A Justification for The Acquisition Functions Being Sampled from GPs

For our Algo. 1, the work of [24] has shown that if $\beta_t = 1$, then after running lines 4-7 of Algo. 1, the resulting function $f_t^i(\mathbf{x}; \theta_t^i)$ corresponds to a function sampled from the GP posterior with the NTK as the kernel function: $\mathcal{GP}(\mu_{t-1}(\cdot), \sigma_{t-1}^2(\cdot,\cdot))$ conditioned on the $(t-1) \times B$ observations from the first $t-1$ iterations. The GP posterior mean and covariance function are expressed as:

$$\mu_{t-1}(\mathbf{x}) \triangleq \mathbf{k}_{t-1}(\mathbf{x})^\top (\mathbf{K}_{t-1} + \sigma^2 \mathbf{I})^{-1} \mathbf{y}_{t-1} \,, \tag{1}$$

$$\sigma_{t-1}^2(\mathbf{x}, \mathbf{x}') \triangleq k(\mathbf{x}, \mathbf{x}') - \mathbf{k}_{t-1}(\mathbf{x})^\top (\mathbf{K}_{t-1} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}_{t-1}(\mathbf{x}') \tag{2}$$

where $\mathbf{k}_{t-1}(\mathbf{x}) \triangleq (k(\mathbf{x}, \mathbf{x}_\tau^i))^\top_{\tau=1,\ldots,t-1, i=1,\ldots,B}$ which is a $(t-1)B-$dimensional vector, $\mathbf{y}_t \triangleq (y_\tau^i)^\top_{\tau=1,\ldots,t-1,i=1,\ldots,B}$ which is also a $(t-1)B-$dimensional vector, and $\mathbf{K}_t \triangleq (k(\mathbf{x}_\tau^i, \mathbf{x}_{\tau'}^{i'}))_{\tau=1,\ldots,t-1,i=1,\ldots,B;\tau'=1,\ldots,t-1,i'=1,\ldots,B}$ which is a $(t-1)B \times (t-1)B-$dimensional squared matrix.

For our Algo. 2, when $\beta_t = 1$, because every run of the procedure in lines 4-6 corresponds to running the sample-then-optimize method [42] while treating the neural tangent features as the input features, therefore, the resulting linear function $f_t^i(\mathbf{x}; \theta_t^i)$ w.r.t. θ_t^i also corresponds to a sampled function from the GP posterior $\mathcal{GP}(\mu_{t-1}(\cdot), \sigma_{t-1}^2(\cdot,\cdot))$ with the NTK (if the NN is infinite-width) or empirical NTK (if the NN if finite-width) as the kernel function according to the work of [42].

Next, for both Algo. 1 and Algo. 2, when $\beta_t = 2\log(\pi^2 t^2 |\mathcal{X}|/\delta)$, since we have multiplied the output of NN by β_t which corresponds to multiplying the gradient of the NN by β_t, therefore, the resulting NTK will be multiplied by β_t^2. Also note that we have also multiplied the noise variance σ^2 by β_t^2 in (1). As a result, after plugging these two changes into the equations for GP posterior mean (2) and variance (3), it is easy to verify that the GP posterior variance will be multiplied by β_t^2 while the GP posteior mean is unchanged.

B GP Posterior Variance with NTK

When using the NTK as the kernel function, the GP posterior variance (3) at any input $\mathbf{x}$ can be easily approximated by

$$\sigma_{t-1}^2(\mathbf{x}, \mathbf{x}) \approx \nabla_\theta f(\mathbf{x}; \theta_0)^\top \left[\Sigma_{t-1} + \sigma^2 I\right]^{-1} \nabla_\theta f(\mathbf{x}; \theta_0), \tag{3}$$

in which

$$\Sigma_{t-1} = \sum_{\tau=0}^{t-1} \sum_{i=1}^{B} \nabla_\theta f(\mathbf{x}_\tau^i; \theta_0) \nabla_\theta f(\mathbf{x}_\tau^i; \theta_0)^\top, \tag{4}$$

and $\theta_0 \sim \text{init}(\cdot)$ are randomly initialized parameters. Therefore, to run the uncertainty sampling algorithm as the initialization stage, we simply need to sequentially maximize equation (4), i.e., in iteration t of the initialization stage, we simply choose the next initial input $\mathbf{x}$ by maximizing equation (4).

