A Theoretical Analysis

In this section, we provide a proof for **Theorem 1** described in Section 4.1.

Lemma 1. *(Lemma C.1 in MES [26]). Pick $\delta \in (0,1)$ and set $\zeta_t = \left(2\log\left(\frac{\pi_t}{2\delta}\right)\right)^{1/2}$, where $\sum_{t=1}^{T}(\pi_t)^{-1} \leq 1, \pi_t > 0$. Then, it holds that for each function f_j, $Pr[\mu_{j,t-1}(\mathbf{x_t}) - f_j(\mathbf{x_t}) \leq \zeta_t\sigma_{j,t-1}(\mathbf{x}), \forall t \in [1,T]] \geq 1-\delta$. Here $\mu_{j,t-1}$ and $\sigma_{j,t-1}(\mathbf{x})$ refers to the predictive mean and variance of j^{th} GP at iteration number t.*

Lemma 2. *(Lemma C.2 in MES [26]) If $\mu_{j,t-1}(\mathbf{x_t}) - f_j(\mathbf{x_t}) \leq \zeta_t\sigma_{j,t-1}(\mathbf{x})$, for each $j \in [1,\cdots,K]$, the quantity $r_t^j = f_j(x^*) - f_j(x_t) \leq \left(v_t^j + \zeta_t\right)\sigma_{j,t-1}(\mathbf{x_t})$, where $v_t^j \doteq \min_{\mathbf{x}\in\mathfrak{X}} \frac{y^{j*}-\mu_{j,t-1}(\mathbf{x})}{\sigma_{j,t-1}(\mathbf{x})}$ and $y^{j*} \geq f_j(x^*) \forall t \in [1,T]$.*

Theorem 1. *Let P be a distribution over vector $[y^{1*},\cdots,y^{K*}]$ where each y^{j*} is the maximum value for function f_j among the vectors in the Pareto front obtained by solving the cheap multi-objective optimization problem over sampled functions from the K Gaussian process models. Let the observation noise for function evaluations is i.i.d $\mathcal{N}(0,\sigma)$ and $w = Pr[\left(y^{1*} > f_1(x^*)\right),\cdots,\left(y^{K*} > f_K(x^*)\right)]$. If $\mathbf{x}_t$ is the candidate input selected by MESMO at the t^{th} iteration according to 4.12 and $[y^{1*},\cdots,y^{K*}]$ is drawn from P, then with probability atleast $1-\delta$, in $T' = \sum_{i=1}^{T}\log_w \frac{\delta}{2\pi_i}$ number of iterations*

$$R(\mathbf{x}^*) = \sqrt{\sum_{j=1}^{K}\left(\left(v_{t^*}^j + \zeta_T\right)^2\left(\frac{2T\gamma_T^j}{log(1+\sigma^{-2})}\right)\right)} \tag{A.1}$$

where $\zeta_T = (2\log(\pi_T/\delta))^{1/2}$, $\pi_i > 0$, and $\sum_{i=1}^{T}\frac{1}{\pi_i} \leq 1$, $v_{t^*}^j = \max_t v_t^j$ with $v_t^j = \min_{x\in\mathfrak{X}} \frac{y^{j*}-\mu_{j,t-1}(\mathbf{x})}{\sigma_{j,t-1}(\mathbf{x})}$, and γ_T^j is the maximum information gain about function f_j after T function evaluations.

Proof. The result for each R^j can be derived from the fact that the corresponding expression for a single sample in Equation 4.12 $\left(\frac{\gamma_s^j(\mathbf{x})\phi(\gamma_s^j(\mathbf{x}))}{2\Phi(\gamma_s^j(\mathbf{x}))} - \ln\Phi(\gamma_s^j(\mathbf{x}))\right)$ is equivalent to EST (optimization as estimation strategy) [27]. This fact is proven as Lemma **3.1** in MES (Max-value entropy search) [26]. Therefore, theoretical results from MES can be leveraged for each R^j provided $y^{j*} > f_j(x^*)$ for all $j \in \{1,\cdots,K\}$.

