Supplementary Material

Lenient Regret and Good-Action Identification in Gaussian Process Bandits (ICML 2021)

Xu Cai, Selwyn Gomes, and Jonathan Scarlett

A. Discussion on GP-UCB with Intersected Confidence Bounds

The reason that the lenient regret bounds in Theorem 1 grow unbounded as $T \to \infty$ is that $\lim_{t\to\infty} \beta_t = \infty$. For the confidence bounds to remain valid uniformly across time, this appears to be unavoidable. On the other hand, one may consider preventing the UCB and LCB scores from growing unbounded by using *intersected confidence bound*, defined as follows:

$$\overline{\mathrm{ucb}}_t(\mathbf{x}) = \min_{t' \le t} \mathrm{ucb}_{t'}(\mathbf{x}), \tag{17}$$

$$\overline{\mathrm{lcb}}_t(\mathbf{x}) = \max_{t' \le t} \mathrm{lcb}_{t'}(\mathbf{x}), \tag{18}$$

with $\mathrm{ucb}_{t'}(\cdot)$ and $\mathrm{lcb}_{t'}(\cdot)$ given in Lemma 1. Since the original confidence bounds hold uniformly across time with high probability, the same is true for these intersected confidence bounds. We note that this intersecting approach has previously been used in works such as (Bogunovic et al., 2020; Sui et al., 2015).

Unfortunately, we expect that even when the UCB algorithm makes use of $\overline{\mathrm{ucb}}_t(\cdot)$ instead of $\mathrm{ucb}_t(\mathbf{x})$, either the lenient regret still grows unbounded as $t \to \infty$, or it is very challenging the prove that it remains bounded. To understand why we expect such difficulties, consider the scenario in which, in some relatively early round, the UCB score of some bad point $\mathbf{x}_{\mathrm{bad}}$ reaches $f(\mathbf{x}^*) + \epsilon$ for some extremely small $\epsilon > 0$, and then remains there for a long time due to the intersecting done in (17). After a long time, points near $\mathbf{x}^*$ will have been sampled enough times for the UCB scores near $\mathbf{x}^*$ to fall below $f(\mathbf{x}^*) + \epsilon$, meaning the algorithm will return to sampling $\mathbf{x}_{\mathrm{bad}}$ (or some similar/nearby point). However, by this stage, β_t may have grown so large that it takes many samples of $\mathbf{x}_{\mathrm{bad}}$ for the UCB score to fall below $f(\mathbf{x}^*)$, incurring significant regret.

One may envision overcoming this difficulty by showing that the these events of UCB scores falling just above $f(\mathbf{x}^*)$ (and staying there) are unlikely enough to be incorporated into the overall error probability. However, this appears to be a highly non-trivial modification to the analysis, and we make no attempt to do so.

Alternatively, following a similar approach (Bogunovic et al., 2020), one could multiply by β_t by a factor of two in the earlier rounds (e.g., for all $t \le N_{\max}$ with $N_{\max}$ defined in (12)), then revert to the original choice from Lemma 1 in the later rounds, while still intersecting the confidence bounds across time. By doing this, the UCB scores of bad actions that are slightly above $f(\mathbf{x}^*)$ with the doubled confidence bounds will fall below $f(\mathbf{x}^*)$ upon halving. This approach can be used to establish a similar regret bound to that of Theorem 2, but it comes with the rather unnatural step of halving the confidence width after a suitably-chosen number of rounds.

Finally, similar to the previous paragraph, one could adopt an *explore-then-commit* strategy (e.g., see Chapter 6 of (Lattimore & Szepesvári, 2020)). While this could provide a bound on the indicator regret similar to Theorem 2, the hinge and large-gap regrets would be significantly higher due to typically incurring $\Omega(1)$ regret for each bad action sampled. Specifically, the dependence on Δ would be $\frac{1}{\Delta^2}$ instead of the improved $\frac{1}{\Delta}$ appearing in Theorem 2.

B. Proofs of Main Results

In this section, we prove Theorems 1, 2, and 3. We start with some auxiliary results for the upper bounds.

B.1. Auxiliary Results

The analyses of (Srinivas et al., 2010) and (Chowdhury & Gopalan, 2017) are based on first bounding the regret in terms of $\sum_{t=1}^{T} \sigma_{t-1}(\mathbf{x}_t)$, upper bounding this quantity by $\sqrt{T \sum_{t=1}^{T} \sigma_{t-1}^2(\mathbf{x}_t)}$ via Cauchy-Schwartz, and then establishing that

$\sum_{t=1}^{T} \sigma_{t-1}^2(\mathbf{x}_t) \le O(\gamma_T)$. The following lemma gives a useful generalization of the latter statement.

Lemma 2. (Bounding a Sum of Sampled Variances) *For any sequence of sampled points $\mathbf{x}_1, \dots, \mathbf{x}_T$ and any subset $\mathcal{T} \subseteq \{1, \dots, T\}$, letting $N = |\mathcal{T}|$, we have*

$$\sum_{t \in \mathcal{T}} \sigma_{t-1}^2(\mathbf{x}_t) \le C_2 \gamma_N, \tag{19}$$

where $C_2 = \frac{2\lambda^{-1}}{\log(1+\lambda^{-1})}$.

