 (±0.2)	5.5 (±0.2)
	VEHICLE DESIGN(NOISELESS)	VEHICLE DESIGN (NOISY)	PENICILLIN (NOISELESS)	PENICILLIN (NOISY)
E-HV-KG	91.5 (±3.1)	86.4 (±1.8)	218.7 (±14.9)	120.8 (±16.5)
HV-KG	34.3 (±1.4)	29.6 (±0.7)	82.0 (±9.8)	45.4 (±10.1)
JES-LB	167.4 (±9.2)	115.6 (±4.9)	172.2 (±27.8)	83.6 (±5.6)
JES-LB2	210.6 (±10.1)	118.5 (±4.3)	187.7 (±18.4)	82.9 (±4.9)
PFES	37.7 (±2.3)	36.2 (±3.6)	-	37.9 (±2.9)
SOBOL	15.8 (±0.5)	12.4 (±0.5)	6.3 (±0.1)	6.6 (±0.3)
QNEHVI	29.7 (±1.3)	27.5 (±1.1)	15.1 (±1.1)	14.7 (±0.8)
	SNAR (NOISELESS)	SNAR (NOISY)	MARINE (NOISELESS)	MARINE (NOISY)
E-HV-KG	67.9 (±3.5)	127.1 (±10.0)	219.4 (±15.4)	194.8 (±10.3)
HV-KG	24.7 (±0.9)	21.4 (±1.7)	74.9 (±2.6)	66.3 (±1.9)
JES-LB	-	-	245.2 (±14.3)	196.0 (±9.8)
JES-LB2	-	-	274.2 (±10.4)	189.7 (±8.2)
PFES	-	-	83.8 (±7.8)	59.0 (±3.2)
SOBOL	13.4 (±1.6)	13.5 (±3.1)	8.0 (±0.3)	7.8 (±0.5)
QNEHVI	19.5 (±0.9)	20.2 (±2.6)	69.9 (±2.5)	76.4 (±3.1)
	CHEMISTRY			
E-HV-KG	47.6 (±2.2)	-	-	-
HV-KG	8.8 (±0.7)	-	-	-
JES-LB	49.6 (±1.9)	-	-	-
JES-LB2	48.6 (±1.2)	-	-	-
PFES	11.9 (±0.4)	-	-	-
SOBOL	3.4 (±0.2)	-	-	-
QNEHVI	5.9 (±0.5)	-	-	-

Table 7: Sequential ($q = 1$) acquisition function optimization wall time in seconds on a Tesla V100 SXM2 GPU (16GB RAM). The mean and two standard errors are reported. PFES and JES-LB(2) failed are missing some values due to numerical errors , which caused the runs to fail.

	DTLZ2 (NOISELESS)	DTLZ2 (NOISY)	ZDT2 (NOISELESS)	ZDT2 (NOISY)
HV-KG	43.1 (±2.4)	49.6 (±1.7)	44.5 (±2.0)	35.4 (±1.6)
JES-LB	249.0 (±9.3)	72.1 (±4.0)	-	-
JES-LB2	333.3 (±15.1)	79.2 (±3.9)	-	-
PFES	1420.2 (±29.8)	1224.8 (±23.8)	-	-
SOBOL	0.3 (±0.0)	0.3 (±0.0)	0.3 (±0.0)	0.3 (±0.0)
QNEHVI	15.6 (±0.4)	18.9 (±0.6)	7.9 (±0.2)	11.1 (±0.5)
	VEHICLE DESIGN(NOISELESS)	VEHICLE DESIGN (NOISY)	PENICILLIN (NOISELESS)	PENICILLIN (NOISY)
HV-KG	69.5 (±2.5)	64.6 (±1.7)	333.7 (±42.5)	210.8 (±40.4)
PFES	1264.0 (±77.3)	1343.7 (±76.5)	-	1376.4 (±50.7)
SOBOL	0.3 (±0.0)	0.2 (±0.0)	0.3 (±0.0)	0.3 (±0.0)
QNEHVI	35.9 (±0.9)	37.1 (±0.9)	35.6 (±1.9)	45.0 (±2.1)
	SNAR (NOISELESS)	SNAR (NOISY)	MARINE (NOISELESS)	MARINE (NOISY)
HV-KG	45.9 (±1.6)	30.3 (±0.8)	252.3 (±8.4)	225.8 (±7.2)
PFES	-	-	1612.1 (±97.1)	1468.8 (±94.6)
SOBOL	0.2 (±0.0)	0.2 (±0.0)	0.2 (±0.0)	0.2 (±0.0)
QNEHVI	11.4 (±0.7)	9.2 (±0.6)	245.4 (±10.4)	264.6 (±10.0)
	CHEMISTRY			
HV-KG	23.6 (±0.6)	-	-	-
PFES	1201.3 (±29.3)	-	-	-
SOBOL	0.3 (±0.0)	-	-	-
QNEHVI	5.7 (±0.3)	-	-	-