Since $[y^{1*},\cdots,y^{K*}]$ is drawn from P, the probability that there exists atleast one vector $[y^{1*},\cdots,y^{K*}]$ in k_i iterations that satisfies $[\left(y^{1*} > f_1(x^*)\right),\cdots,\left(y^{K*} > f_K(x^*)\right)]$ is given by:

$$\Rightarrow w + (1-w)w + (1-w)^2w\cdots + (1-w)^{k_i-1}w \tag{A.2}$$

$$= w\cdot\left(\frac{1-(1-w)^{k_i}}{1-(1-w)}\right) \tag{A.3}$$

$$= 1-(1-w)^{k_i} \tag{A.4}$$

$$\geq (1-(1-w))^{k_i} \quad \text{since } w \in (0,1) \tag{A.5}$$

$$\geq w^{k_i} \tag{A.6}$$

Suppose $T' = \sum_{i=1}^{T} k_i$ be the total number of iterations (function evaluations). Following Theorem 3.2 from MES [26], splitting the total number of iterations into T parts, where each part has k_i iterations, there exists at least one iteration t_i in each of the T parts with probability $1-\sum_{i=1}^{T} w^{k_i}$ such that $[\left(y^{1*} > f_1(x^*)\right),\cdots,\left(y^{K*} > f_K(x^*)\right)]$.

Let $\sum_{i=1}^{T} w^{k_i} = \frac{\delta}{2}$ and setting $k_i = \log_w \frac{\delta}{2\pi_i}$ for any $\sum_{i=1}^{T}\frac{1}{\pi_i} = 1$. A standard choice for π_i is $\pi_i = \pi^2 i^2/6$. Using this transformation of variables, the probability that there exists sampled functions such that $[\left(y^{1*} > f_1(x^*)\right),\cdots,\left(y^{K*} > f_K(x^*)\right)]$ is atleast $1-\delta/2, \forall i \in [1,T]$.

By lemma 1 and 2,

$$r_{t_i}^j = \left(v_{t_i}^j + \zeta_{t_i}\right)\sigma_{j,t_i-1}(\mathbf{x_{t_i}}) \tag{A.7}$$

From Lemma C.3 in MES[26], $\sum_{i=1}^{T} \sigma^2_{j,t_i-1}(\mathbf{x}_{t_i}) \leq \frac{2}{log(1+\sigma^{-2})}\gamma^j_T$, where γ^j_T is the maximum information gain about function f_j and is an important theoretical quantity related to regret bounds in bayesian optimization literature[25]. By Cauchy-Schwarz inequality, $\sum_{i=1}^{T} \sigma_{j,t_i-1}(\mathbf{x}_{t_i}) \leq \sqrt{T\sum_{i=1}^{T} \sigma^2_{j,t_i-1}(\mathbf{x}_{t_i})} \leq \sqrt{2T\gamma^j_T/\log(1+\sigma^{-2})}$. Therefore, with probability $1-\delta$,

$$R^j(\mathbf{x}^*) = \sum_{i=1}^{T} r^j_{t_i} \leq \left(v^j_{t^*} + \zeta_T\right)\sqrt{\frac{2T\gamma^j_T}{log(1+\sigma^{-2})}} \tag{A.8}$$

Consequently,

$$R(x^*) = \sqrt{\sum_{j=1}^{K}\left(\left(v^j_{t^*} + \zeta_T\right)^2\left(\frac{2T\gamma^j_T}{log(1+\sigma^{-2})}\right)\right)} \tag{A.9}$$

□

B Detailed derivation of acquisition function

The complete derivation of Equation 4.12 from Equation 4.10 is given below.