Proof. Denote the N points indexed by $\mathcal{T}$ (i.e., $\{\mathbf{x}_t\}_{t \in \mathcal{T}}$) as $\widetilde{\mathbf{x}}_1, \dots, \widetilde{\mathbf{x}}_N$, where the indexing is done in the order that the points were sampled. For $i = 1, \dots, N$, let $\widetilde{\sigma}_i^2(\mathbf{x})$ be the (hypothetical) GP posterior variance that would arise from sampling $\widetilde{\mathbf{x}}_1, \dots, \widetilde{\mathbf{x}}_i$ alone (note that posterior variance only depends on the sampled locations, not the observations (Rasmussen, 2006)). It is well-known from (Srinivas et al., 2010) that $\sum_{i=1}^{N} \widetilde{\sigma}_{i-1}^2(\widetilde{\mathbf{x}}_i) \le C_2 \gamma_N$, so we only need to show that $\sum_{t \in \mathcal{T}} \sigma_{t-1}^2(\mathbf{x}_t) \le \sum_{i=1}^{N} \widetilde{\sigma}_{i-1}^2(\widetilde{\mathbf{x}}_i)$. Indexing the entries of $\mathcal{T}$ in order by $t_1, \dots, t_N$, the latter claim in turn holds as long as $\sigma_{t_i-1}^2(\mathbf{x}_{t_i}) \le \widetilde{\sigma}_{i-1}^2(\widetilde{\mathbf{x}}_i)$ for all $i = 1, \dots, N$.

By definition, $\mathbf{x}_{t_i}$ is precisely $\widetilde{\mathbf{x}}_i$. Moreover, the posterior variance $\sigma_{t_i-1}^2(\cdot)$ is computed using $t_i - 1$ sampled points, $i-1$ of which are $\widetilde{\mathbf{x}}_1, \dots, \widetilde{\mathbf{x}}_{i-1}$. In contrast, $\widetilde{\sigma}_{i-1}^2(\cdot)$ is computed based on $\widetilde{\mathbf{x}}_1, \dots, \widetilde{\mathbf{x}}_{i-1}$ alone. Since adding points to the set of sampled points cannot increase the posterior variance in a GP model (Rasmussen, 2006), the desired claim $\sigma_{t_i-1}^2(\mathbf{x}_{t_i}) \le \widetilde{\sigma}_{i-1}^2(\widetilde{\mathbf{x}}_i)$ follows, and the proof is complete. □

B.2. Bounding the Number of Bad Actions for GP-UCB

Let $\mathcal{T}_{\rm bad}$ denote the set of times at which GP-UCB chooses a bad action, and let $N = |\mathcal{T}_{\rm bad}|$. By Lemma 2, we have

$$\frac{1}{N} \sum_{t \in \mathcal{T}_{\rm bad}} \sigma_{t-1}^2(\mathbf{x}_t) \le \frac{C_2 \gamma_N}{N}, \tag{20}$$

where we multiplied by $\frac{1}{N}$ on both sides for convenience. Since the minimum is upper bounded by the average, it follows that

$$\min_{t \in \mathcal{T}_{\rm bad}} \sigma_{t-1}^2(\mathbf{x}_t) \le \frac{C_2 \gamma_N}{N}. \tag{21}$$

Now, letting τ denote the time index attaining the minimum in (21), and supposing that the high-probability confidence bound event in Lemma 1 holds, we have

$$\begin{aligned}
{\rm ucb}_\tau(\mathbf{x}_\tau) &= {\rm lcb}_\tau(\mathbf{x}_\tau) + 2\beta_\tau^{1/2} \sigma_{\tau-1}(\mathbf{x}_\tau) & (22)\\
&\le f(\mathbf{x}_\tau) + 2\beta_\tau^{1/2} \sigma_{\tau-1}(\mathbf{x}_\tau) & (23)\\
&\le f(\mathbf{x}^*) - \Delta + 2\beta_\tau^{1/2} \sigma_{\tau-1}(\mathbf{x}_\tau) & (24)\\
&\le f(\mathbf{x}^*) + 2\beta_T^{1/2} \sqrt{\frac{C_2 \gamma_N}{N}} - \Delta & (25)\\
&\le {\rm ucb}_\tau(\mathbf{x}^*) + \sqrt{\frac{C_1 \beta_T \gamma_N}{N}} - \Delta, & (26)
\end{aligned}$$

where:

- (22) follows since the upper and lower confidence bounds differ by $2\beta_\tau^{1/2} \sigma_{\tau-1}(\mathbf{x}_\tau)$;
- (23) and (26) follow from the validity of the confidence bounds, and the latter also defines $C_1 = 4C_2$;
- (24) follows since $f(\mathbf{x}_\tau) \le f(\mathbf{x}^*) - \Delta$ due to $\mathbf{x}_\tau$ being a bad point;
- (25) applies (21), along with $\beta_\tau \le \beta_T$ due to monotonicity.

