Probability Distribution of Hypervolume Improvement in Bi-objective Bayesian Optimization

Hao Wang [1] Kaifeng Yang [2] Michael Affenzeller [2]

Abstract

Hypervolume improvement (HVI) is commonly employed in multi-objective Bayesian optimization algorithms to define acquisition functions due to its Pareto-compliant property. Rather than focusing on specific statistical moments of HVI, this work aims to provide the exact expression of HVI's probability distribution for bi-objective problems. Considering a bi-variate Gaussian random variable resulting from Gaussian process (GP) modeling, we derive the probability distribution of its hypervolume improvement via a cell partition-based method. Our exact expression is superior in numerical accuracy and computation efficiency compared to the Monte Carlo approximation of HVI's distribution. Utilizing this distribution, we propose a novel acquisition function - ε-probability of hypervolume improvement (ε-PoHVI). Experimentally, we show that on many widely-applied bi-objective test problems, ε-PoHVI significantly outperforms other related acquisition functions, e.g., ε-PoI, and expected hypervolume improvement, when the GP model exhibits a large the prediction uncertainty.

1. Introduction

For solving black-box multi-objective optimization problems (MOPs), the hypervolume indicator (HV) (Zitzler & Thiele, 1999) is extensively employed for assessing the quality of the Pareto front approximation or guiding the search direction. HV is defined as the Lebesgue measure of the subset of $\mathbb{R}^m$ dominated by an approximation set to the Pareto front. It is extensively applied in many multi-objective optimization algorithms, e.g., indicator-based evolutionary algorithms (Beume et al., 2007; Deb et al., 2002) and Bayesian optimization (Emmerich et al., 2016; Yang et al., 2019b; Daulton et al., 2020; Emmerich et al., 2020; Daulton et al., 2020; 2021; Suzuki et al., 2020; Garrido-Merchán et al., 2023; Zuluaga et al., 2016; Yang et al., 2016; 2022). In multi-objective Bayesian optimization, HV induces the famous hypervolume improvement (HVI) function (Emmerich, 2005), which quantifies the benefit of appending a new data point to the approximation set - the increment of the HV value caused by the new point. HVI generalizes the notion of "improvement" in the single-objective scenarios, and therefore, it serves as the base of many successful multi-objective acquisition functions, e.g., the probability of improvement (PoI) (Emmerich et al., 2006; Keane, 2006) that measures the chance of realizing nonzero HVI, ε-PoI that computes the probability of objective points having least ε distance to the approximation set, and the expected hypervolume improvement (EHVI) (Emmerich et al., 2006; Yang et al., 2019b) that generalizes expected improvement (EI) (Jones et al., 1998) from single-objective Bayesian optimization.

Motivation At an unknown decision point, the predictive/posterior distribution of its objective value follows a multivariate normal distribution in Bayesian optimization. Consequently, HVI defined on this predictive distribution is a real-valued random variable. To the best of our knowledge, the existing multi-objective acquisition functions either only consider the first moment of HVI's distribution, e.g., EHVI for the mean, or completely disregard the distribution, e.g., ε-PoI, which is not related to the quantile of HVI's distribution (see Sec. 4). However, when HVI shows a large dispersion (e.g., large variance), only relying on the mean of HVI makes the acquisition function less trustworthy/meaningful. In this sense, higher moments or at least the quantiles can help quantify the risk/uncertainty of the acquisition value (Mehlawat et al., 2021; Schonlau et al., 1998), which is difficult to obtain without the exact expression of HVI's distribution.

Contributions To answer the issue we raise above, we aim to provide the exact distribution function of HVI and demonstrate its usefulness in Bayesian optimization for speeding up empirical convergence. Our contributions are

[1] Leiden University, Leiden, The Netherlands [2] University of Applied Sciences, Hagenberg, Austria. Correspondence to: Kaifeng Yang <kaifeng.yang@fh-hagenberg.at>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

summarized as follows:

1. We derive the exact expression of HVI's distribution function in the bi-objective optimization scenario and numerically validate it against the Monte Carlo (MC) method. The exact distribution exhibits better computational efficiency and numerical accuracy (Sec. 3)

2. We propose a novel acquisition function, ε-Probability of Hypervolume Improvement (ε-PoHVI), which utilizes HVI's distribution function directly. It computes the probability of making at least ε hypervolume improvement to the current approximation set (Sec. 4).

3. We compare ε-PoHVI, ε-PoI, and EHVI on 14 selected test problems, where we observe ε-PoHVI substantially improves the empirical convergence of Bayesian optimization over ε-PoI and EHVI (Sec. 5).

2. Preliminaries

Multi-objective optimization A real-valued multi-objective optimization problem (MOP) involves minimizing multiple objective functions simultaneously, i.e., $\mathbf{f} = (f_1, \ldots, f_m), f_i : \mathcal{X} \rightarrow \mathbb{R}, \mathcal{X} \subseteq \mathbb{R}^d, i \in [1..m]$. For every $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)} \in \mathbb{R}^m$, we say $\mathbf{y}^{(1)}$ weakly dominates $\mathbf{y}^{(2)}$ (written as $\mathbf{y}^{(1)} \preceq \mathbf{y}^{(2)}$) iff. $y_i^{(1)} \leq y_i^{(2)}, i \in [1..m]$. The Pareto order $\prec$ on $\mathbb{R}^m$ is defined: $\mathbf{y}^{(1)} \prec \mathbf{y}^{(2)}$ iff. $\mathbf{y}^{(1)} \preceq \mathbf{y}^{(2)}$ and $\mathbf{y}^{(1)} \neq \mathbf{y}^{(2)}$. A point $\mathbf{x} \in \mathcal{X}$ is efficient iff. $\nexists \mathbf{x}' \in \mathcal{X}(\mathbf{f}(\mathbf{x}') \prec \mathbf{f}(\mathbf{x}))$. The set of all efficient points of $\mathcal{X}$ is called the *efficient set*. The image of the efficient set under $\mathbf{f}$ is called the *Pareto front*. Multi-objective optimization algorithms often employ a finite multiset $\mathrm{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ to approximate the efficient set, whose image under $\mathbf{f}$ is denoted by Y. The *non-dominated subset* of Y is a finite approximation to the Pareto front, which is denoted by $\mathcal{P}$. *Non-dominated space* w.r.t. $\mathcal{P}$ is the subset of $\mathbb{R}^m$ that is not dominated by $\mathcal{P}$, i.e., $\mathrm{ndom}(\mathcal{P}) := \{\mathbf{y} \in \mathbb{R}^m : \nexists \mathbf{p} \in \mathcal{P}(\mathbf{p} \prec \mathbf{y})\}$. Similarly, the *dominated space* w.r.t. $\mathcal{P}$, denoted by $\mathrm{dom}(\mathcal{P})$, is the complement of $\mathrm{ndom}(\mathcal{P})$.

Bayesian Optimization BO (Mockus, 1974; Jones et al., 1998; Shahriari et al., 2016) is a sequential model-based optimization algorithm for solving black-box optimization problems that are expensive to evaluate. BO starts with sampling a small initial set of data points $\mathrm{X} \subseteq \mathcal{X}$ (with Latin Hypercube Sampling). After evaluating X with $\mathbf{f}$, it constructs a probabilistic model $\Pr(\mathbf{f} \mid \mathrm{X}, \mathrm{Y})$ (e.g., Gaussian process regression). BO quantifies the quality of unseen points with an acquisition function, which targets balancing exploration and exploitation of the search process. BO chooses the next point to evaluate by maximizing the acquisition function. Please see (Knowles, 2006; Emmerich et al., 2016; 2020; Daulton et al., 2020; Belakaria et al., 2019; Zhang & Golovin, 2020; Tu et al., 2022; Garrido-Merchán et al., 2023) for more details and developments on this topic.

Gaussian process regression In this work, we model each objective function independently as the realization of a centered Gaussian process (GP) prior (Rasmussen & Williams, 2006), i.e., $f_i \sim \mathrm{gp}(0, k_i), i \in [1..m]$, where $k_i : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a kernel function that models the auto-covariance of f_i, $\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}, \mathrm{Cov}\{f_i(\mathbf{x}), f_i(\mathbf{x}')\} = k_i(\mathbf{x}, \mathbf{x}')$. Given a data set $\mathcal{D} = (\mathrm{X}, \mathrm{Y})$, $\mathrm{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\} \subset \mathcal{X}$, and $\mathrm{Y} = \{\mathbf{f}(\mathbf{x}^{(1)}), \ldots, \mathbf{f}(\mathbf{x}^{(n)})\}$, the posterior GPs are independent: $f_i \mid \mathcal{D} \sim \mathrm{gp}(\hat{f}_i, \hat{k}_i)$, where $\hat{f}_i$ and $\hat{k}_i$ are the posterior means and kernel functions, respectively. Many works have been devoted to model cross-correlations among GPs (Álvarez et al., 2012), e.g., multi-task GP (Bonilla et al., 2007) and dependent GP (Boyle & Frean, 2004).

Hypervolume Improvement The hypervolume (HV) indicator of a set $\mathcal{P} \subseteq \mathbb{R}^m$ is defined as the Lebesgue measure λ of the set that is dominated by $\mathcal{P}$ and bounded from above by a reference point $\mathbf{r} \in \mathbb{R}^m$, i.e., $\mathrm{HV}(\mathcal{P}, \mathbf{r}) = \lambda(\{\mathbf{y} \in \mathbb{R}^m : \mathbf{y} \prec \mathbf{r} \wedge \exists \mathbf{p} \in \mathcal{P}, \mathbf{p} \prec \mathbf{y}\})$. The hypervolume indicator is often taken as a performance metric for comparing the empirical performance for multi-objective optimization algorithms (Zitzler et al., 2003) or used in the indicator-based optimization algorithms (Beume et al., 2007). In Bayesian optimization, the set $\mathcal{P}$ is typically the approximation set - the non-dominated subset of the data $\mathcal{D}$. The contribution of a single objective vector $\mathbf{y}$ to $\mathcal{P}$ can be quantified by the well-known *hypervolume improvement* (HVI):

$$\Delta^+(\mathbf{y}; \mathcal{P}, \mathbf{r}) = \mathrm{HV}(\mathcal{P} \cup \{\mathbf{y}\}, \mathbf{r}) - \mathrm{HV}(\mathcal{P}, \mathbf{r}). \quad (1)$$

Note that, for a dominated point, i.e., $\mathbf{y} \in \mathrm{dom}(\mathcal{P})$, its HVI value is zero. Here, we introduce a "plus" symbol in Δ^+ to indicate that HVI is non-negative. The purpose of this notation shall become clear when we propose a generalization to the definition of HVI (see Sec. 3). HVI underpins many useful acquisition functions in Bayesian optimization. For instance, *Probability of Improvement* (PoI) (Stuckman, 1988) is originally devised for single-objective optimization cases and later generalized to multi-objective optimization (Emmerich et al., 2006; Keane, 2006). It quantifies the probability that $\mathbf{y}$ lies in the non-dominated space w.r.t. $\mathcal{P}$: $\mathrm{PoI}(\mathbf{x}; \mathcal{P}) = \mathrm{E}\{\mathbb{1}_{\mathrm{ndom}(\mathcal{P})}(\mathbf{y}) \mid \mathcal{D}, \mathbf{x}\}$. Also, *$\varepsilon$-Probability of Improvement* (ε-PoI) is proposed recently (Emmerich et al., 2020) to make the search less exploitative. It measures the probability of the non-dominated points that are at least ε away from the approximation set $\mathcal{P}$:

$$\varepsilon\text{-}\mathrm{PoI}(\mathbf{x}; \mathcal{P}, \varepsilon) = \mathrm{E}\left\{\mathbb{1}_{\mathrm{ndom}(\mathcal{P})}(\mathbf{y} + \varepsilon \mathbf{1}_m) \mid \mathcal{D}, \mathbf{x}\right\},$$

where $\mathbf{1}_m$ is an m-dimensional vector of 1's. The computational complexity of ε-PoI is $\Theta(n \log n)$ for $m = 2, 3$ (Yang et al., 2017; Emmerich et al., 2020) and

$\mathcal{O}(2^{m-1}n^{\lfloor \frac{m}{2} \rfloor})$ for $m \geq 4$ (Yang et al., 2019a), where $n = |\mathcal{P}|$. Furthermore, *Expected hypervolume improvement* (Emmerich et al., 2006) calculates the expectation of the HVI value of a multivariate Gaussian random variable: $\mathrm{EHVI}(\mathbf{x}; \mathcal{P}, \mathbf{r}) := \mathrm{E}\{\Delta^+(\mathbf{y}; \mathcal{P}, \mathbf{r}) \mid \mathcal{D}, \mathbf{x}\}$. The time complexity of EHVI's computation is $\Theta(n \log n)$ for $m = 2, 3$ (Emmerich et al., 2016; Yang et al., 2017; 2019a) and $\mathcal{O}(2^{m-1}n^{\lfloor \frac{m}{2} \rfloor})$ for $m \geq 4$ (Yang et al., 2019a).

