A. Theoretical Analysis and Asymptotic Bounds

This section provides the proofs of Theorem 1, which follows from Lemmas 3 to 12.

Lemma 3. *Given $\delta \in (0,1)$ and $\beta_t = 2\log(n|E|\pi_t/\delta)$, the following holds with probability $\geq 1-\delta$:*

$$f_i(\boldsymbol{x}) - \mu_{t-1,i}(x) \leq \beta_t^{1/2}\sigma_{t-1,i}(\boldsymbol{x}) \quad \text{for all } 1 \leq i \leq n, \boldsymbol{x} \in E, \text{ for all } t \geq 1. \tag{12}$$

In other words, with probability $\geq 1-\delta$:

$$\boldsymbol{f}(\boldsymbol{x}) \in R_t(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in E, \text{ for all } t \geq 1$$

Proof. According to Lemma 5.1 in (Srinivas et al., 2012), the following inequality holds:

$$Pr\left\{f_i(\boldsymbol{x}) - \mu_{t-1,i}(\boldsymbol{x}) > \beta_t^{1/2}\sigma_{t-1,i}(\boldsymbol{x})\right\} \leq e^{-\beta_t/2}$$

Applying the union bound for $i, t \in \mathbb{N}$, we obtain that the following holds with probability $\geq 1 - n|E|e^{-\beta_t/2}$:

$$f_i(\boldsymbol{x}) - \mu_{t-1,i}(\boldsymbol{x}) \leq \beta_t^{1/2}\sigma_{t-1,i}(\boldsymbol{x}) \quad \text{for all } 1 \leq i \leq n, \text{ for all } \boldsymbol{x} \in E. \tag{13}$$

The lemma holds by choosing $n|E|e^{-\beta_t/2} = \delta/\pi_t$. As suggested in (Srinivas et al., 2010), we can use $\pi_t = \pi^2 t^2/6$. □

Lemma 4. *If $n = 1$ and $\boldsymbol{f}_T = (\boldsymbol{f}(\boldsymbol{x}_t))_{1\leq t\leq T}$, then*

$$I(\boldsymbol{y}_T; \boldsymbol{f}_T) = \frac{1}{2}\sum_{t=1}^{T}\log(1+\sigma^{-2}\sigma_{t-1}^2(\boldsymbol{x}_t)).$$

This is directly taken from Lemma 5.3 in (Srinivas et al., 2010). $I(\boldsymbol{y}_T; \boldsymbol{f}_T)$ defines the mutual information between $\boldsymbol{f}$ and observations $\boldsymbol{y}_T = \boldsymbol{f}_T + \epsilon_T$, where $\epsilon_T \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$.

Lemma 5. *Given $\delta \in (0,1)$ and $\beta_t = 2\log(n|E|\pi_t/\delta)$, the following holds with probability at least $1-\delta$:*

$$\sum_{t=1}^{T}\overline{w}_t^2 \leq \beta_T C_1 I(\boldsymbol{y}_T; \boldsymbol{f}_T) \leq C_1\beta_T\gamma_T \text{ for all } T \geq 1,$$

where $C_1 = 8/\log(1+\sigma^{-2})$.

Proof. One of the rectangles of which $R_t(\boldsymbol{x}_t)$ is the intersection has a diagonal length of $2\beta_t^{1/2}\|\boldsymbol{\sigma}_{t-1}(\boldsymbol{x}_t)\|_2$: as a consequence,

$$\overline{w}_t^2 \leq 4\beta_t\|\boldsymbol{\sigma}_{t-1}(\boldsymbol{x}_t)\|_2^2.$$