Since $\mathbf{x}_\tau$ is the point at time τ with the highest UCB score by definition, we observe from (26) that we must have $\sqrt{\frac{C_1 \beta_T \gamma_N}{N}} - \Delta \ge 0$ in order to avoid a contradiction. Re-arranging, we obtain the equivalent condition

$$N \le \frac{C_1 \gamma_N \beta_T}{\Delta^2}. \tag{27}$$

Since this was proved only assuming the validity of the confidence bounds in Lemma 1, which in turn holds with probability at least $1-\delta$, the claim on $\widetilde{R}_T^{\rm ind}$ in Theorem 1 follows.

B.3. Bounding the Large Gap Regret for GP-UCB

Since $\widetilde{R}_T^{\rm hinge} \le \widetilde{R}_T^{\rm gap}$ (see Figure 1), it suffices to upper bound $\widetilde{R}_T^{\rm gap}$. We first write

$$\widetilde{R}_T^{\rm gap} = \sum_{t=1}^T r_t \cdot \mathbb{1}(r_t > \Delta) = \sum_{t \in \mathcal{T}_{\rm bad}} r_t. \tag{28}$$

Following the steps of (Srinivas et al., 2010), and again conditioning on the validity of the confidence bounds in Lemma 1, we have

$$\begin{aligned} r_t &= f(\mathbf{x}^*) - f(\mathbf{x}_t) & (29)\\ &\le \mathrm{ucb}_t(\mathbf{x}^*) - \mathrm{lcb}_t(\mathbf{x}_t) & (30)\\ &= \mathrm{ucb}_t(\mathbf{x}^*) - \mathrm{ucb}_t(\mathbf{x}_t) + 2\beta_t^{1/2}\sigma_{t-1}(\mathbf{x}_t) & (31)\\ &\le 2\beta_t^{1/2}\sigma_{t-1}(\mathbf{x}_t), & (32) \end{aligned}$$

where (30) uses the confidence bounds, (31) follows since the upper and lower confidence bounds differ by $2\beta_t^{1/2}\sigma_{t-1}(\mathbf{x}_t)$, and (32) uses the fact that $\mathbf{x}_t$ is the point with the highest UCB score.

Summing (32) over $t \in \mathcal{T}_{\rm bad}$, upper bounding $\beta_t \le \beta_T$, and applying the Cauchy-Schwartz inequality, we obtain

$$\widetilde{R}_T^{\rm gap} \le \sqrt{4\beta_T |\mathcal{T}_{\rm bad}| \sum_{t \in \mathcal{T}_{\rm bad}} \sigma_{t-1}^2(\mathbf{x}_t)}. \tag{33}$$

Again letting $N = |\mathcal{T}_{\rm bad}|$ denote the number of bad points selected, it follows from Lemma 2 that

$$\widetilde{R}_T^{\rm gap} \le \sqrt{C_1 \beta_T N \gamma_N}. \tag{34}$$

Since we already established that N satisfies (27) when the confidence bounds are valid, we can further bound

$$\widetilde{R}_T^{\rm gap} \le \frac{C_1 \beta_T \gamma_N}{\Delta}. \tag{35}$$

The bound on $\widetilde{R}_T^{\rm gap}$ in Theorem 1 follows by substituting $N \le N_{\max}$ and using the monotonicity of γ_N.

B.4. Bounding the Number of Bad Actions for the Elimination Algorithm

Our analysis uses similar ingredients as in (Bogunovic et al., 2016; Contal et al., 2013; Srinivas et al., 2010). We first note the well-known fact that as long as the confidence bounds in Lemma 1 are valid, the algorithm never eliminates $\mathbf{x}^*$. This is because having the UCB of $\mathbf{x}^*$ be below another point's LCB would contradict the optimality of $\mathbf{x}^*$.

Suppose that the elimination algorithm has run up to some number of rounds N. Using Lemma 2 with $\mathcal{T} = \{1, \dots, N\}$, we have

$$\frac{1}{N}\sum_{t=1}^N \sigma_{t-1}^2(\mathbf{x}_t) \le \frac{C_2 \gamma_N}{N}, \tag{36}$$

where we again divided both sides by N for convenience. Using the standard property that the GP posterior variance always decreases as more points are selected, and noting the algorithm chooses the point with the highest variance, we find that $\sigma_{N-1}^2(\mathbf{x}_N)$ is the smallest summand in (36), and hence

$$\sigma_{N-1}^2(\mathbf{x}_N) \le \frac{C_2 \gamma_N}{N}. \tag{37}$$

Moreover, since $\mathbf{x}_N$ is defined to maximize $\sigma_{N-1}^2(\cdot)$, it follows that

$$\max_{\mathbf{x} \in M_{N-1}} \sigma_{N-1}^2(\mathbf{x}) \le \frac{C_2 \gamma_N}{N}. \tag{38}$$

That is, all non-eliminated points have posterior variance at most $\frac{C_2\gamma_N}{N}$ after time N.