3. The Distribution of Hypervolume Improvement

Generalized hypervolume improvement Prior to deriving the distribution functions, we first propose to generalize the definition of HVI in Eq. (1) for assigning nonzero values to the dominated points. We define the negative hypervolume "improvement" of a dominated point $\mathbf{y}$ as the *negative* volume of the intersection of the set that dominates $\mathbf{y}$ with the set dominated by $\mathcal{P}$, namely,

$$\Delta^-(\mathbf{y}) = -\lambda\left(\{\mathbf{p} \in \mathbb{R}^m \colon \mathbf{p} \preceq \mathbf{y}\} \cap \mathrm{dom}(\mathcal{P})\right),$$

which penalizes the dominated points that are located far from $\mathcal{P}$. In Fig. 1, we depict an example for the negative HVI (point $\mathbf{b}$). We claim that the negative HVI is desirable/useful to compute for two major reasons: (1) It makes the acquisition optimization step more tractable when a large subset of objective points are dominated by $\mathcal{P}$ w.h.p., typically when $\mathcal{P}$ is quite close to the Pareto front. Since the HVI is nearly zero on this subset, the acquisition function defined on HVI (e.g., PoI and EHVI) will exhibit a large plateau on its optimization landscape, making it harder to maximize. The negative HVI practically mitigates this issue by turning plateaus into valleys. (2) HVI is also extensively employed in evolutionary multi-objective algorithms (EMOAs) (Deb et al., 2002; Beume et al., 2007), where it is necessary to assign nonzero values to dominated points to move such points to the Pareto front. Our extension, the negative HVI, facilitates a sensible comparison among dominated points, which is a more natural extension compared to other existing proposals (Wang et al., 2015). We combine the negative HVI with Eq. (1), resulting in the *generalized hypervolume improvement*:

$$\Delta(\mathbf{y}) = \mathbb{1}_{\mathrm{ndom}(\mathcal{P})}(\mathbf{y})\Delta^+(\mathbf{y}) + \mathbb{1}_{\mathrm{dom}(\mathcal{P})}(\mathbf{y})\Delta^-(\mathbf{y}). \quad (2)$$

Importantly, if one wishes to focus only on the non-dominated points, then the above generalization is exactly the same with Eq. (1).

Assume a data set $\mathcal{D} = (\mathrm{X}, \mathrm{Y})$ observed on the vector-valued objective function. The approximation set $\mathcal{P} \subset \mathbb{R}^m$ to the Pareto front is the non-dominated subset of Y. Also, we assume a reference point $\mathbf{r} \in \mathbb{R}^m$. For the bi-objective case ($m = 2$), we depict an example of the HVI in Fig. 1. In this example, the random point $\mathbf{y}$ follows the posterior distribution of a Gaussian process model, i.e., $\mathbf{y} \mid \mathcal{D}, \mathbf{x} \sim \mathcal{N}(\hat{\mathbf{f}}(\mathbf{x}), \mathrm{diag}(\hat{k}_1(\mathbf{x}, \mathbf{x}), \hat{k}_2(\mathbf{x}, \mathbf{x})))$. It can be observed from the figure that the expression of $\Delta^+(\mathbf{y})$ depends on the subset of $\mathcal{P}$ that it dominates, indicating the expression of $\Delta^+(\mathbf{y})$ varies across realizations of $\mathbf{y}$, which brings the difficulty of deriving the distribution function. Note that HVI on $\mathbb{R}^m$ is actually a piecewise-defined function. It suffices to first identify the set on which the restriction of Δ^+ admits a fixed expression and then derive the conditional distribution function of HVI on such a set. As the first step, we provide a characterization of such a set.

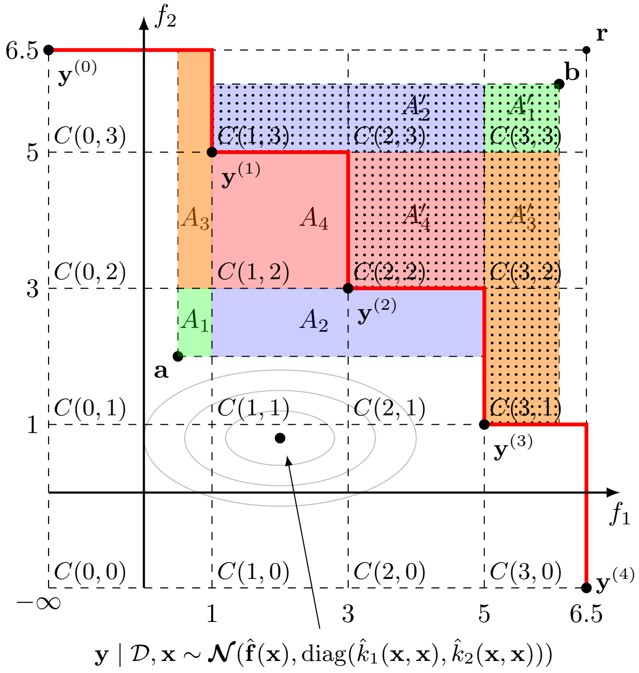


Figure 1. For a two-dimensional objective space, we picture the augmented Pareto approximation set $\widetilde{\mathcal{P}}$ by the black dots $\mathbf{y}^{(0)}, \ldots, \mathbf{y}^{(4)}$ and the attainment boundary by the red curve. The posterior distribution of $\mathbf{y}$ at a point $\mathbf{x} \in \mathcal{X}$ is illustrated by the light gray ellipsoids. The generalized hypervolume improvement of two realizations $\mathbf{a}$ and $\mathbf{b}$ are depicted in the shaded area. The objective space $[-\boldsymbol{\infty}, \mathbf{r}]$ is partitioned into cells (e.g., $C(1, 0)$). When restricting the random point $\mathbf{y}$ to a cell, its hypervolume can always be expressed in four terms: $\Delta^+(\mathbf{a})|_{C(0,1)} = \lambda(\boxed{A_1}) + \lambda(\boxed{A_2}) + \lambda(\boxed{A_3}) + \lambda(\boxed{A_4})$ (λ is the Lebesgue measure on $\mathbb{R}^2$). When a point is dominated by $\widetilde{\mathcal{P}}$, its negative hypervolume improvement can be written similarly: $\Delta^-(\mathbf{b})|_{C(3,3)} = -\lambda(\boxed{A'_1}) - \lambda(\boxed{A'_2}) - \lambda(\boxed{A'_3}) - \lambda(\boxed{A'_4})$.

Lemma 3.1. *Given a Pareto approximation set $\mathcal{P} \subset \mathbb{R}^m$ and a compact and connected set $\mathrm{S} \subset \mathbb{R}^m$ that dominates $\mathcal{P}$. If every point in S dominates the same subset of $\mathcal{P}$, then the restriction $\Delta^+|_S$ is not a piecewise function and is continuous.*

Proof. Let $U = \{\mathbf{p} \in \mathcal{P} \colon \forall \mathbf{s} \in S, \mathbf{s} \prec \mathbf{p}\}$. Based on the assumption that every point in S dominates the same subset U of $\mathcal{P}$, we could reformulate for every point $\mathbf{s}$, its hypervolume improvement, namely $\Delta^+(\mathbf{s}) = \mathrm{HV}((\mathcal{P} \setminus U) \cup \{\mathbf{s}\}) - \mathrm{HV}(\mathcal{P})$. Since $\mathcal{P} \setminus U$ and $\{\mathbf{s}\}$ are mutually non-dominated by construction, $\mathrm{HV}((\mathcal{P} \setminus U) \cup \{\mathbf{s}\})$ is not piecewise-defined and continuous. □

Remark 3.2. Obviously, the largest sets satisfying Lemma 3.1 are the hyperboxes constructed by the intersection of the $m-1$ dimensional hyperplanes passing through each point in $\mathcal{P}$. It is easy to verify that the above lemma also applies to the negative HVI. For the bi-objective case ($m = 2$), we show an example of those cells in Fig. 1.

In the paper, we focus on the bi-objective case, and we use the following notations for convenience: $\forall \mathbf{x} \in \mathcal{X}$, $\mu_1 = \hat{f}_1(\mathbf{x}), \mu_2 = \hat{f}_2(\mathbf{x})$, and $\sigma_1^2 = \hat{k}_1(\mathbf{x}, \mathbf{x}), \sigma_2^2 = \hat{k}_2(\mathbf{x}, \mathbf{x})$.

Cell partition of the objective space In bi-objective scenarios, we partition the objective space w.r.t. the approximation points. For a finite approximation set $\mathcal{P}$ of n points, we consider the extended real line $\mathbb{R} \cup \{-\infty, \infty\}$ and use it to augment $\mathcal{P}$ with two extreme points $(-\infty, r_2)^\top$ and $(r_1, -\infty)^\top$, i.e., $\widetilde{\mathcal{P}} = \mathcal{P} \cup \{(r_1, -\infty), (-\infty, r_2)\}$. Without loss of generality, we index the approximation points of $\widetilde{\mathcal{P}}$ in the increasing order w.r.t. their first objective values, i.e., $\mathbf{y}^{(0)}, \mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(n+1)}$ where $-\infty = y_1^{(0)} < y_1^{(1)} < \cdots < y_1^{(n+1)} = r_1$. We denote by

$$C(i,j) = \left[y_1^{(i)}, y_1^{(i+1)}\right] \times \left[y_2^{(n-j+1)}, y_2^{(n-j)}\right], i, j \in [0..n],$$

the cell area bounded by axis-parallel lines that pass through the points in $\mathcal{P}$. For instance, in Fig. 1, where $n = 3$, $C(0,1) = [y_1^{(0)}, y_1^{(1)}] \times [y_2^{(3)}, y_2^{(2)}]$. In the following discussion, we will use the minimum $\mathbf{l}^{(i,j)}$ and maximum $\mathbf{u}^{(i,j)}$ of $C(i,j)$, namely,

$$\mathbf{l}^{(i,j)} = \left(y_1^{(i)}, y_2^{(n-j+1)}\right)^\top, \ \mathbf{u}^{(i,j)} = \left(y_1^{(i+1)}, y_2^{(n-j)}\right)^\top.$$

In this manner, the entire objective space is partitioned into $(n+1)^2$ cells: $[-\boldsymbol{\infty}, \mathbf{r}] = \cup_{i,j \in [0..n]} C(i,j)$. Note that, for $i + j \leq n$, the union of the cells $C(i,j)$ represents the subset that is not dominated by $\widetilde{\mathcal{P}}$, i.e., $\mathrm{ndom}(\widetilde{\mathcal{P}}) = \cup_{i+j \leq n} C(i,j)$ while for $i + j > n$, it indicates the subset dominated by $\widetilde{\mathcal{P}}$.