As β_t is increasing, we have that

$$\begin{aligned}\overline{w}_t^2 &\leq 4\beta_T\sigma^2\sum_{i=1}^{n}\sigma^{-2}\sigma_{t-1,i}^2(\boldsymbol{x}_t)\\ &\leq 4\beta_T\sigma^2 C_2\sum_{i=1}^{n}\log(1+\sigma^{-2}\sigma_{t-1,i}^2(\boldsymbol{x}_t))\end{aligned}$$

with $C_2 = \sigma^{-2}/\log(1+\sigma^{-2}) \geq 1$, since $s^2 \leq C_2\log(1+s^2)$ for $0 \leq s \leq \sigma^{-2}$, and $\sigma^{-2}\sigma_{t-1,i}^2(\boldsymbol{x}_t) \leq \sigma^{-2}k_i(\boldsymbol{x}_t, \boldsymbol{x}_t) \leq \sigma^{-2}$.
Using $C_1 = 8\sigma^2 C_2$ and Lemma 4 we have that

$$\begin{aligned}\sum_{t=1}^{T}\overline{w}_t^2 &\leq \beta_T C_1\sum_{i=1}^{n}I(\boldsymbol{y}_T; f_{T,i})\\ &\leq \beta_T C_1 I(\boldsymbol{y}_T; \boldsymbol{f}_T)\end{aligned}$$

□

Lemma 6. *Given $\delta \in (0,1)$ and $\beta_t = 2\log(n|E|\pi_t/\delta)$, the following holds with probability $\geq 1-\delta$:*

$$\sum_{t=1}^{T}\overline{w}_t \leq \sqrt{C_1 T\beta_T\gamma_T} \text{ for all } T \geq 1$$

Proof. This follows from Lemma 5, since $(\sum_t^T \overline{w}_t)^2 \leq T\sum_{t=1}^T\overline{w}_t^2$ by the Cauchy-Schwarz inequality. □

Lemma 7. *Running* PAL *with a monotonic classification, it holds that $\overline{w}_t$ decreases with t.*

Proof. As a direct consequence of the sample picking rule, $w_{t-1}(\boldsymbol{x}_t) \leq \overline{w}_{t-1}$. On the other hand, $w_t(\boldsymbol{x}) \leq w_{t-1}(\boldsymbol{x})$ and thus, $\overline{w}_t \leq w_{t-1}(\boldsymbol{x}_t)$. The lemma follows. □

Lemma 8. *Running* PAL *with $\delta \in (0,1)$ and $\beta_t = 2\log(n|E|\pi_t/\delta)$, the following holds:*

$$Pr\left\{\overline{w}_T \leq \sqrt{\frac{C_1\beta_T\gamma_T}{T}} \text{ for all } T \geq 1\right\} \geq 1-\delta, \tag{14}$$

where $C_1 = 8/\log(1-\sigma^{-2})$ and $\pi_t = \pi^2t^2/6$.

Proof. This is derived from Lemmas 6 and 7, since $\sum_{t=1}^T\overline{w}_t/T \geq \overline{w}_T$.

□

Corollary 9. *When running* PAL *with squared exponential kernels k_i for all $1 \leq i \leq n$, the following holds with probability $\geq 1-\delta$:*

$$\begin{aligned}\overline{w}_T &= O\left(\sqrt{\frac{n\log^{d+1}T(\log T + \log n - \log\delta)}{T}}\right)\\ &= O^*(n^{\frac{1}{2}}T^{-\frac{1}{2}})\end{aligned} \tag{15}$$

Lemma 10. *If when running* PAL *at iteration t', a point $\boldsymbol{x}$ is classified as not Pareto-optimal, i.e. $\boldsymbol{x} \in N_{t'}$, it can be removed from E as it is not needed for the classification of points in $U_{t'}$. This means that no point $\boldsymbol{x}' \in U_t$ for $t > t'$ has to be compared with $\boldsymbol{x}$ to attempt its classification at time $t > t'$.*

Proof. To attempt the classification of a point $\boldsymbol{x}' \in U_t$, PAL searches for other points in E that may dominate $\boldsymbol{x}'$ under different outcomes, considering their corresponding uncertainty regions R_t. If a point $\boldsymbol{x}$ is classified as not Pareto-optimal, there exist at least a point $\boldsymbol{x}'' \in (P_t \cup U_t)$ such that $\boldsymbol{x}'' \succeq \boldsymbol{x}$. Then $\boldsymbol{x}$ can be ignored since if $\boldsymbol{x} \succeq \boldsymbol{x}'$ then $\boldsymbol{x}'' \succeq \boldsymbol{x}'$. □