We now fix an arbitrary non-eliminated bad point $\mathbf{x}_{\rm bad}$, and note the following analogous steps to (22)–(26) (whose explanations are similar and thus mostly omitted):

$$\begin{aligned}
\mathrm{ucb}_N(\mathbf{x}_{\rm bad}) &= \mathrm{lcb}_N(\mathbf{x}_{\rm bad}) + 2\beta_N^{1/2}\sigma_{N-1}(\mathbf{x}_{\rm bad}) && (39)\\
&\leq f(\mathbf{x}_{\rm bad}) + 2\beta_N^{1/2}\sigma_{N-1}(\mathbf{x}_{\rm bad}) && (40)\\
&\leq f(\mathbf{x}^*) - \Delta + 2\beta_N^{1/2}\sigma_{N-1}(\mathbf{x}_{\rm bad}) && (41)\\
&\leq \mathrm{lcb}_N(\mathbf{x}^*) - \Delta + 2\beta_N^{1/2}\sigma_{N-1}(\mathbf{x}_{\rm bad}) + 2\beta_N^{1/2}\sigma_{N-1}(\mathbf{x}^*) && (42)\\
&\leq \mathrm{lcb}_N(\mathbf{x}^*) + 2\sqrt{\frac{C_1\beta_N\gamma_N}{N}} - \Delta, && (43)
\end{aligned}$$

where (43) applies (38) for both $\mathbf{x} \in \{\mathbf{x}_{\rm bad}, \mathbf{x}^*\}$.

Since (43) applies to an arbitrary non-eliminated bad point, we find that in order for any bad points to remain non-eliminated after time N, it must be the case that $2\sqrt{\frac{C_1\beta_T\gamma_N}{N}} - \Delta \geq 0$, or equivalently,

$$N \leq \frac{4C_1\gamma_N\beta_N}{\Delta^2}. \qquad (44)$$

In other words, all bad points are eliminated after time $N'_{\max}$, with $N'_{\max}$ defined in (13). This proves the first part of Theorem 2.

B.5. Bounding the Large Gap Regret for the Elimination Algorithm

While we performed the analysis leading to (44) considering the number of pulls of Δ-suboptimal points, we can similarly replace Δ by any positive value $\widetilde{\Delta}$ and reach a similar conclusion. In the following, it is more convenient to rephrase (44) by expressing Δ in terms of N as $\Delta \leq \sqrt{\frac{4C_1\gamma_N\beta_N}{N}}$. Replacing Δ by a generic value of $\widetilde{\Delta}$, and replacing N by a generic time index t, it follows that after t iterations, all non-eliminated arms have regret upper bounded by $\widetilde{\Delta}_t$, where

$$\widetilde{\Delta}_t = \sqrt{\frac{4C_1\gamma_t\beta_t}{t}}. \qquad (45)$$

To bound the large gap regret, we simply sum the regret over all time indices up to $N'_{\max}$, after which we already know from the above analysis that no further (lenient) regret is incurred. We additionally treat $t = 1$ as a special case, noting that the regret incurred is at most $2B$ since $\|f\|_k \leq B$ (and thus $|f(\mathbf{x})| \leq B$ for all $\mathbf{x}$), yielding

$$\begin{aligned}
\widetilde{R}_T^{\rm gap} &\leq 2B + \sum_{t=2}^{N'_{\max}} \widetilde{\Delta}_{t-1} && (46)\\
&\leq 2B + \sum_{t=1}^{N'_{\max}} \widetilde{\Delta}_t && (47)\\
&\leq 2B + \sum_{t=1}^{N'_{\max}} \sqrt{\frac{4C_1\gamma_t\beta_t}{t}} && (48)\\
&\leq 2B + \sqrt{4C_1\gamma_{N'_{\max}}\beta_{N'_{\max}}} \sum_{t=1}^{N'_{\max}} \frac{1}{\sqrt{t}} && (49)\\
&\leq 2B + 4\sqrt{C_1 N'_{\max}\gamma_{N'_{\max}}\beta_{N'_{\max}}}, && (50)
\end{aligned}$$

where (48) uses the definition of $\widetilde{\Delta}_t$, (49) uses the monotonicity of γ_t and β_t, and (50) uses the fact that $\sum_{t=1}^N \frac{1}{\sqrt{t}} \leq 2\sqrt{N}$. Finally, by definition in (13), we have $N'_{\max} \leq \frac{4C_1\gamma_{N'_{\max}}\beta_{N'_{\max}}}{\Delta^2}$, and substituting into (50) yields $\widetilde{R}_T^{\rm gap} \leq 2B + \frac{8C_1\gamma_{N'_{\max}}\beta_{N'_{\max}}}{\Delta}$, as desired.

B.6. Proofs of the Lower Bounds

Since our lower bounds follow in a fairly straightforward manner from the analysis in (Cai & Scarlett, 2021), we do not attempt to give a self-contained analysis (which would require considerable repetition with (Cai & Scarlett, 2021; Scarlett et al., 2017)), and instead only state the differences.

The analysis depends on a parameter $\epsilon > 0$ that is initially arbitrary, and that we will set differently to (Cai & Scarlett, 2021) to account for the different regret notion. A *hard subset* of functions $\{f_1, \dots, f_M\} \in \mathcal{F}_k(B/3)$ is constructed in a manner such that any given action $x \in D$ is ϵ-optimal for at most one function. It is shown in (Scarlett et al., 2017) that such a subset exists with the following choices of M depending on the kernel:

- For the SE kernel, we can set
$$M = \left\lfloor \left(\frac{c_1 \sqrt{\log \frac{B(2\pi l^2)^{d/4}}{\epsilon}}}{l} \right)^d \right\rfloor, \tag{51}$$
where c_1 is a universal positive constant, and l denotes the length-scale.
- For the Matérn kernel, we can set
$$M = \left\lfloor \left(\frac{Bc_3}{\epsilon} \right)^{d/\nu} \right\rfloor, \tag{52}$$
where $c_3 := \big(\frac{1}{\zeta}\big)^\nu \cdot \big(\frac{c_2^{-1/2}}{2(8\pi^2)^{(\nu+d/2)/2}}\big)$, and where $\zeta > 0$ and $c_2 > 0$ are constants.