Conditional probability density function Taking the cell decomposition, we can express the distribution functions of generalized HVI by marginalizing the conditional distributions over cells:

$$\begin{aligned} F_{\Delta(\mathbf{y})|\mathcal{D}}(\delta) &= \sum_{i,j \in [0..n]} F_{\Delta(\mathbf{y})}^{(i,j)}(\delta) \Pr\left(\mathbf{y} \in C(i,j) \mid \mathcal{D}, \mathbf{x}\right), \\ &= \sum_{i,j \in [0..n]} F_{\Delta(\mathbf{y})}^{(i,j)}(\delta) \left[\Phi_{\mu_1,\sigma_1}\left(u_1^{(i,j)}\right) - \Phi_{\mu_1,\sigma_1}\left(l_1^{(i,j)}\right)\right] \\ &\times \left[\Phi_{\mu_2,\sigma_2}\left(u_2^{(i,j)}\right) - \Phi_{\mu_2,\sigma_2}\left(l_2^{(i,j)}\right)\right], \end{aligned} \tag{3}$$

where $F_{\Delta(\mathbf{y})}^{(i,j)} \in \{\mathrm{PDF}_{\Delta(\mathbf{y})}^{(i,j)}, \mathrm{CDF}_{\Delta(\mathbf{y})}^{(i,j)}\}$ denotes either the cumulative distribution function (CDF) or the probability density function (PDF) of HVI conditioned on cell $C(i,j)$. Note that within a non-dominated cell $C(i,j)$ ($i + j \leq n$), the conditional HVI takes it value in $\left[\Delta^+(\mathbf{u}^{(i,j)}), \Delta^+(\mathbf{l}^{(i,j)})\right]$. Similarly, within dominated cell $C(i,j)$ ($i + j > n$), the HVI's range is $\left[\Delta^-(\mathbf{u}^{(i,j)}), \Delta^-(\mathbf{l}^{(i,j)})\right]$. We shall proceed to derive the conditional PDF as follows.

Theorem 3.3. *Assume the above cell partition of the objective space w.r.t. augmented approximation set $\widetilde{\mathcal{P}}$. For $0 \leq i + j \leq n$, where $i, j \in [0..n]$, the hypervolume improvement of $\mathbf{y} = (y_1, y_2)^\top \in \mathbb{R}^2$ restricted to cell $C(i,j)$ can be expressed as follows:*

$$\Delta^+(\mathbf{y})|_{C(i,j)} = \lambda(A_1) + \lambda(A_2) + \lambda(A_3) + \lambda(A_4),$$

where λ is the Lebesgue measure on $\mathbb{R}^2$ and

$$\begin{aligned} A_1 &= \left[y_1, u_1^{(i,j)}\right] \times \left[y_2, u_2^{(i,j)}\right], \\ A_2 &= \left[u_1^{(i,j)}, y_1^{(n-j+1)}\right] \times \left[y_2, u_2^{(i,j)}\right], \\ A_3 &= \left[y_1, u_1^{(i,j)}\right] \times \left[u_2^{(i,j)}, y_2^{(i)}\right], \\ A_4 &= \mathrm{dom}\left(\left\{\mathbf{u}^{(i,j)}\right\}\right) \setminus \mathrm{dom}\left(\widetilde{\mathcal{P}}\right). \end{aligned}$$

Proof. An illustration has been given in Fig. 1. Let $S = \mathrm{dom}(\{\mathbf{y}\}) \setminus \mathrm{dom}(\widetilde{\mathcal{P}})$ denote the set that is dominated by $\mathbf{y}$ but not by $\widetilde{\mathcal{P}}$. We first consider the set A_4, the subset dominated by the maximum point $\mathbf{u}^{(i,j)}$, which is clearly contained in S as $\mathbf{y}|_{C(i,j)} \preceq \mathbf{u}^{(i,j)}$. Note that, when $\mathbf{y}$ is restricted to $C(i,j)$, its projection along f_1 onto the attainment boundary is always $(y_1^{(n-j+1)}, y_2)^\top$; its projection along f_2 onto the attainment boundary is always $(y_1, y_2^{(i)})^\top$. Hence, we can express the reminder set $S \setminus A_4 = [y_1, y_1^{(n-j+1)}] \times [y_2, u_2^{(i,j)}] + [y_1, u_1^{(i,j)}] \times [y_2, y_2^{(i)}] - [y_1, u_1^{(i,j)}] \times [y_2, u_2^{(i,j)}] = A_1 + A_2 + A_3$. □

Based on Thm. 3.3, we can express the HVI restricted to cell $C(i,j)$ as:

$$\begin{aligned}\Delta^+(\mathbf{y})|_{C(i,j)} &= (u_1^{(i,j)} - y_1)(u_2^{(i,j)} - y_2)\\ &\quad + (y_1^{(n-j+1)} - u_1^{(i,j)})(u_2^{(i,j)} - y_2)\\ &\quad + (u_1^{(i,j)} - y_1)(y_2^{(i)} - u_2^{(i,j)}) + \Delta^+(\mathbf{u}^{(i,j)})\\ &= z_1 z_2 + \gamma^{(i,j)},\end{aligned}$$

where $z_1 = y_1^{(n+1-j)} - y_1, z_2 = (y_2^{(i)} - y_2), \gamma^{(i,j)} = \Delta^+(\mathbf{u}^{(i,j)}) - (y_1^{(n+1-j)} - u_1^{(i,j)})$. Note that $z_1 \sim \mathcal{N}(y_1^{(n-j+1)} - \mu_1, \sigma_1^2)$ and $z_2 \sim \mathcal{N}(y_2^{(i)} - \mu_2, \sigma_2^2)$ are Gaussian random variables truncated to $[L_1, U_1] \coloneqq [y_1^{(n-j+1)} - u_1^{(i,j)}, y_1^{(n-j+1)} - l_1^{(i,j)}]$ and $[L_2, U_2] \coloneqq [y_2^{(i)} - u_2^{(i,j)}, y_2^{(i)} - l_2^{(i,j)}]$, respectively. Thereby, the distribution of $\Delta^+(\mathbf{y})|_{C(i,j)}$ can be expressed via that of products of truncated Gaussians.

Let $\mu_1' = y_1^{(n+1-j)} - \mu_1$, $\mu_2' = y_2^{(i)} - \mu_2$. We have the following expression of HVI's distribution for the non-dominated cells: $\forall \delta \in \left[\Delta^+(\mathbf{u}^{(i,j)}), \Delta^+(\mathbf{l}^{(i,j)})\right]$,

$$\begin{aligned}\mathrm{PDF}^{(i,j)}_{\Delta^+(\mathbf{y})}(\delta) &= \mathrm{PDF}_{z_1 z_2}(\delta - \gamma^{(i,j)})\\ &= D_1 D_2 \int_{\alpha(p)}^{\beta(p)} \phi_{\mu_1',\sigma_1}(x) \phi_{\mu_2',\sigma_2}\left(\frac{p}{x}\right) x^{-1} \mathrm{d}x, \quad (4)\\ p &= \delta - \gamma^{(i,j)},\\ D_1 &= [\Phi_{\mu_1',\sigma_1}(U_1) - \Phi_{\mu_1',\sigma_1}(L_1)]^{-1},\\ D_2 &= [\Phi_{\mu_2',\sigma_2}(U_2) - \Phi_{\mu_2',\sigma_2}(L_2)]^{-1},\end{aligned}$$

where $\phi_{\mu,\sigma}$ and $\Phi_{\mu,\sigma}$ denote the PDF and CDF of a Gaussian random variable with mean μ and standard deviation σ, respectively. The integration bounds are determined as follows. If $L_1 U_2 < U_1 L_2$ [1], we define:

$$[\alpha(p), \beta(p)] = \begin{cases} [L_1, p/L_2], & L_1 L_2 \leq p < L_1 U_2 \\ [p/U_2, p/L_2], & L_1 U_2 \leq p < U_1 L_2 \\ [p/U_2, U_1], & U_1 L_2 \leq p \leq U_1 U_2 \end{cases} \quad (5)$$

For a cell in the dominated space, i.e., $C(i,j)$ with $i + j > n$, we could invert the coordinate system to treat it in the same way as the non-dominated part (notice that $\mathbf{l}^{(i,j)}$ and $\mathbf{u}^{(i,j)}$ are swamped after inversion). Namely, we take the following inverted quantities:

$$\begin{aligned}&[L_1^-, U_1^-] \coloneqq [-U_1, -L_1],\\ &[L_2^-, U_2^-] \coloneqq [-U_2, -L_2],\\ &\mu_1^- \coloneqq -\mu_1', \quad \mu_2^- \coloneqq -\mu_2',\\ &\gamma^{(i,j)} = -\Delta^-(\mathbf{l}^{(i,j)}) - (l_1^{(i,j)} - y_1^{(n+1-j)}).\end{aligned}$$

It is straightforward to verify that the above derivation still holds for the inverted quantities. In this case, the conditional density function can be computed with Eq. (4): $\forall \delta \in [\Delta^-(\mathbf{u}^{(i,j)}), \Delta^-(\mathbf{l}^{(i,j)})]$,

$$\mathrm{PDF}^{(i,j)}_{\Delta^-(\mathbf{y})}(\delta) = \mathrm{PDF}^{(i,j)}_{\Delta^+(\mathbf{y})}(-\delta) = \mathrm{PDF}_{z_1 z_2}(-\delta - \gamma^{(i,j)}),$$

where we substitute the quantities (e.g., L_1) with the inverted ones (e.g., L_1^-) in $\mathrm{PDF}_{z_1 z_2}$.

Conditional cumulative distribution function Taking the conditional density function, the cumulative distribution of the hypervolume improvement can be derived for a non-dominated cell $C(i,j)$ $(i + j \leq n)$. For $\delta \in \left[\Delta^+(\mathbf{u}^{(i,j)}), \Delta^+(\mathbf{l}^{(i,j)})\right]$ and $p = \delta - \gamma^{(i,j)}$, we have:

$$\begin{aligned}\mathrm{CDF}^{(i,j)}_{\Delta^+(\mathbf{y})}(\delta) &= \int_{L_1 L_2}^{p} \mathrm{PDF}_{z_1 z_2}(x) \, \mathrm{d}x\\ &= D_1 D_2 \left(G + \int_{\alpha(p)}^{\beta(p)} \phi_{\mu_1',\sigma_1}(\zeta) \Phi_{\mu_2',\sigma_2}\left(\frac{p}{x}\right) \mathrm{d}x \right), \quad (6)\\ G &= \Phi_{\mu_2',\sigma_2}(U_2)\left[\Phi_{\mu_1',\sigma_1}(\alpha(p)) - \Phi_{\mu_1',\sigma_1}(L_1)\right] +\\ &\quad \Phi_{\mu_2',\sigma_2}(L_2)\left[\Phi_{\mu_1',\sigma_1}(L_1) - \Phi_{\mu_1',\sigma_1}(\beta(p))\right],\end{aligned}$$

where the integration bounds α, β are defined in Eq. (5). For a cell $C(i,j)$ in the dominated space $(i + j > n)$, we take the same trick of inverting the coordinate system as above: $\forall \delta \in [\Delta^-(\mathbf{u}^{(i,j)}), \Delta^-(\mathbf{l}^{(i,j)})]$,

$$\mathrm{CDF}^{(i,j)}_{\Delta^-(\mathbf{y})}(\delta) = 1 - \mathrm{CDF}^{(i,j)}_{\Delta^+(\mathbf{y})}(-\delta),$$

where we have to take the inverted quantities in Eq. (6).

Numerical computation We use the numerical integration[2] to compute the distribution function in each cell and set the absolute error of the integration to 10^{-8}. The time complexity of the PDF and CDF of HVI is quadratic w.r.t. the number of approximation points in $\mathcal{P}$ due to the quadratic number terms in Eq. (3). To reduce the time complexity, we propose to prune the computation on some cells: (1) the ones on which the probability mass of $\mathbf{y}$ is sufficiently small. We only include the cells that overlap with the range of $\mu \pm 3\sigma$ in the computation; (2) if the value of HVI to compute is out of the range of conditional HVI on a cell. In the left plot of Fig 2, we have illustrated an example of the CDF of HVI computed from both the exact distribution and the Monte Carlo (MC) method. It is necessary to compare the computational cost of the exact distribution to that of the MC method to approximate the cumulative distribution function. Generally, for achieving an accuracy of τ, the numerical integration requires $O(\tau^{-1})$ (Novak, 2014), resulting in an overall complexity of $O(\tau^{-1}(n+1)^2)$ for the exact method. In contrast, the MC method requires sampling $O(\tau^{-2})$ realizations of $\mathbf{y}$ and calculating the hypervolume improvement

[1] When $L_1 U_2 > U_1 L_2$, it suffices to swap variables z_1 and z_2, and apply Eq. (5).