$$H(\mathbf{y} \mid D, \mathbf{x}, \mathcal{Y}^*_s) \simeq \sum_{j=1}^{K} H(y^j|D, \mathbf{x}, \max\{z^j_1, \cdots z^j_m\}) \tag{B.1}$$

The r.h.s is a summation over entropies of K variables $\{y^1, \cdots, y^K\}$. Let $y^{j*}_s = \max\{z^j_1, \cdots z^j_m\}$ and $\gamma^j_s(x) = \frac{y^{j*}_s - \mu_j(\mathbf{x})}{\sigma_j(\mathbf{x})}$. The probability distribution of each variable y^j is a truncated Gaussian with upper bound y^{j*}_s. The differential entropy for each y^j is given as:

$$H(y^j \mid D, \mathbf{x}, \mathcal{Y}^*_s) \simeq \left[\frac{(1+\ln(2\pi))}{2} + \ln(\sigma_j(\mathbf{x})) + \ln\Phi(\gamma^j_s(\mathbf{x})) - \frac{\gamma^j_s(\mathbf{x})\phi(\gamma^j_s(\mathbf{x}))}{2\Phi(\gamma^j_s(\mathbf{x}))}\right] \tag{B.2}$$

Summing over all K variables gives:

$$H(\mathbf{y} \mid D, \mathbf{x}, \mathcal{Y}^*_s) \simeq \sum_{j=1}^{K}\left[\frac{(1+\ln(2\pi))}{2} + \ln(\sigma_j(\mathbf{x})) + \ln\Phi(\gamma^j_s(\mathbf{x})) - \frac{\gamma^j_s(\mathbf{x})\phi(\gamma^j_s(\mathbf{x}))}{2\Phi(\gamma^j_s(\mathbf{x}))}\right] \tag{B.3}$$

Equation 4.8 is given below:

$$\mathbb{E}_{\mathcal{Y}^*}[H(\mathbf{y} \mid D, \mathbf{x}, \mathcal{Y}^*)] \simeq \frac{1}{S}\sum_{s=1}^{S}[H(\mathbf{y} \mid D, \mathbf{x}, \mathcal{Y}^*_s)] \tag{B.4}$$

Using B.3 in B.4:

$$\mathbb{E}_{\mathcal{Y}^*}[H(\mathbf{y} \mid D, \mathbf{x}, \mathcal{Y}^*)] \simeq \frac{1}{S}\sum_{s=1}^{S}\sum_{j=1}^{K}\left[\frac{(1+\ln(2\pi))}{2} + \ln(\sigma_j(\mathbf{x})) + \ln\Phi(\gamma^j_s(\mathbf{x})) - \frac{\gamma^j_s(\mathbf{x})\phi(\gamma^j_s(\mathbf{x}))}{2\Phi(\gamma^j_s(\mathbf{x}))}\right] \tag{B.5}$$

Equation 4.7 is given below:

$$H(\mathbf{y} \mid D, \mathbf{x}) = \frac{K(1+\ln(2\pi))}{2} + \sum_{i=1}^{K}\ln(\sigma_i(\mathbf{x})) \tag{B.6}$$

Recall that MESMO acquisition function is given by Equation 4.6 which is composed of B.5 and B.6:

$$\alpha(\mathbf{x}) = H(\mathbf{y} \mid D, \mathbf{x}) - \mathbb{E}_{\mathcal{Y}^*}[H(\mathbf{y} \mid D, \mathbf{x}, \mathcal{Y}^*)]$$

$$\alpha(\mathbf{x}) \simeq \left[\frac{K(1+\ln(2\pi))}{2} + \sum_{i=1}^{K} \ln(\sigma_i(\mathbf{x}))\right] - \left[\frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{K}\left[\frac{(1+\ln(2\pi))}{2} + \ln(\sigma_i(\mathbf{x})) + \ln\Phi(\gamma_s^i(\mathbf{x})) - \frac{\gamma_s^i(\mathbf{x})\phi(\gamma_s^i(\mathbf{x}))}{2\Phi(\gamma_s^i(\mathbf{x}))}\right]\right]$$