Once the existence of this function class is established, the analysis in (Cai & Scarlett, 2021) shows that there exists a function $f \in \mathcal{F}_k(B)$ and constant c_0 such that when the time horizon satisfies

$$T < \frac{(M-1)\sigma^2}{2c_0\epsilon^2} \log \frac{1}{2.4\delta}, \tag{53}$$

it must hold with probability at least δ that ϵ-suboptimal actions are selected in at least $\frac{T}{2}$ rounds.

We now turn to the part of the analysis that differs from (Cai & Scarlett, 2021). We first use the trivial fact that the cumulative regret up to time T is lower bounded by that up to any $\widetilde{T} \le T$. We consider $\widetilde{T}$ being slightly below the threshold in (53) (or capped to T):

$$\widetilde{T} = \min\left\{ T, \frac{M\sigma^2}{4c_0\epsilon^2} \log \frac{1}{2.4\delta} \right\}, \tag{54}$$

and since this choice is smaller than the right-hand side of (53), we know that ϵ-suboptimal actions must be played at least $\frac{\widetilde{T}}{2}$ times.

To lower bound the lenient regret in the case that $\Phi = \Phi^{\rm ind}$, we simply set $\epsilon = \Delta$, so that being ϵ-suboptimal is exactly equivalent to being a bad action. In this case, the desired lower bounds follow directly by substituting (51) and (52) into (54) and lower bounding the lenient regret by $\frac{\widetilde{T}}{2}$. Note that the assumption $\frac{\Delta}{B} = O(1)$ (with a small enough implied constant) implies that (51) and (52) scale as $\Theta\big(\big(\log\frac{B}{\epsilon}\big)^{d/2}\big)$ and $\Theta\big(\big(\frac{B}{\Delta}\big)^{d/\nu}\big)$ respectively.

To lower bound the lenient regret in the case that $\Phi = \Phi^{\rm hinge}$, we notice from the definition of the hinge function that if a 2Δ-suboptimal point is selected, then the contribution to the lenient regret is still at least Δ. Hence, the desired lower bounds follow by setting $\epsilon = 2\Delta$, substituting (51) and (52) into (54), and lower bounding the lenient regret by $\frac{\widetilde{T}\Delta}{2}$. Finally, the inequality $\widetilde{R}_T^{\rm gap} \ge \widetilde{R}_T^{\rm hinge}$ is trivial by definition (see Figure 1).

C. Additional Good-Action Identification Algorithms

C.1. Satisficing Thompson Sampling (STS)

Thompson sampling (TS) samples actions randomly according to the posterior probability of being optimal (Russo et al., 2018). To adapt TS to the good-action identification problem, we follow an idea proposed in (Russo & Van Roy, 2018) for multi-armed bandits, termed *satisficing Thompson sampling* (STS). In the finite-arm setting, the STS approach samples according to the probability of being the good arm *with the lowest index*.

In our continuous-domain setting, there is no natural order over the arms, so we instead consider the following natural analog: Seek the good action *closest to some fixed point* $\mathbf{x}^{\mathrm{c}}$ (with the default value being the domain center). The resulting algorithm is as follows:

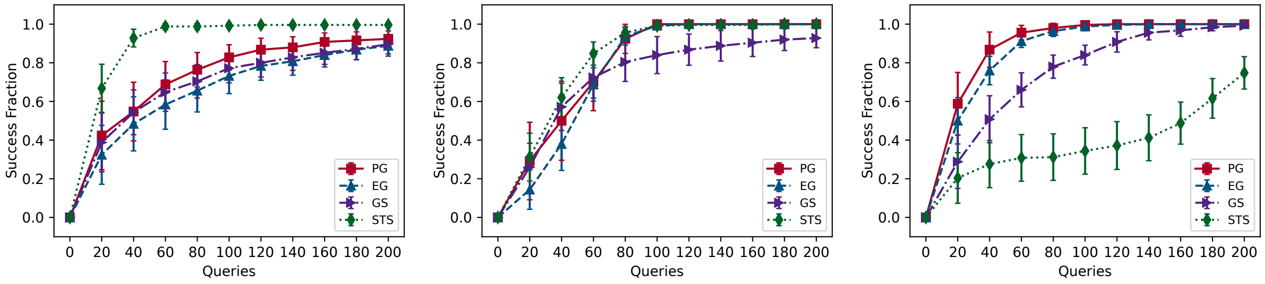
- Let $\tilde{f}_t$ be a sample from the GP posterior distribution given the first $t-1$ observations;
- Choose $\mathbf{x}_t$ to maximize the following acquisition function:

$$\alpha_t^{\mathrm{STS}}(\mathbf{x}) = \begin{cases} -\|\mathbf{x}-\mathbf{x}^{\mathrm{c}}\| & \tilde{f}_t(\mathbf{x}) \geq \eta \\ -\infty & \text{otherwise.} \end{cases} \tag{55}$$

It may be that none of the points in the domain satisfy $\tilde{f}_t(\mathbf{x}) \geq \eta$, in which case we simply let $\mathbf{x}_t$ be a maximizer of $\tilde{f}_t$ (i.e., revert to regular TS).