[2] We employed the 21-point Gauss–Kronrod quadrature method with maximally 50 sub-intervals.

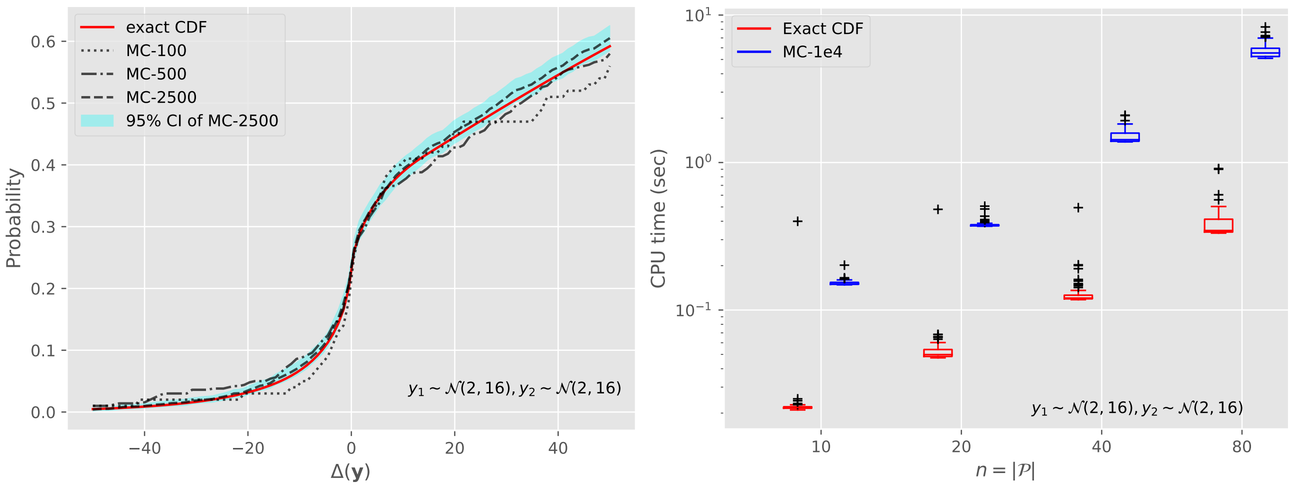


Figure 2. **Left**: For the Pareto front in Fig. 1, we show the CDF of $\mathbf{y}$ computed from the exact and MC methods (using 100, 500, and 2 500 samples). **Right**: The CPU time for the exact and the MC method (with 10^4 sample points) w.r.t. an increasing number of points of $\mathcal{P}$.

thereof[3], giving rise to the complexity of $O(\tau^{-2} n \ln n)$. Also, in the right plot of Fig. 2, we compare the CPU time of the exact expression to the MC computation when varying the number of points in the Pareto front, which shows that under a comparable numerical accuracy, the CPU time consumed by the exact computation is roughly one order of magnitude lower than that of the MC method, for a wide range of the cardinality of $\mathcal{P}$.

4. ε-Probability of Hypervolume Improvement

In this section, we leverage the commonly used ε-PoI function with HVI's distribution in order to demonstrate the usefulness thereof in Bayesian optimization. The ε-PoI function translates the mean of posterior distribution towards $\mathcal{P}$ by $\varepsilon \mathbf{1}_m$ and computes the probability of improving the current $\mathcal{P}$ (see its definition in Sec. 2). When computing ε-PoI, all objective points that are taken into account have a minimal distance of ε to the attainment boundary of $\mathcal{P}$. Despite its simplicity, we find it difficult to relate ε to the quantiles of HVI's distribution: from the posterior distribution, we can generate two sample points that both have ε minimal distance to the attainment boundary but differ hugely in their hypervolume improvements.

To mitigate this issue, we propose the *ε-Probability of Hypervolume Improvement* (ε-PoHVI) function, which computes the probability of making at least ε hypervolume improvement to $\mathcal{P}$. This probability can be computed directly by HVI's CDF function defined in Eq. (6):

$$\varepsilon\text{-PoHVI}(\mathbf{x}; \mathcal{P}, \mathbf{r}, \varepsilon) = 1 - \mathrm{CDF}_{\Delta(\mathbf{y})|\mathcal{D},\mathbf{x}}(\varepsilon), \quad (7)$$

where ε can be considered a lower quantile of HVI's distribution. This definition gives rise to two advantages: (1) When HVI exhibits a huge dispersion, ε-PoHVI is still safe to use, compared to EHVI, which would become less meaningful in this circumstance; (2) the user of Bayesian optimization can precisely control the level of the minimal improvement in each iteration. Compared to ε-PoI (time complexity: $\Theta(n \log n)$), our new acquisition function suffers from relatively higher computation overheads (quadratic; see "Numerical computation" above), and it requires the user to specify a reference point (for computing the hypervolume improvement), which is not needed in ε-PoI. Nevertheless, we are interested in whether the losses in the computation cost of ε-PoHVI can speed up the empirical convergence of Bayesian optimization on challenging problems. In practice, the free parameter ε in those two acquisition functions is either manually determined or tuned via hyperparameter tuning. In this paper, we also propose two control schemes:

- *ε-PoHVI-scaling* determines the parameter ε_t at iteration t with the schedule: $\varepsilon_t = \varepsilon_0 \exp(-ct)$, where $\varepsilon_0 = 0.05, c = 0.02$. We include, in the supplementary material, a rule-of-thumb to set the hyperparameter ε_0 and c. This schedule is motivated by the fact that when converging to the Pareto front, the steps of each approximation point tend to decrease.

- *ε-PoHVI-smoothing* exponentially smooths of the hypervolume improvement measured in the optimization: $\varepsilon_{t+1} = \alpha\left(\mathrm{HV}(\mathcal{P}_t, \mathbf{r}) - \mathrm{HV}(\mathcal{P}_{t-1}, \mathbf{r})\right) + (1-\alpha)\varepsilon_t$, where $\alpha = 0.5$ and $\varepsilon_0 = 0.05$. The intuition is to take the hypervolume improvement realized in the optimization history to set the ε value for the next iteration.

[3] It requires computing the hypervolume of the Pareto front approximation, which has a time complexity of $\Theta(n \log n)$ when $m = 2, 3$ (Beume et al., 2009).

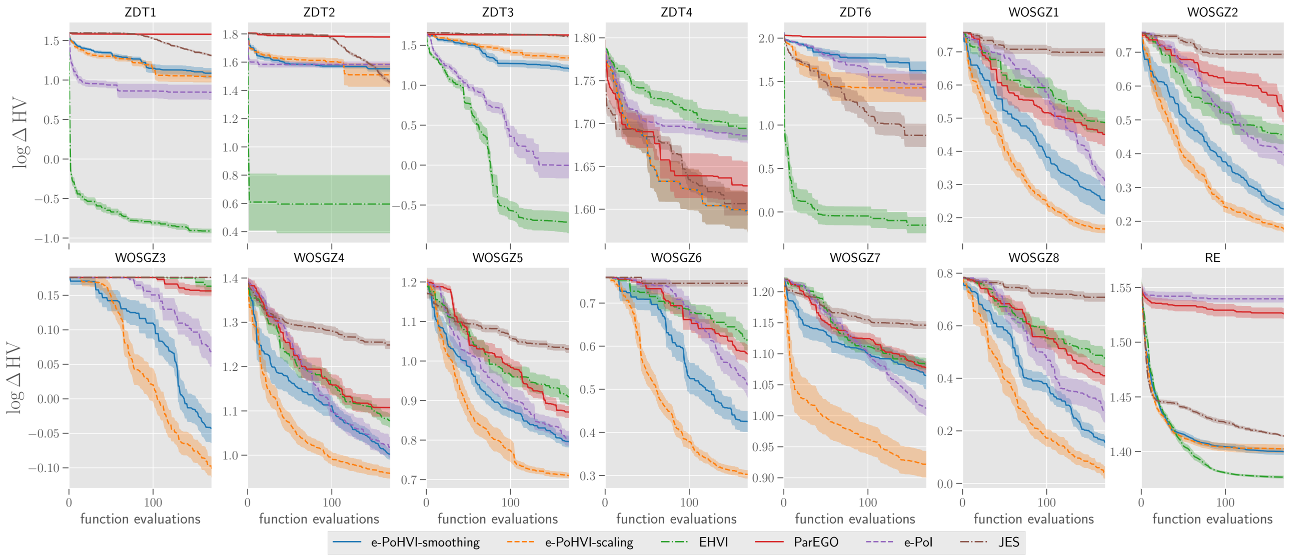


Figure 3. The log difference between the hypervolume of the best-so-far approximation set $\mathcal{P}$ and the target hypervolume over function evaluations. The target hypervolume is obtained with 1 000 points evenly sampled on the Pareto front of each problem. We show the mean and standard error of the log differences measured from 15 independent runs of each acquisition function on each problem.

5. Experiments

Experimental setup We investigate the empirical performance of ε-PoHVI against ε-PoI and EHVI on three sets of test problems: (1) the classical bi-objective ZDT problems (Zitzler et al., 2000), which have regular-shaped Pareto front (either convex or concave). We selected problems ZDT1-4 and 6 (ZDT5 is a discrete optimization problem); (2) WOSGZ1-8 problems (Wang et al., 2019) whose Pareto fronts are more difficult to approximate (WOSGZ9-16 are tri-objective problems); (3) a real-world problem - *four bar truss design* (Cheng & Li, 1999; Tanabe & Ishibuchi, 2020) (denoted as RE), which has a convex Pareto front and the ranges of two objective functions differ drastically. The decision space is set to $[0, 1]^{30}$ for ZDT1-3 and ZDT6, $[0, 1] \times [-5, 5]^{29}$ for ZDT4, $[0, 1] \times [-1, 1]^{29}$ for all WOSGZ problems, and $[1, 3] \times [1, \sqrt{2}] \times [1, \sqrt{2}] \times [1, 3]$ for RE. In the objective space, we take the reference point $\mathbf{r} = [15, 15]$ as recommended in (Zitzler et al., 2000) for ZDT problems when computing the hypervolume. The reference points are set to $[1.2, 1.2]$ and $[3\,000, 0.0383]$ for the WOSGZ and the RE problem, respectively, as suggested in (Wang et al., 2019; Lukovic et al., 2020).

We implement ε-PoI, ε-PoHVI, EHVI in the DGEMO (Lukovic et al., 2020) algorithmic framework[4] and test each of acquisition functions with 15 independent runs on each test problem. Also, we compare PoHVI with two commonly-used multi-objective Bayesian optimization algorithms: Joint Entropy Search (JES) (Tu et al., 2022) and ParEGO (Knowles, 2006).

We initialize the BO algorithm with $\min(60, 6d)$ points generated with Latin Hypercube sampling and terminate the algorithm at 170 iterations. We build two independent Gaussian processes for each objective with Matérn 5/2 kernel. We maximize the acquisition function in each iteration with covariance matrix adaptation evolution strategy (CMA-ES) algorithm (Hansen, 2006). Also, we take the above "scaling" control scheme to set the free parameter of ε-PoI.

Results and discussion In Fig. 3, we show the mean convergence and 95% confidence interval of the best-so-far hypervolume value of $\mathcal{P}$ for BO equipped with different bi-objective acquisition functions. Overall, we see (1) two versions of ε-PoHVI outperform ε-PoI and EHVI substantially on WOSGZ problems while EHVI takes the lead on ZDT and RE problems; (2) ε-PoHVI outperforms the classic ParEGO algorithm across all functions; (3) the Joint Entropy search (JES) only outperforms PoHVI on ZDT6.

Compared to WOSGZ, ZDT problems are considered relatively easier to solve since (1) the Pareto fronts are very regular - either convex or concave; (2) there are no local Pareto fronts (except ZDT4); (3) on the Pareto front, the optimal distribution (Auger et al., 2010) of points (w.r.t. HV) is mostly uniform. Similarly, the real-world problem RE has a much lower search dimensionality (four) with a convex Pareto front. WOSGZ problems are, however, designed to have more realistic Pareto fronts (non-convex/non-concave with non-uniform optimal distribution), which is shown to be challenging to solve or model (Wang et al., 2019).