Lemma 11. *If when running* PAL*, at iteration t $\overline{w}_t \leq 2\epsilon$, then $U_{t+1} = \emptyset$.*

Proof. We show that if a point is not classified as Pareto-optimal, then it is classified as not Pareto-optimal, when $\overline{w}_t \leq 2\epsilon$.
If a point $\boldsymbol{x}$ is not classified as Pareto-optimal, then there is a point $\boldsymbol{x}'$ such that

$$\min(R_t(\boldsymbol{x})) + \boldsymbol{\epsilon} \preceq \max(R_t(\boldsymbol{x}')) - \boldsymbol{\epsilon}, \tag{16}$$

with $\boldsymbol{\epsilon} = (\epsilon, \ldots, \epsilon)$. We define a point $\boldsymbol{v}(\boldsymbol{x}) \in \mathbb{R}^n = \max(R_t(\boldsymbol{x})) - \min(R_t(\boldsymbol{x})) = (v_1(\boldsymbol{x}), \ldots, v_n(\boldsymbol{x}))$. Thus, 16 is equivalent to

$$\min(R_t(\boldsymbol{x})) + \boldsymbol{\epsilon} \preceq \min(R_t(\boldsymbol{x}')) + \boldsymbol{v}(\boldsymbol{x}') - \boldsymbol{\epsilon}.$$

If there is a point $\boldsymbol{x}'$ that meets the relation in (16) and $\overline{w}_t \leq 2\epsilon$, then $\boldsymbol{x}'$ meets the following condition that classifies $\boldsymbol{x}$ as not Pareto-optimal

$$\min(R_t(\boldsymbol{x})) + \boldsymbol{v}(\boldsymbol{x}) - \boldsymbol{\epsilon} \preceq \min(R_t(\boldsymbol{x}')) + \boldsymbol{\epsilon},$$

since $v_i(\boldsymbol{x}) \leq \overline{w}_t$, for all $1 \leq i \leq n$ and for all $\boldsymbol{x} \in P_t \cup U_t$, and $\overline{w}_t - \epsilon \leq \epsilon$. $\overline{w}_t$ is an upper bound of $||\max(R_t(\boldsymbol{x})) - \min(R_t(\boldsymbol{x}))||_2$ for all points that are in P_t and U_t. As shown in Lemma 10, points in N_t do not have to be considered in the classification. □

Lemma 12. *Let $\delta \in (0,1)$, $\beta_t = 2\log(n|E|\pi^2 t^2/(6\delta))$, $a_i = \max_{\boldsymbol{x}\in E}\{\sqrt{\beta_1 k_i(\boldsymbol{x},\boldsymbol{x})}\}$, and $a = \max_{1\leq i\leq n}\{a_i\}$. The following holds with probability $1-\delta$.*
The hypervolume error obtained by PAL *at iteration T when all points have been classified is bounded as:*

$$\eta_T \leq \frac{n a^{n-1}}{(n-1)!}(\overline{w}_T), \tag{17}$$

In particular,

$$\eta_T = O^*\left(\frac{n^{3/2} a^{n-1}}{T^{1/2}(n-1)!}\right).$$

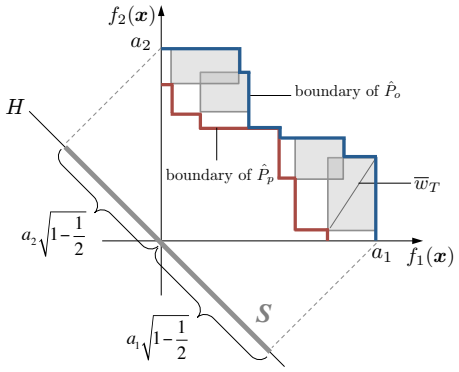

Figure 6. Example of hypervolume error bound for $n = 2$.