This approach is primarily suited to scenarios where prior knowledge is available on the approximate location of the maximizer or a good region (captured by $\mathbf{x}^{\mathrm{c}}$). Since such knowledge is typically unavailable, we only investigate STS in some proof-of-concept experiments here; further studies of TS-type methods for good-action identification is left for future work. The experimental details are as described in Section 5, and the results shown in Figure 6.

For the Dropwave function the optimal action is precisely at the domain center ($\mathbf{x}^* = \mathbf{0}$), and accordingly, STS performs much better than the other methods. For the Keane function it is near the center ($\mathbf{x}^* = (1.39, 0)$ or $(0, 1.39)$), and STS remains competitive with PG. Finally, when we shift the Dropwave function so that the good actions are near the boundary ($\mathbf{x}^* = (-5.12, 5.12)$), we find that STS performs significantly worse. Thus, these experiments provide evidence that prior knowledge of an approximate function maximizer (or at least a "good region") is important for our version of STS to perform well.



(a) Dropwave (b) Keane (c) Shifted Dropwave

Figure 6. Experimental results for good-action identification with Satisficing Thompson Sampling (STS).

C.2. Elimination Algorithm

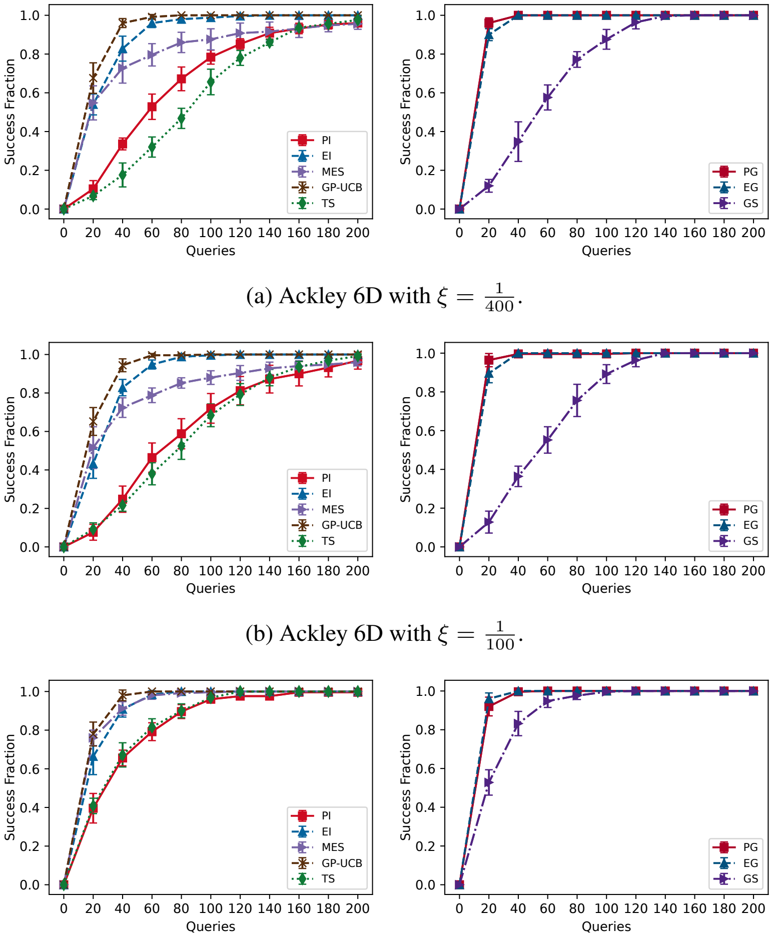
We briefly mention that one can modify the elimination algorithm described in Section 2.4 by eliminating all actions whose UCB score is below η, rather than those whose UCB is below the highest LCB. That is, we modify (11) as follows:

$$M_t = \big\{\mathbf{x} \in M_{t-1} \,:\, \mathrm{ucb}_t(\mathbf{x}) \geq \eta\big\}. \tag{56}$$

At the times of primary interest where no good action has been found yet, η will typically be significantly above the highest LCB score, and hence, more bad actions will be eliminated earlier compared to when using (11). However, as discussed in Section 3.2, elimination algorithms are susceptible to complete failure under kernel misspecification, and we thus do not include this approach in our experiments, in which the kernel hyperparameters are learned online.

D. Additional Experiments

Here we present further experiments for good-action identification, adopting the same setup as described in Section 5.2 except where stated otherwise.



(a) Ackley 6D with $\xi = \frac{1}{400}$.

(b) Ackley 6D with $\xi = \frac{1}{100}$.

(c) Ackley 6D with $\xi = \frac{1}{50}$.