Table 8: Batch ($q = 4$) acquisition function optimization wall time in seconds on a Tesla V100 SXM2 GPU (16GB RAM). The mean and two standard errors are reported. Most of the PFES and JES-LB(2) runs to failed with numerical errors.

D.6. Fidelity Selection Behavior

In this section, we examine the how different algorithms select fidelities. We examine the fidelity levels which MOMF and MF-HV-KG choose to evaluate at each iteration. We observe that MOMF tends to evaluate many more higher fidelities in early iterations and therefore exhausts its cost budget very quickly. In contrast, MF-HV-KG evaluates many more low-fidelity points early on and therefore collects many more observations (at lower fidelities).

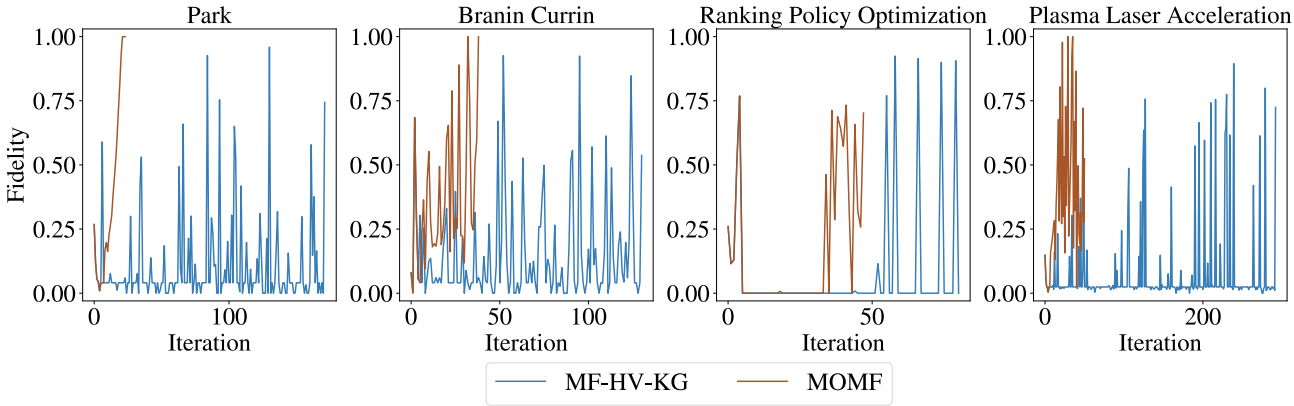

Figure 17: A comparison of which fidelities each algorithm chooses to query at each iteration.

E. On Pareto Subset Selection

In general, the quality of a finite approximation of a larger (potentially infinite) Pareto frontier is often assessed by additive (and multiplicative) approximation ratios (Bringmann and Friedrich, 2013), which are the minimum added value (and

multiplier, respectively) that when applied to all points in the approximate Pareto frontier yield a frontier that is at least a good as all points on the true Pareto frontier. In the bi-objective case, the hypervolume maximizing set enjoys the optimal additive (and multiplicative) approximation ratio(s) asymptotically in N_p (Bringmann and Friedrich, 2013).