[4] Our source code is available at `https://github.com/wangronin/HVI-distribution`

Table 1. Over all test problems, we perform the pairwise Wilcoxon's Rank-Sum test matrix at significance level 0.05 with p-value correction, where $+/\approx/-$ indicates that the algorithm in the column is significantly better/not different/worst than the ones in rows.

	e-PoI	**EHVI**	**e-PoHVI-smoothing**	**e-PoHVI-scaling**
e-PoI	n.a.	5/2/7	6/6/2	9/3/2
EHVI	7/2/5	n.a.	7/2/5	8/1/5
e-PoHVI-smoothing	2/6/6	5/2/7	n.a.	8/5/1
e-PoHVI-scaling	2/3/9	5/1/8	1/6/7	n.a.
Sum of $+/\approx/-$	11/11/20	15/5/22	14/14/14	**25/9/8**

With a small data set, we expect a much higher GP's prediction uncertainty on WOSGZ problems than ZDTs, which is validated with numerical observations: in Table 2, we list GP's prediction uncertainty (posterior standard deviations) for each objective function over decision points sampled u.a.r. from the decision space. We see, for instance, on WOSGZ1, the posterior standard deviation of the first objective is about $\sigma_1 = 1.5524$, which is orders of magnitude higher than that on ZDT1 ($\sigma_1 = 0.00837$). As a result, HVI's distribution on WOSGZs exhibits a high dispersion, and the EHVI function becomes less characteristic of the distribution. In contrast, ε-PoHVI is not affected by the large dispersion since it utilizes the quantile of HVI's distribution. Similarly, ε-PoI also suffers less from the high dispersion of HVI despite not connecting to the quantiles of HVI's distribution. *We conclude that ε-PoHVI is advantageous when the prediction uncertainty of GP is high, which occurs if the objective functions are challenging to model with a small data set.*

Between ε-PoI and ε-PoHVI, we see that ε-PoHVI (with both control schemes of ε) substantially improves the convergence speed and achieves better hypervolume values of the final approximation set $\mathcal{P}$ on all test problems except ZDT1, 2, and 4. ZDT1 is one of the simplest bi-objective problems. On ZDT2, although ε-PoI shows a faster initial convergence, ε-PoHVI manages to hit about the same mean hypervolume value in the last few iterations. On ZDT4, all acquisition functions fail to get close to the Pareto front since this problem has many local Pareto fronts. Much more function evaluations are needed for all acquisition functions. Also, we see that the scaling control (ε-PoHVI-scaling) performs better than the exponential smoothing (ε-PoHVI-smoothing) on all problems except ZDT3.
Furthermore, we perform a pairwise Wilcoxon's Rank-Sum test (with p-value correction for multiple testing) on all test problems to verify the statistical significance of the result in the convergence chart. In table 1, we show the test's outcome - each entry of the table ($+/\approx/-$) indicates out of 14 test problems, the number of cases where the algorithm in the first row is significantly better/not different/worst than the ones in the first column. Overall, ε-PoHVI-scaling outperforms both ε-PoI, ε-PoHVI-smoothing, and EHVI. Since ε-PoHVI-scaling is a comparable parameter control compared to that of ε-PoI. We can conclude that ε-PoHVI has better empirical performance than ε-PoI on the selected test problems. Finally, in the supplementary material, we have included the detailed descriptive statistics of the convergence of hypervolume values and the best, median, and worst $\mathcal{P}$ obtained by each method at the last iteration.

6. Conclusions

We first propose a generalization to hypervolume improvement (HVI), which assigns nonzero value to the dominated points. Then, we derive the exact expression of the distribution functions of the hypervolume improvement (HVI) in the Bayesian optimization setup, in which we utilize a cell partition of the objective space. Compared to the Monte Carlo approach, the numerical computation of the exact expression is computationally faster and numerically more accurate. Taking this distribution function, we propose a novel acquisition function, ε-Probability of Hypervolume Improvement, which shows a large empirical advantage over its counterparts.

The limitation of this work is two-fold: (1) the distribution functions are derived in the bi-objective scenario with uncorrelated Gaussian processes for each objective. In practice, users of Bayesian optimization often wish to solve more than two objectives with a multi-output Gaussian process model. For correlated Gaussian posteriors, our analytical results (Eq. (4) and (6)) can be directly applied by substituting the probability density in Eq. (4) with the one of the correlated Gaussian. As for more than two objectives, the cell partition approach can be applied in principle. However, the exact formulation of HVI's distribution is very complicated to accommodate in this work, given the page limit. (2) The experimental study of the proposed acquisition function can be enhanced by applying it to more real-world problems, such as industrial optimization and hyperparameter tuning tasks.

Acknowledgement

The authors would like to thank Michael Emmerich for various discussions and suggestions. This work is supported by the Austrian Science Fund (FWF – Der Wissenschaftsfonds) under the project (I 5315, 'ML Methods for Feature Identification Global Optimization).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Álvarez, M. A., Rosasco, L., and Lawrence, N. D. Kernels for Vector-Valued Functions: A Review. *Found. Trends Mach. Learn.*, 4(3):195–266, 2012. doi: 10.1561/2200000036. URL https://doi.org/10.1561/2200000036.

Auger, A., Bader, J., and Brockhoff, D. Theoretically Investigating Optimal μ-Distributions for the Hypervolume Indicator: First Results for Three Objectives. In Schaefer, R., Cotta, C., Kolodziej, J., and Rudolph, G. (eds.), *Parallel Problem Solving from Nature - PPSN XI, 11th International Conference, Kraków, Poland, September 11-15, 2010, Proceedings, Part I*, volume 6238 of *Lecture Notes in Computer Science*, pp. 586–596. Springer, 2010. doi: 10.1007/978-3-642-15844-5_59. URL https://doi.org/10.1007/978-3-642-15844-5_59.

Belakaria, S., Deshwal, A., and Doppa, J. R. Max-value Entropy Search for Multi-Objective Bayesian Optimization. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 7823–7833, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/82edc5c9e21035674d481640448049f3-Abstract.html.

Beume, N., Naujoks, B., and Emmerich, M. T. M. SMS-EMOA: multiobjective selection based on dominated hypervolume. *Eur. J. Oper. Res.*, 181(3):1653–1669, 2007. doi: 10.1016/j.ejor.2006.08.008. URL https://doi.org/10.1016/j.ejor.2006.08.008.

Beume, N., Fonseca, C. M., López-Ibáñez, M., Paquete, L., and Vahrenhold, J. On the Complexity of Computing the Hypervolume Indicator. *IEEE Trans. Evol. Comput.*, 13 (5):1075–1082, 2009. doi: 10.1109/TEVC.2009.2015575. URL https://doi.org/10.1109/TEVC.2009.2015575.

Bonilla, E. V., Chai, K. M. A., and Williams, C. K. I. Multi-task Gaussian Process Prediction. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T. (eds.), *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pp. 153–160. Curran Associates, Inc., 2007.

Boyle, P. and Frean, M. R. Dependent Gaussian Processes. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, pp. 217–224, 2004.

Cheng, F. Y. and Li, X. S. Generalized center method for multiobjective engineering optimization. *Engineering Optimization*, 31(5):641–661, 1999. doi: 10.1080/03052159908941390. URL https://doi.org/10.1080/03052159908941390.

Daulton, S., Balandat, M., and Bakshy, E. Differentiable Expected Hypervolume Improvement for Parallel Multi-Objective Bayesian Optimization. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Daulton, S., Balandat, M., and Bakshy, E. Parallel Bayesian Optimization of Multiple Noisy Objectives with Expected Hypervolume Improvement. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 2187–2200, 2021.

Deb, K., Agrawal, S., Pratap, A., and Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.*, 6(2):182–197, 2002. doi: 10.1109/4235.996017. URL https://doi.org/10.1109/4235.996017.

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

Emmerich, M., Yang, K., Deutz, A. H., Wang, H., and Fonseca, C. M. A Multicriteria Generalization of Bayesian Global Optimization. In Pardalos, P. M., Zhigljavsky, A., and Zilinskas, J. (eds.), *Advances in Stochastic and Deterministic Global Optimization*, volume 107 of *Springer Optimization and Its Applications*, pp. 229–242. Springer, 2016. doi: 10.1007/

978-3-319-29975-4_12. URL https://doi.org/10.1007/978-3-319-29975-4_12.

Emmerich, M. T. M., Giannakoglou, K. C., and Naujoks, B. Single- and multiobjective evolutionary optimization assisted by Gaussian random field metamodels. *IEEE Trans. Evol. Comput.*, 10(4):421–439, 2006. doi: 10.1109/TEVC.2005.859463. URL https://doi.org/10.1109/TEVC.2005.859463.

Emmerich, M. T. M., Yang, K., and Deutz, A. H. Infill Criteria for Multiobjective Bayesian Optimization. In Bartz-Beielstein, T., Filipic, B., Korosec, P., and Talbi, E. (eds.), *High-Performance Simulation-Based Optimization*, volume 833 of *Studies in Computational Intelligence*, pp. 3–16. Springer, 2020. doi: 10.1007/978-3-030-18764-4_1. URL https://doi.org/10.1007/978-3-030-18764-4_1.

Garrido-Merchán, E. C., Fernández-Sánchez, D., and Hernández-Lobato, D. Parallel predictive entropy search for multi-objective Bayesian optimization with constraints applied to the tuning of machine learning algorithms. *Expert Syst. Appl.*, 215:119328, 2023. doi: 10.1016/j.eswa.2022.119328. URL https://doi.org/10.1016/j.eswa.2022.119328.

Hansen, N. The CMA evolution strategy: A comparing review. In Lozano, J. A., Larrañaga, P., Inza, I., and Bengoetxea, E. (eds.), *Towards a New Evolutionary Computation - Advances in the Estimation of Distribution Algorithms*, volume 192 of *Studies in Fuzziness and Soft Computing*, pp. 75–102. Springer, 2006. doi: 10.1007/3-540-32494-1_4. URL https://doi.org/10.1007/3-540-32494-1_4.

Jones, D. R., Schonlau, M., and Welch, W. J. Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, 13(4):455–492, 1998. doi: 10.1023/A:1008306431147.

Keane, A. J. Statistical improvement criteria for use in multiobjective design optimization. *AIAA journal*, 44(4): 879–891, 2006.

Knowles, J. D. ParEGO: a hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE Trans. Evol. Comput.*, 10(1):50–66, 2006. doi: 10.1109/TEVC.2005.851274. URL https://doi.org/10.1109/TEVC.2005.851274.

Lukovic, M. K., Tian, Y., and Matusik, W. Diversity-guided multi-objective bayesian optimization with batch evaluations. *Advances in Neural Information Processing Systems*, 33:17708–17720, 2020.

Mehlawat, M. K., Gupta, P., and Khan, A. Z. Portfolio optimization using higher moments in an uncertain random environment. *Inf. Sci.*, 567:348–374, 2021. doi: 10.1016/j.ins.2021.03.019. URL https://doi.org/10.1016/j.ins.2021.03.019.

Mockus, J. On Bayesian Methods for Seeking the Extremum. In Marchuk, G. I. (ed.), *Optimization Techniques, IFIP Technical Conference, Novosibirsk, USSR, July 1-7, 1974*, volume 27 of *Lecture Notes in Computer Science*, pp. 400–404. Springer, 1974. doi: 10.1007/3-540-07165-2_55. URL https://doi.org/10.1007/3-540-07165-2_55.

Novak, E. Some Results on the Complexity of Numerical Integration. In Cools, R. and Nuyens, D. (eds.), *Monte Carlo and Quasi-Monte Carlo Methods, MCQMC 2014, Leuven, Belgium, April 2014*, volume 163 of *Springer Proceedings in Mathematics and Statistics*, pp. 161–183. Springer, 2014. doi: 10.1007/978-3-319-33507-0_6. URL https://doi.org/10.1007/978-3-319-33507-0_6.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006. ISBN 026218253X. URL https://www.worldcat.org/oclc/61285753.

Schonlau, M., Welch, W. J., and Jones, D. R. Global versus local search in constrained optimization of computer models. *Lecture notes-monograph series*, pp. 11–25, 1998.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and de Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE*, 104(1):148–175, 2016. doi: 10.1109/JPROC.2015.2494218. URL https://doi.org/10.1109/JPROC.2015.2494218.