Figure 7. Ackley 6D function for different values of η dictated by $\xi \in (0, 1)$, the approximate proportion of points that are good.

D.1. Comparison of Different Threshold Values

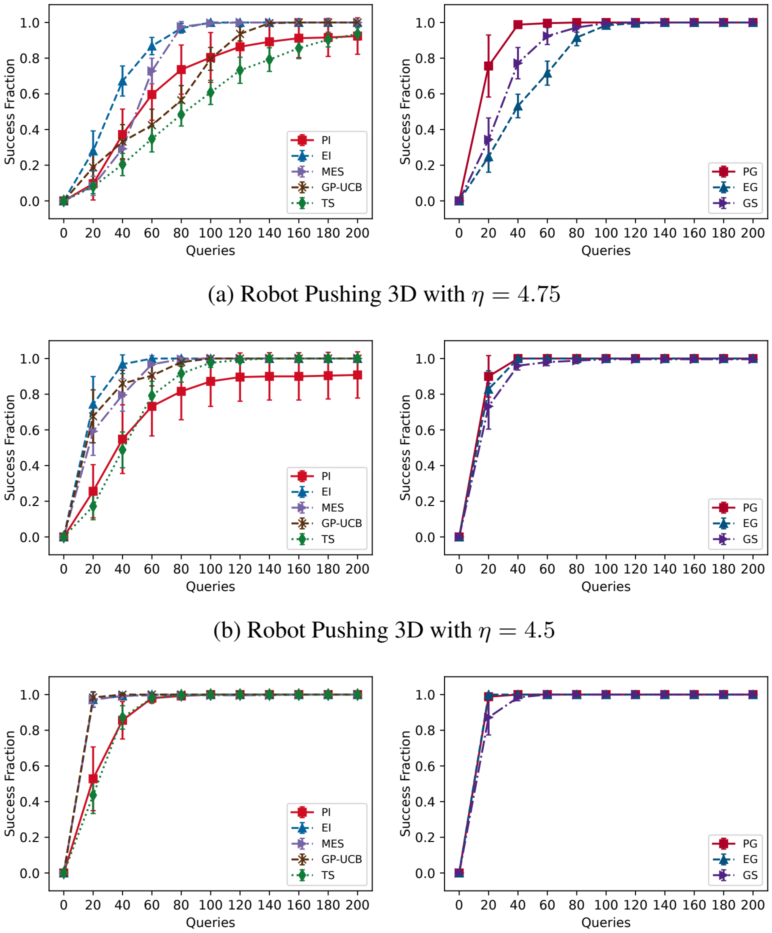
We explore the effect of varying η using the Ackley function and the robot pushing function. For the Ackley function, we consider choosing η such that roughly a fraction ξ of points are good, as detailed in Section 5.2. The results for $\eta \in \left\{\frac{1}{400}, \frac{1}{100}, \frac{1}{50}\right\}$ are shown in Figure 7. For the robot pushing objective, we choose $\eta \in \{4.0, 4.5, 4.75\}$, and the results are shown in Figures 8 and 9 (3D and 4D versions, respectively).

In each experiment, we observe fairly similar behavior for each good-action threshold, but we find that increasing ξ (or equivalently, decreasing η) naturally makes all algorithms find good points faster. A somewhat less obvious finding is that this also tends to bring all of the curves closer together, suggesting that most "reasonable" algorithms can quickly find a good action when sufficiently many of them exist.

D.2. Cases When No Good Action Exists

A potential concern of the good-action identification perspective is whether the algorithms can still be expected to behave in a reasonable manner when no good actions exist. Here we provide evidence that, in fact, one can still maintain robustness, in the sense that even when $\eta > f(\mathbf{x}^*)$, the algorithms introduced in Section 4 can still find an action with function value close to $f(\mathbf{x}^*)$. To demonstrate this, we revert to the standard simple regret notion (since the "fraction found" notion used previously will always be zero here).

Figure 10 plots the simple regret for the 3D Hartmann function (with $f(\mathbf{x}^*) = 3.863$). In sub-figure (a), we consider both η