C Proof of Theorem 1

Here, to simplify the analysis, we follow the work of [17] and reparameterize the iterations to view our algorithms in the sequential setting. Specifically, in the main text (Algos. 1 and 2), every B function evaluations are counted as an iteration t; however, we reparameterize the iterations such that every query selection is counted as an iteration t. That is, every time an input query is selected, we increment the number of iterations by 1. As a result, before the reparameterization, the cumulative regret is expressed as $R_T = \sum_{t=1}^{T/B} \sum_{i=1}^{B} (f(\mathbf{x}^*) - f(\mathbf{x}_t^i))$; after reparameterization, the same cumulative regret is now expressed as $R_T = \sum_{t=1}^{T} (f(\mathbf{x}^*) - f(\mathbf{x}_t))$. Note that when $B = 1$, the two parameterizations are the same. Therefore, in the entire proof in this section, we index the iterations sequentially by $1, 2, \ldots, t, t+1, \ldots, T$.

At iteration t, we use fb[t] to denote the largest iteration index whose observation has been collected. For example, if the batch size is $B = 3$, assuming that after the most recent batch of inputs

have been collected, we have in total gathered $t-1$ observations; then when selecting the input queries in iterations t, $t+1$ and $t+2$, we have that $\text{fb}[t] = \text{fb}[t+1] = \text{fb}[t+2] = t-1$, because the index of the most recent observation is fixed at $t-1$ since we do not collect any new observations during this process. Next, when choosing the input query at iterations $t+3$, $t+4$ and $t+5$, we have that $\text{fb}[t+3] = \text{fb}[t+4] = \text{fb}[t+5] = t+2$. As a result of our reparameterization here, the requirement on the constant C from Theorem 1 should be slightly modified into: $\max_{A\subset\mathcal{X},|A|\leq B-1} \mathbb{I}(f;\mathbf{y}_A|\mathbf{y}_{1:\text{fb}[t]}) \leq C, \forall t \geq 1$, where $\mathbf{y}_{1:\text{fb}[t]}$ represents the output observations from iterations 1 to $\text{fb}[t]$.

Our reparameterization mentioned above allows us to derive more general theoretical results which hold for both synchronous and asynchronous batch BO. In the setting of synchronous batch evaluations which we have focused on in the main text, $\max\{t-B,0\} \leq \text{fb}[t] \leq t-1$. In this setting, $\text{fb}[t]$ is a deterministic function of t, which is determined before the algorithm starts. Of note, although we focus on the setting of synchronous batch BO in our theoretical analysis, the only requirement of our theoretical analysis on $\text{fb}[t]$ is that $t - \text{fb}[t] \leq B$, i.e., the number of pending observations $t - \text{fb}[t] - 1$ should be upper-bounded by $B-1$. Therefore, our theoretical results also hold in the setting of asynchronous batch BO, because the number of pending observations in asynchronous batch BO is always equal to $B-1$ [17].

Denote as $\mu_{\text{fb}[t]}$ and $\sigma_{\text{fb}[t]}$ the GP posterior mean and standard deviation conditioned on the observations from iteration 1 to $\text{fb}[t]$. Define $\mathcal{F}_{t-1} = \{\mathbf{x}_1, y_1, \ldots, \mathbf{x}_{\text{fb}[t]}, y_{\text{fb}[t]}, \mathbf{x}_{\text{fb}[t]+1}, \ldots, \mathbf{x}_{t-1}\}$ as the history of selected inputs and observed outputs for those completed observations, as well as the selected inputs of those pending observations. Define $\beta_t = 2\log(\pi^2 t^2 |\mathcal{X}|/(3\delta))$, and $c_t = \beta_t(1+\sqrt{2\log(|\mathcal{X}|t^2)})$.