Stuckman, B. E. A global search method for optimizing nonlinear systems. *IEEE Trans. Syst. Man Cybern.*, 18(6): 965–977, 1988. doi: 10.1109/21.23094. URL https://doi.org/10.1109/21.23094.

Suzuki, S., Takeno, S., Tamura, T., Shitara, K., and Karasuyama, M. Multi-objective Bayesian Optimization using Pareto-frontier Entropy. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9279–9288. PMLR, 2020. URL http://proceedings.mlr.press/v119/suzuki20a.html.

Tanabe, R. and Ishibuchi, H. An easy-to-use real-world multi-objective optimization problem suite. *Applied Soft Computing*, 89:106078, 2020. ISSN 1568-4946.

doi: https://doi.org/10.1016/j.asoc.2020.106078. URL https://www.sciencedirect.com/science/article/pii/S1568494620300181.

Tu, B., Gandy, A., Kantas, N., and Shafei, B. Joint Entropy Search for Multi-Objective Bayesian Optimization. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Wang, H., Ren, Y., Deutz, A. H., and Emmerich, M. T. M. On Steering Dominated Points in Hypervolume Indicator Gradient Ascent for Bi-Objective Optimization. In Schütze, O., Trujillo, L., Legrand, P., and Maldonado, Y. (eds.), *NEO 2015 - Results of the Numerical and Evolutionary Optimization Workshop NEO 2015 held at September 23-25 2015 in Tijuana, Mexico*, volume 663 of *Studies in Computational Intelligence*, pp. 175–203. Springer, 2015. doi: 10.1007/978-3-319-44003-3_8. URL https://doi.org/10.1007/978-3-319-44003-3_8.

Wang, Z., Ong, Y., Sun, J., Gupta, A., and Zhang, Q. A Generator for Multiobjective Test Problems With Difficult-to-Approximate Pareto Front Boundaries. *IEEE Trans. Evol. Comput.*, 23(4):556–571, 2019. doi: 10.1109/TEVC.2018.2872453. URL https://doi.org/10.1109/TEVC.2018.2872453.

Yang, K., Deutz, A. H., Yang, Z., Bäck, T., and Emmerich, M. T. M. Truncated expected hypervolume improvement: Exact computation and application. In *IEEE Congress on Evolutionary Computation, CEC 2016, Vancouver, BC, Canada, July 24-29, 2016*, pp. 4350–4357. IEEE, 2016. doi: 10.1109/CEC.2016.7744343. URL https://doi.org/10.1109/CEC.2016.7744343.

Yang, K., Emmerich, M., Deutz, A., and Fonseca, C. M. Computing 3-D expected hypervolume improvement and related integrals in asymptotically optimal time. In Trautmann, H., Rudolph, G., Klamroth, K., Schütze, O., Wiecek, M., Jin, Y., and Grimme, C. (eds.), *International Conference on Evolutionary Multi-Criterion Optimization*, pp. 685–700, Cham, 2017. Springer.

Yang, K., Emmerich, M., Deutz, A., and Bäck, T. Efficient computation of expected hypervolume improvement using box decomposition algorithms. *Journal of Global Optimization*, 75(1):3–34, Sep 2019a. ISSN 1573-2916. doi: 10.1007/s10898-019-00798-7. URL https://doi.org/10.1007/s10898-019-00798-7.

Yang, K., Emmerich, M., Deutz, A. H., and Bäck, T. Multi-Objective Bayesian Global Optimization using expected hypervolume improvement gradient. *Swarm Evol. Comput.*, 44:945–956, 2019b. doi: 10.1016/j.swevo.2018.10.007. URL https://doi.org/10.1016/j.swevo.2018.10.007.

Yang, K., Affenzeller, M., and Dong, G. A parallel technique for multi-objective bayesian global optimization: Using a batch selection of probability of improvement. *Swarm and Evolutionary Computation*, 75:101183, 2022. ISSN 2210-6502. doi: https://doi.org/10.1016/j.swevo.2022.101183. URL https://www.sciencedirect.com/science/article/pii/S221065022200150X.

Zhang, Q. R. and Golovin, D. Random Hypervolume Scalarizations for Provable Multi-Objective Black Box Optimization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11096–11105. PMLR, 2020. URL http://proceedings.mlr.press/v119/zhang20i.html.

Zitzler, E. and Thiele, L. Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. *IEEE Transactions on Evolutionary Computation*, 3(4):257–271, 1999.

Zitzler, E., Deb, K., and Thiele, L. Comparison of multiobjective evolutionary algorithms: Empirical results. *Evol. Comput.*, 8(2):173–195, 2000. doi: 10.1162/106365600568202. URL https://doi.org/10.1162/106365600568202.

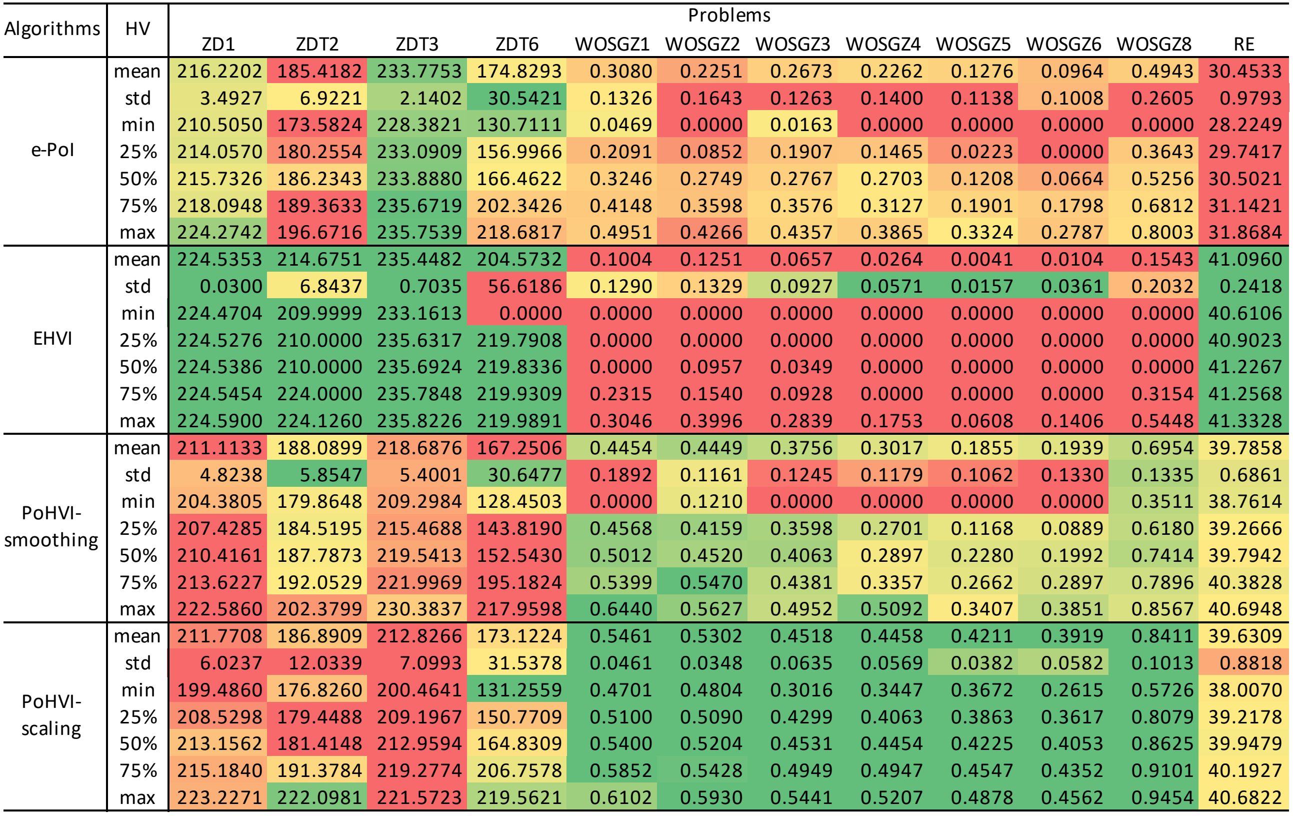
Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C. M., and da Fonseca, V. G. Performance assessment of multiobjective optimizers: an analysis and review. *IEEE Trans. Evol. Comput.*, 7(2):117–132, 2003. doi: 10.1109/TEVC.2003.810758. URL https://doi.org/10.1109/TEVC.2003.810758.

Zuluaga, M., Krause, A., and Püschel, M. e-PAL: An Active Learning Approach to the Multi-Objective Optimization Problem. *J. Mach. Learn. Res.*, 17:104:1–104:32, 2016. URL http://jmlr.org/papers/v17/15-047.html.

A. Appendix

We illustrate how to determine the hyperparameters of ε-PoHVI-scaling: $\varepsilon_t = \varepsilon_0 \exp\left(-ct\right)$. We use the following reasoning to determine the hyperparameters in the scaling scheme. For example, on WOSGZ problems: (1) ε_0 controls the maximal HV improvement across all iterations; (2) we take $[1.2, 1.2]$ as the reference point, which gives rise to an upper bound of 1.44 on the maximal HV value, provided that the ideal point on those problems is $[0, 0]$; (3) We assume that the initial HV value after random sampling is one-half the maximal HV value, i.e., $1.44/2$; (4) If the BO algorithm were to realize $\varepsilon_0 \exp\left(-ct\right)$ HV improvement in each iteration, then the total sum of all such improvements should be bounded above the maximal HV value to realize, i.e., $\varepsilon_0 \sum_t \exp\left(-ct\right) \leq 1.44/2$. The hyperparameter of ε-PoHVI-smoothing function can be determined in a similar way.

Figure 4. On each test problem, we show some descriptive statistics: min, max, mean, median, standard deviation (std), and 25%- and 75%-quantiles of the hypervolume (HV) value observed at the last iteration of the BO algorithm. The entries are color-coded relative to the corresponding ones (e.g., we take all the mean values in a column) in the same column/problem, where a more greenish color indicates better performance and vice versa.