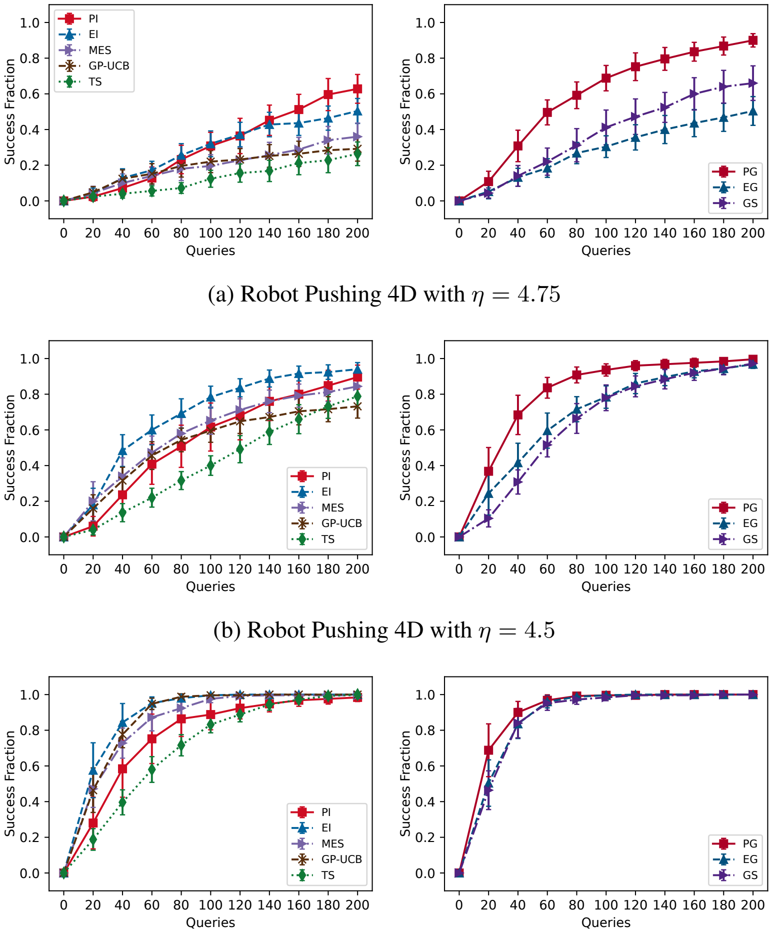


(a) Robot Pushing 3D with $\eta = 4.75$

(b) Robot Pushing 3D with $\eta = 4.5$

(c) Robot Pushing 3D with $\eta = 4.0$

Figure 8. Robot Pushing 3D function for different values of η

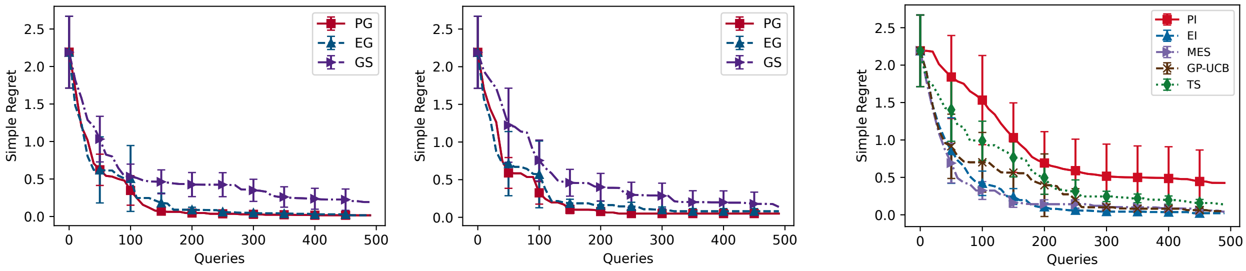


(a) Robot Pushing 4D with $\eta = 4.75$

(b) Robot Pushing 4D with $\eta = 4.5$

(c) Robot Pushing 4D with $\eta = 4.0$

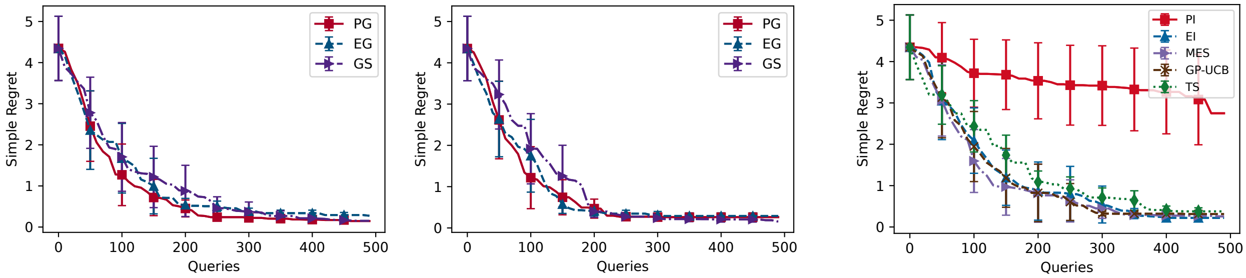
Figure 9. Robot Pushing 4D function for different values of η



(a) Our algorithms with $\eta = f(\mathbf{x}^*) + 0.1$ (Left) and $\eta = f(\mathbf{x}^*) + 0.5$ (Right).

(b) Standard optimization algorithms.

Figure 10. Simple regret plots for the 3D Hartmann function when no good action exists.



(a) Our algorithms with $\eta = f(\mathbf{x}^*) + 0.1$ (Left) and $\eta = f(\mathbf{x}^*) + 0.5$ (Right).

(b) Standard optimization algorithms.

Figure 11. Simple regret plots for the Robot Pushing 3D function when no good action exists.

slightly above the threshold, and significantly above. Even in the latter case, PG and EG are able to attain simple regret tending to zero, indicating their robustness in the case that no good points exist. While GS appears to be somewhat less robust, this could potentially be remedied by modifying how the algorithm behaves when all acquisition functions are zero, as discussed in Section 4.3.

An analogous plot for the robot pushing experiment is given in Figure 11, with similar findings. We note that the poor performance of PI here is due to the existence of a small number of runs in which the algorithm gets stuck in a highly suboptimal local minimum. These runs significantly impact the average regret, but only have a minor impact on the cumulative fraction found in Figure 4 (due to occurring on few runs).