Proof. Let $\mathbf{1}_n = (1, \ldots, 1)^T$ and let $\boldsymbol{e}_i$ denote the ith canonical base vector, all assumed $\in \mathbb{R}^n$. The length of every (one-dimensional) side of a hyperrectangle associated with a point in $\hat{P}$ is bounded by the length of its diagonal $\overline{w}_T$. Hence the distance between the boundaries defined by $\hat{P}_o$ and $\hat{P}_p$ along the direction $\mathbf{1}_n$ is bounded by $\sqrt{n}\overline{w}_T$ (the diagonal of a hypercube with side length $\overline{w}_t$).
a_i is the maximum value that $f_i(\boldsymbol{x})$ can have, with probability $1-\delta$. $a_i = \max_{\boldsymbol{x}\in E}\{\sqrt{\beta_1 k_i(\boldsymbol{x},\boldsymbol{x})}\}$ is a bound obtained from the width of the confidence regions, given the Gaussian process prior distribution. Let $\boldsymbol{a}_i = a_i \boldsymbol{e}_i$. The projection $\boldsymbol{a}_i$, $1 \leq i \leq n$, onto the hyperplane $H_n = \langle \mathbf{1}_n \rangle^\perp$ is an n-simplex S_n. $\mathrm{V}(\hat{P}_o) - \mathrm{V}(\hat{P}_p)$, and hence η_T, are bounded by the volume of S_n times $\sqrt{n}\overline{w}_t$.
We compute an upper bound on the volume of S_n. The projection of $\boldsymbol{a}_i$ onto H_n is $\tilde{\boldsymbol{a}}_i = a_i(\boldsymbol{e}_i - 1/n\mathbf{1})$, which has length $a_i\sqrt{1-1/n}$. The $\tilde{\boldsymbol{a}}_i$ enclose pairwise the same angle; hence the volume of S_n is bounded by the volume of a regular n-simplex with radius $a\sqrt{1-1/n}$, $a = \max_{1\leq i\leq n}\{a_i\}$. Using known formulas, this volume is $\sqrt{n}a^{n-1}/(n-1)!$. Multiplying by $\sqrt{n}\overline{w}_t$ yields the desired result.
The second assertion is immediate from (15). □

Note that we can get a bound independent of n by summing over all n in (17) to get

$$\eta_T \leq e^a \overline{w}_T.$$

Figure 6 shows an example for $n = 2$, where the simplex is a line. The area between the boundaries of $\hat{P}_o$ and $\hat{P}_p$ is hence bounded by $\sqrt{2}\overline{w}_T$ multiplied by the length of the simplex S, which is formed by the two sides of length: $a_1\sqrt{1-1/2}$ and $a_2\sqrt{1-1/2}$. Therefore, $\eta_T \leq 2a\overline{w}_T$ for $n = 2$.

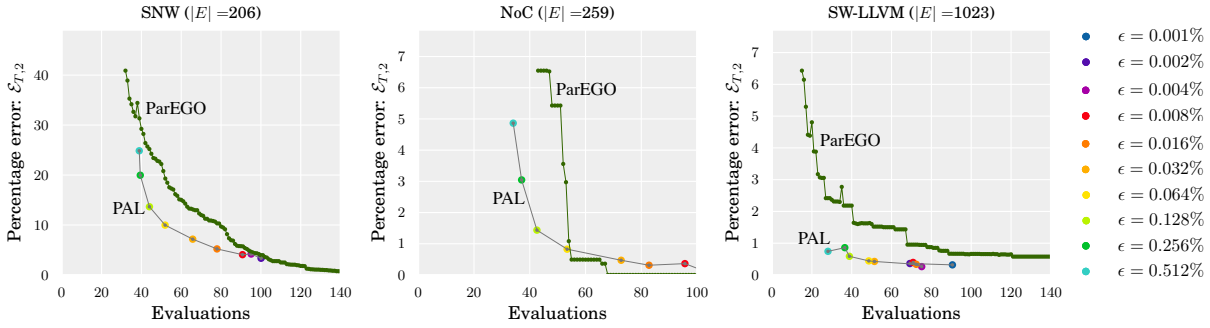

Figure 7. Avg. percentage error in f_2 vs. number of evaluations after termination; for PAL different values for ϵ are used.