$$\alpha(\mathbf{x}) \simeq \frac{1}{S}\sum_{s=1}^{S}\left[\frac{K(1+\ln(2\pi))}{2} + \sum_{i=1}^{K} \ln(\sigma_i(\mathbf{x}))\right] - \left[\frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{K}\left[\frac{(1+\ln(2\pi))}{2} + \ln(\sigma_i(\mathbf{x})) + \ln\Phi(\gamma_s^i(\mathbf{x})) - \frac{\gamma_s^i(\mathbf{x})\phi(\gamma_s^i(\mathbf{x}))}{2\Phi(\gamma_s^i(\mathbf{x}))}\right]\right]$$

Re-arranging similar terms together:

$$\alpha(\mathbf{x}) \simeq \frac{1}{S}\left[\sum_{s=1}^{S}\frac{K(1+\ln(2\pi))}{2} - \frac{K(1+\ln(2\pi))}{2} + \sum_{i=1}^{K}\ln(\sigma_i(\mathbf{x})) - \sum_{i=1}^{K}\ln(\sigma_i(\mathbf{x}))\right] + \left[\sum_{s=1}^{S}\sum_{i=1}^{K}\left[\frac{\gamma_s^i(\mathbf{x})\phi(\gamma_s^i(\mathbf{x}))}{2\Phi(\gamma_s^i(\mathbf{x}))} - \ln\Phi(\gamma_s^i(\mathbf{x}))\right]\right]$$

$$\alpha(\mathbf{x}) \simeq \left[\sum_{s=1}^{S}\sum_{i=1}^{K}\left[\frac{\gamma_s^i(\mathbf{x})\phi(\gamma_s^i(\mathbf{x}))}{2\Phi(\gamma_s^i(\mathbf{x}))} - \ln\Phi(\gamma_s^i(\mathbf{x}))\right]\right]$$

C Additional Experimental Results

Figure 3 shows the results for MESMO and baseline algorithms on two benchmarks from the general multi-objective optimization literature. Figure 4 shows the results comparing the acquisition function optimization time of MESMO and baseline algorithms. We fix the input space dimensions to $d = 5$ and vary the number of objective functions to show how different algorithms scale with increasing number of objectives.

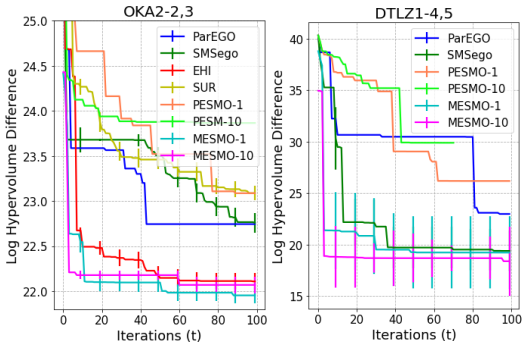

Figure 3: Results of different multi-objective BO algorithms including MESMO on synthetic benchmarks from general MO literature. The log of the hypervolume difference is shown with different number of function evaluations. The mean and variance of 10 different runs are plotted. The title of each figure refers to the name of benchmark. (Figures better seen in color.)

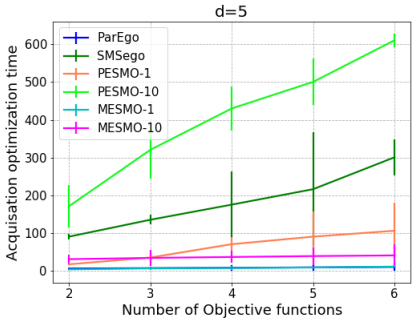

Figure 4: Results for acquisition function optimization time of different multi-objective BO algorithms including MESMO with increasing number of objective functions for fixed input space dimension $d = 5$.