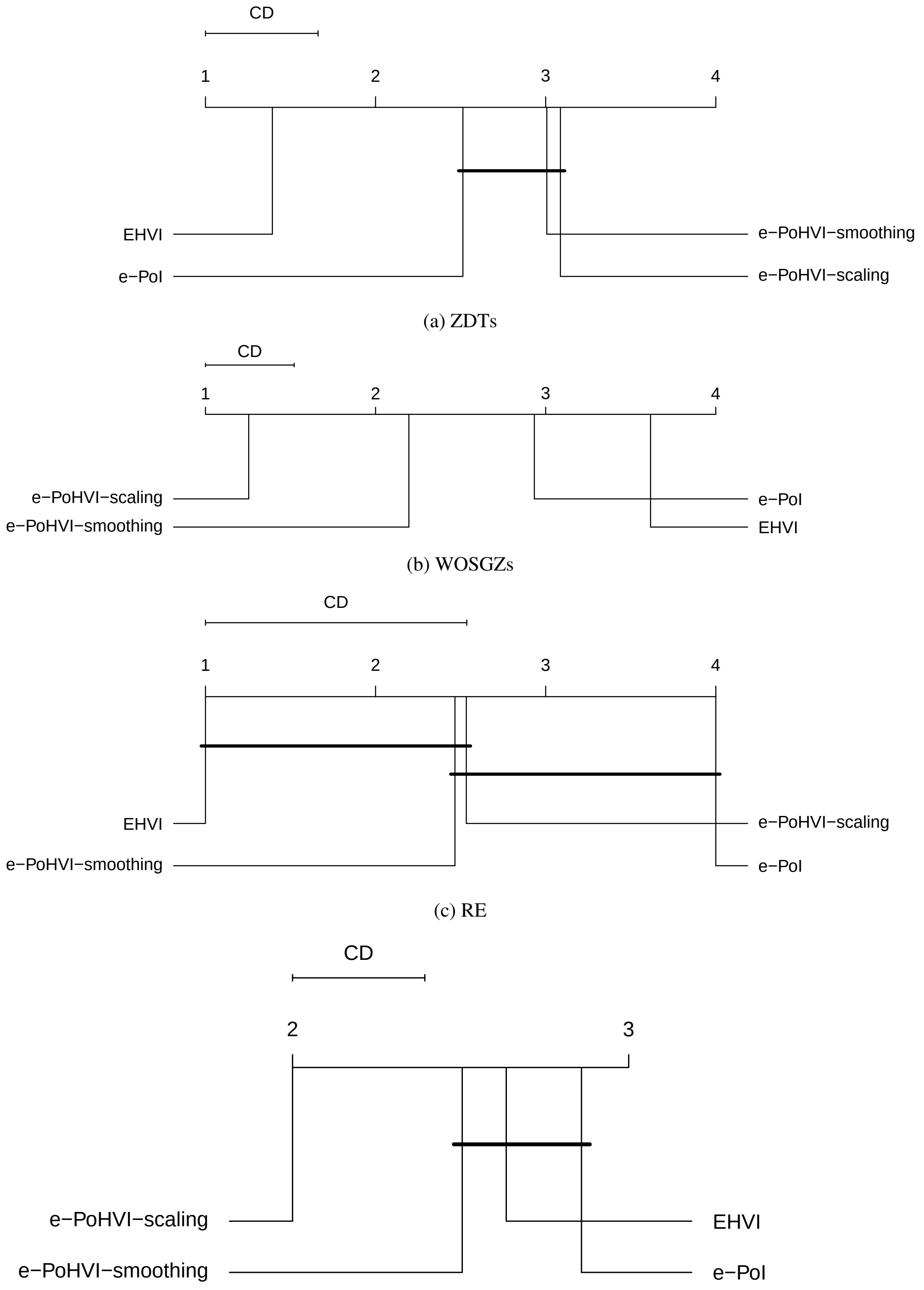


Algorithms	HV	Problems											
		ZD1	ZDT2	ZDT3	ZDT6	WOSGZ1	WOSGZ2	WOSGZ3	WOSGZ4	WOSGZ5	WOSGZ6	WOSGZ8	RE
e-PoI	mean	216.2202	185.4182	233.7753	174.8293	0.3080	0.2251	0.2673	0.2262	0.1276	0.0964	0.4943	30.4533
	std	3.4927	6.9221	2.1402	30.5421	0.1326	0.1643	0.1263	0.1400	0.1138	0.1008	0.2605	0.9793
	min	210.5050	173.5824	228.3821	130.7111	0.0469	0.0000	0.0163	0.0000	0.0000	0.0000	0.0000	28.2249
	25%	214.0570	180.2554	233.0909	156.9966	0.2091	0.0852	0.1907	0.1465	0.0223	0.0000	0.3643	29.7417
	50%	215.7326	186.2343	233.8880	166.4622	0.3246	0.2749	0.2767	0.2703	0.1208	0.0664	0.5256	30.5021
	75%	218.0948	189.3633	235.6719	202.3426	0.4148	0.3598	0.3576	0.3127	0.1901	0.1798	0.6812	31.1421
	max	224.2742	196.6716	235.7539	218.6817	0.4951	0.4266	0.4357	0.3865	0.3324	0.2787	0.8003	31.8684
EHVI	mean	224.5353	214.6751	235.4482	204.5732	0.1004	0.1251	0.0657	0.0264	0.0041	0.0104	0.1543	41.0960
	std	0.0300	6.8437	0.7035	56.6186	0.1290	0.1329	0.0927	0.0571	0.0157	0.0361	0.2032	0.2418
	min	224.4704	209.9999	233.1613	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	40.6106
	25%	224.5276	210.0000	235.6317	219.7908	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	40.9023
	50%	224.5386	210.0000	235.6924	219.8336	0.0000	0.0957	0.0349	0.0000	0.0000	0.0000	0.0000	41.2267
	75%	224.5454	224.0000	235.7848	219.9309	0.2315	0.1540	0.0928	0.0000	0.0000	0.0000	0.3154	41.2568
	max	224.5900	224.1260	235.8226	219.9891	0.3046	0.3996	0.2839	0.1753	0.0608	0.1406	0.5448	41.3328
PoHVI-smoothing	mean	211.1133	188.0899	218.6876	167.2506	0.4454	0.4449	0.3756	0.3017	0.1855	0.1939	0.6954	39.7858
	std	4.8238	5.8547	5.4001	30.6477	0.1892	0.1161	0.1245	0.1179	0.1062	0.1330	0.1335	0.6861
	min	204.3805	179.8648	209.2984	128.4503	0.0000	0.1210	0.0000	0.0000	0.0000	0.0000	0.3511	38.7614
	25%	207.4285	184.5195	215.4688	143.8190	0.4568	0.4159	0.3598	0.2701	0.1168	0.0889	0.6180	39.2666
	50%	210.4161	187.7873	219.5413	152.5430	0.5012	0.4520	0.4063	0.2897	0.2280	0.1992	0.7414	39.7942
	75%	213.6227	192.0529	221.9969	195.1824	0.5399	0.5470	0.4381	0.3357	0.2662	0.2897	0.7896	40.3828
	max	222.5860	202.3799	230.3837	217.9598	0.6440	0.5627	0.4952	0.5092	0.3407	0.3851	0.8567	40.6948
PoHVI-scaling	mean	211.7708	186.8909	212.8266	173.1224	0.5461	0.5302	0.4518	0.4458	0.4211	0.3919	0.8411	39.6309
	std	6.0237	12.0339	7.0993	31.5378	0.0461	0.0348	0.0635	0.0569	0.0382	0.0582	0.1013	0.8818
	min	199.4860	176.8260	200.4641	131.2559	0.4701	0.4804	0.3016	0.3447	0.3672	0.2615	0.5726	38.0070
	25%	208.5298	179.4488	209.1967	150.7709	0.5100	0.5090	0.4299	0.4063	0.3863	0.3617	0.8079	39.2178
	50%	213.1562	181.4148	212.9594	164.8309	0.5400	0.5204	0.4531	0.4454	0.4225	0.4053	0.8625	39.9479
	75%	215.1840	191.3784	219.2774	206.7578	0.5852	0.5428	0.4949	0.4947	0.4547	0.4352	0.9101	40.1927
	max	223.2271	222.0981	221.5723	219.5621	0.6102	0.5930	0.5441	0.5207	0.4878	0.4562	0.9454	40.6822

Table 2. GP's prediction uncertainty (posterior standard deviations σ_1 and σ_2 for each objective function, respectively) aggregated over uniformly sampled decision points on ZDTs (top) and WOSGZs (bottom) and BO's iterations. The average and standard error of the uncertainty is estimated over 15 repetitions of BO.

	ZDT1		ZDT2		ZDT3		ZDT4		ZDT6	
	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2
average	0.008378	0.044345	0.008819	0.008107	0.003032	0.400788	0.023912	22.251792	0.043914	0.173658
standard error	0.043334	0.015357	0.041234	0.003639	0.004248	0.028704	0.019340	0.8503408	0.030066	0.199384

	WOSGZ1		WOSGZ2		WOSGZ3		WOSGZ4		WOSGZ5		WOSGZ6		WOSGZ7		WOSGZ8	
	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2
average	1.5524	0.0915	3.1203	0.1797	3.1615	0.1834	3.9904	0.2614	4.4331	0.3117	5.4329	0.3487	3.5918	0.2047	3.8567	0.2651
standard error	0.5048	0.0352	0.9660	0.0654	1.0346	0.0690	1.3334	0.1030	1.4432	0.1061	1.7562	0.1255	0.8340	0.0628	1.0050	0.0861



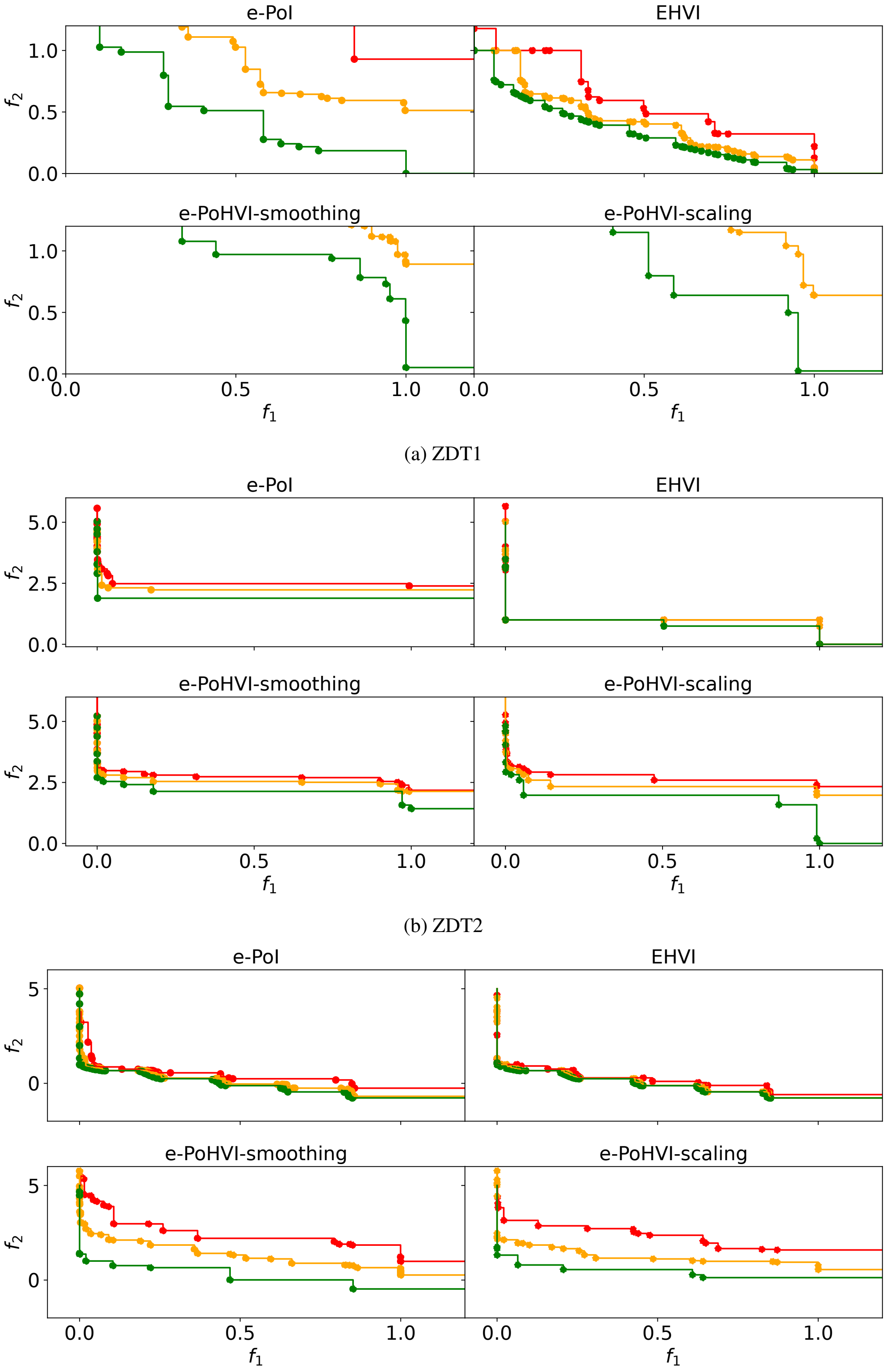
(a) ZDTs

(b) WOSGZs

(c) RE

(d) All problems

Figure 5. The critical difference (CD) chart obtained with the Nemenyi posthoc testing procedure to a Friedman test. The performance of two acquisition functions significantly differs on a problem set if their average ranks of HV values differ by at least the critical difference shown as the interval atop each chart. The thick horizontal line indicates a clique of acquisition functions with no significant difference.