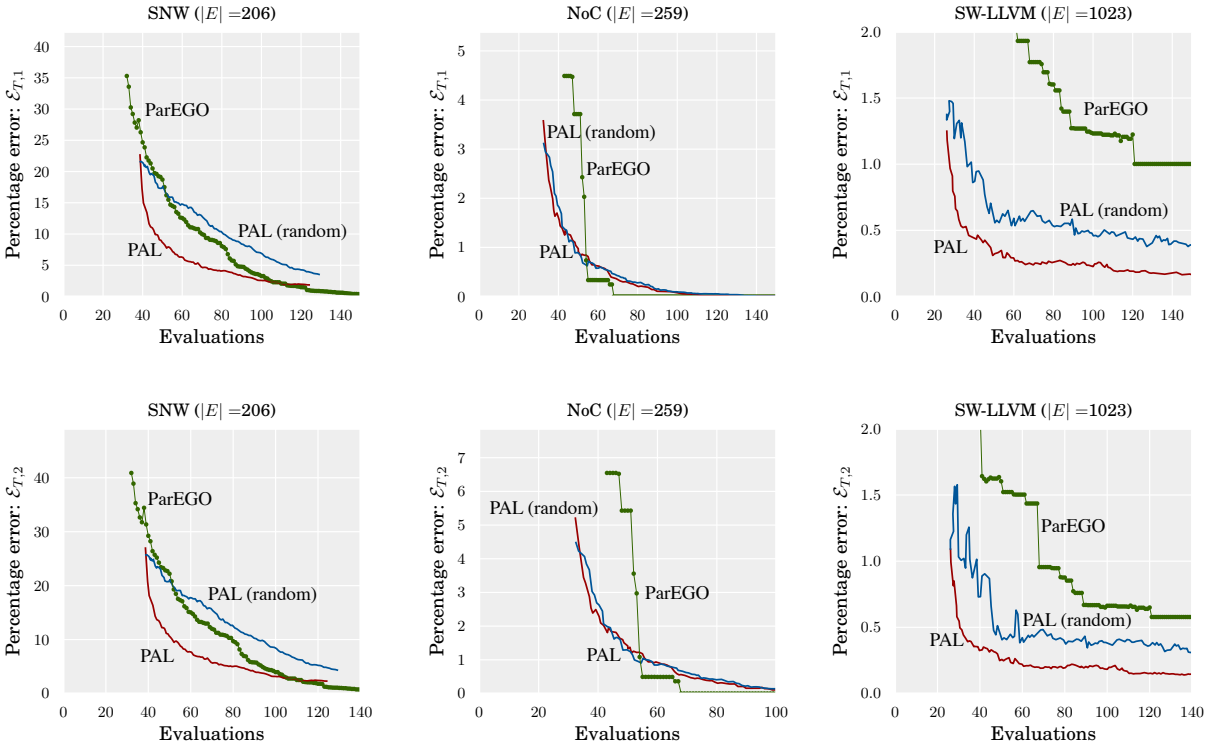

Figure 8. Percentage hypervolume error vs. number of evaluations required by every $\hat{P}_t$.

B. Experiments and Comparison with Random Sampling.

Fig. 7 shows the error obtained in f_2 when PAL stops, for different values of ϵ. The results for f_1 are shown in Fig. 5. The x-axis shows the total number of evaluations of $\boldsymbol{f}$ required to obtain the percentage error on the Pareto prediction displayed on the y-axis of the plots.

We also compare PAL with a variation of PAL that selects the points to evaluate at random from the points that have not been evaluated. At every iteration t after initialization, we generate a prediction $\hat{P}_t$ by adding the Pareto-optimal points of the unclassified points (using predictions $\boldsymbol{\mu}_t$) to the set P_t that contains the points that have been classified as Pareto-optimal at iteration t. We then calculate the error and the cost of this prediction as it has been done in Section 6. Figure 8 shows the results for our three data sets when using an $\epsilon = 0.001\%$ of each range of $\boldsymbol{f}_i$. PAL shows for all data sets better results than PAL with Random sampling, with significantly better results found with the SNW data set. This clearly shows the effectiveness of our sampling strategy in evaluating points that are relevant to achieve the goal of predicting the Pareto-frontier of a design space.