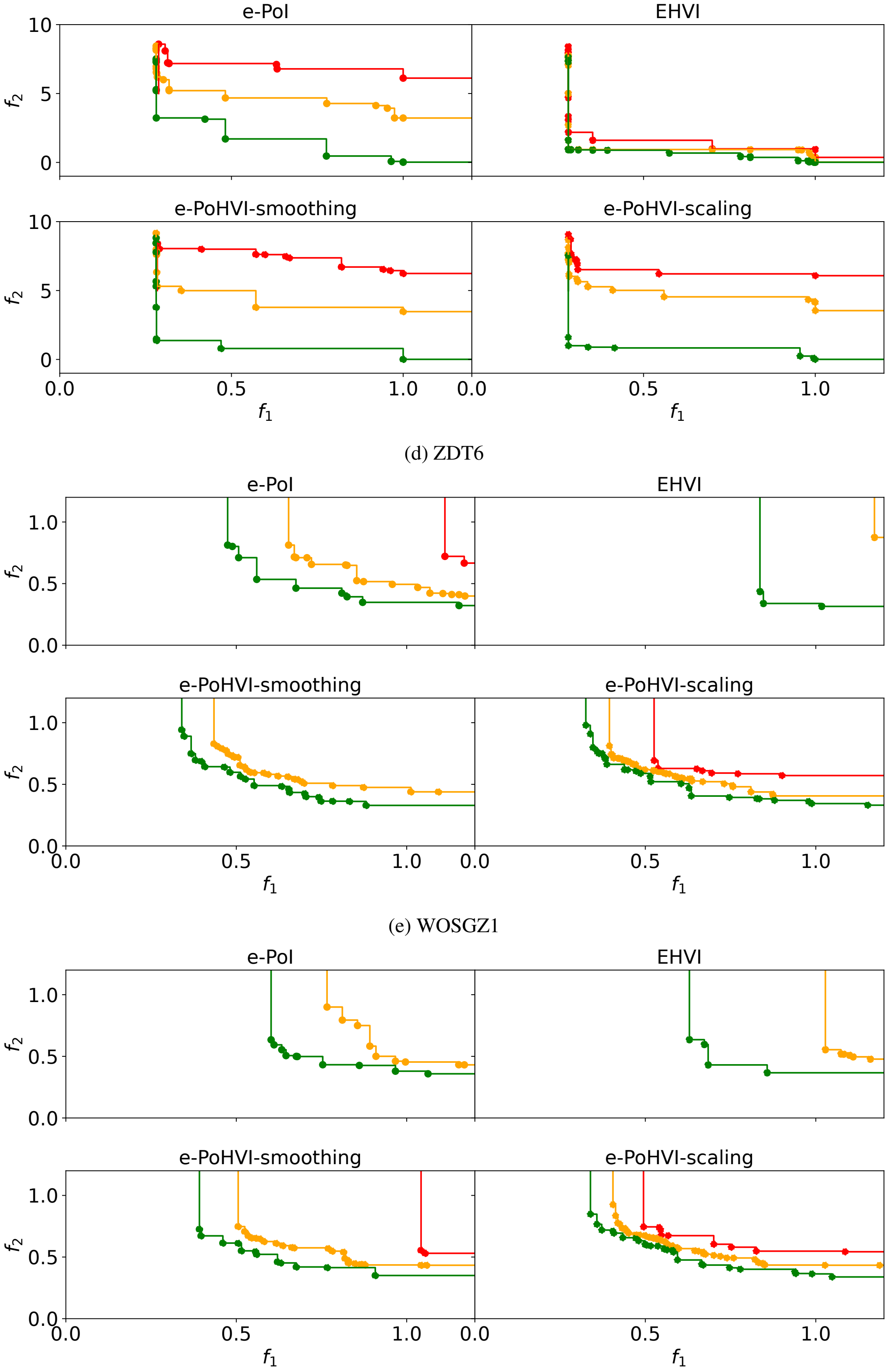


(a) ZDT1

(b) ZDT2

(c) ZDT3

Figure 6. The best, median, and the worst empirical attainment curves on all the test problems, where ●, ●, and ● represent the best, the worst, and the median Pareto-front approximation sets over 15 independent runs, respectively.

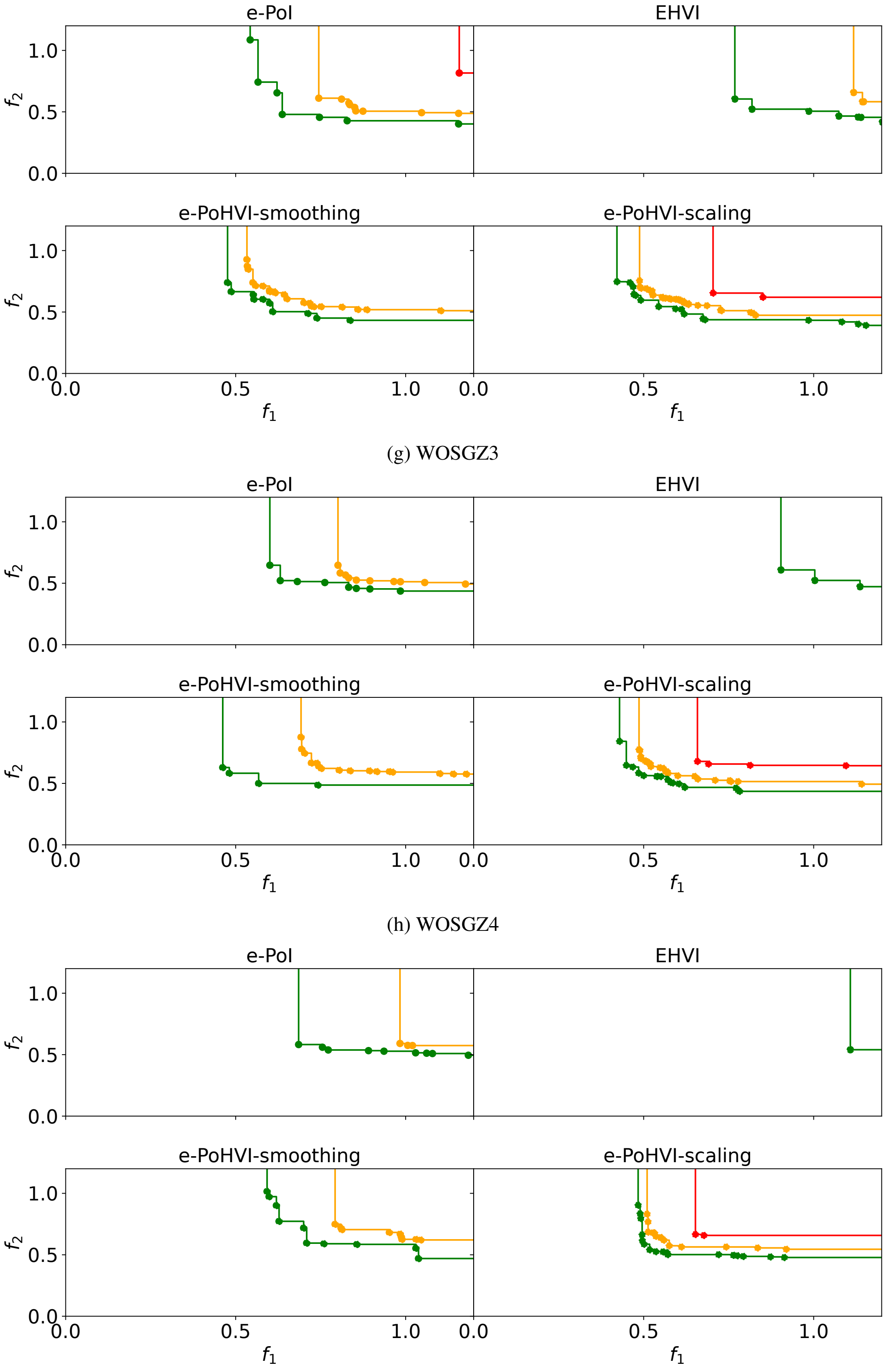


(d) ZDT6

(e) WOSGZ1

(f) WOSGZ2

Figure 6. The best, median, and the worst empirical attainment curves on all the test problems, where 🟢, 🟠, and 🔴 represent the best, the worst, and the median Pareto-front approximation sets over 15 independent runs, respectively.

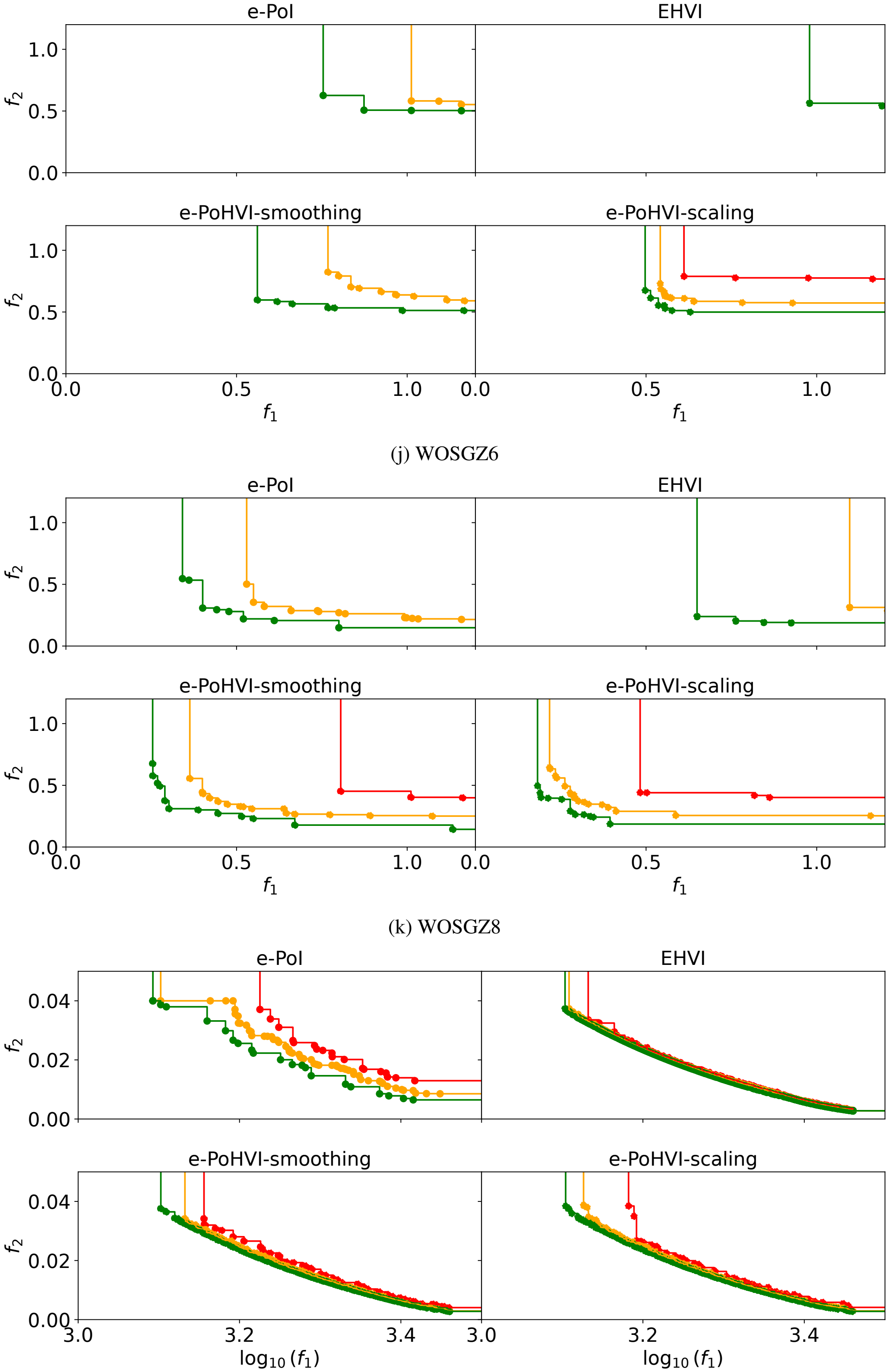


(g) WOSGZ3

(h) WOSGZ4

(i) WOSGZ5

Figure 6. The best, median, and the worst empirical attainment curves on all the test problems, where ●, ●, and ● represent the best, the worst, and the median Pareto-front approximation sets over 15 independent runs, respectively.



(j) WOSGZ6

(k) WOSGZ8

(l) RE

Figure 6. The best, median, and the worst empirical attainment curves on all the test problems, where ●, ●, and ● represent the best, the worst, and the median Pareto-front approximation sets over 15 independent runs, respectively.