Lemma 1. *Choose $\delta \in (0,1)$. Define $E^f(t)$ as the event that $|\mu_{\text{fb}[t]}(\mathbf{x}) - f(\mathbf{x})| \leq \beta_t \sigma_{\text{fb}[t]}(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}$. We have that $\mathbb{P}(E^f(t)) \geq 1-\delta/2, \forall t \geq 1$.*

The proof of Lemma 1, which follows from the proof of Lemma 5.1 of [57], makes use of our assumption that f is sampled from a GP and relies on simple applications of the concentration of Gaussian distributions and union bounds. Denote by f_t the acquisition function in iteration t, which is sampled from the GP posterior with the NTK as the kernel function: $f_t \sim \mathcal{GP}(\mu_{\text{fb}[t]}(\cdot), \beta_t^2 \sigma^2_{\text{fb}[t]}(\cdot,\cdot))$ as we have justified in Appendix A. Note that the query in iteration t is selected by $\mathbf{x}_t = \arg\max_{\mathbf{x}\in\mathcal{X}} f_t(\mathbf{x})$, which corresponds to line 8 of Algo. 1 and line 7 of Algo. 2 respectively.

Lemma 2. *Define $E^{f_t}(t)$ as the event that $|\mu_{\text{fb}[t]}(\mathbf{x}) - f_t(\mathbf{x})| \leq \beta_t \sqrt{2\log(|\mathcal{X}|t^2)}\sigma_{\text{fb}[t]}(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}$. We have that $\mathbb{P}(E^{f_t}(t)) \geq 1 - 1/t^2, \forall t \geq 1$.*

The proof of Lemma 2 follows from Lemma 5 of the work of [7]. Importantly, conditioned on both events $E^f(t)$ and $E^{f_t}(t)$, we have that

$$|f(\mathbf{x}) - f_t(\mathbf{x})| \leq c_t \sigma_{\text{fb}[t]}(\mathbf{x}). \tag{5}$$

We next define the set of *saturated points*, which can be understood as the set of undesirable points in every iteration.

Definition 1. *Define the set of saturated inputs in iteration t as*

$$S_t = \{\mathbf{x} \in \mathcal{X} : \Delta(\mathbf{x}) > c_t \sigma_{\text{fb}[t]}(\mathbf{x})\},$$

in which $\Delta(\mathbf{x}) = f(\mathbf{x}^*) - f(\mathbf{x})$.

An important consequence of the definition above is that $\mathbf{x}^*$ is always unsaturated, because $\Delta(\mathbf{x}^*) = 0 < c_t \sigma_{\text{fb}[t]}(\mathbf{x}^*)$.

Lemma 3. *For any $\mathcal{F}_{t-1}$, conditioned on the events $E^f(t)$, we have that $\forall \mathbf{x} \in \mathcal{X}$,*

$$\mathbb{P}\left(f_t(\mathbf{x}) > f(\mathbf{x}) | \mathcal{F}_{t-1}\right) \geq p, \tag{6}$$

in which $p = \frac{1}{4e\sqrt{\pi}}$.

Proof.

$$\begin{aligned}
\mathbb{P}\left(f_t(\mathbf{x}) > f(\mathbf{x})|\mathcal{F}_{t-1}\right) &= \mathbb{P}\left(\frac{f_t(\mathbf{x}) - \mu_{\text{fb}[t]}(\mathbf{x})}{\beta_t \sigma_{\text{fb}[t]}(\mathbf{x})} > \frac{f(\mathbf{x}) - \mu_{\text{fb}[t]}(\mathbf{x})}{\beta_t \sigma_{\text{fb}[t]}(\mathbf{x})}\Big|\mathcal{F}_{t-1}\right) \\
&\geq \mathbb{P}\left(\frac{f_t(\mathbf{x}) - \mu_{\text{fb}[t]}(\mathbf{x})}{\beta_t \sigma_{\text{fb}[t]}(\mathbf{x})} > \frac{|f(\mathbf{x}) - \mu_{\text{fb}[t]}(\mathbf{x})|}{\beta_t \sigma_{\text{fb}[t]}(\mathbf{x})}\Big|\mathcal{F}_{t-1}\right) \\
&\overset{(a)}{\geq} \mathbb{P}\left(\frac{f_t(\mathbf{x}) - \mu_{\text{fb}[t]}(\mathbf{x})}{\beta_t \sigma_{\text{fb}[t]}(\mathbf{x})} > 1\Big|\mathcal{F}_{t-1}\right) \\
&\overset{(b)}{\geq} \frac{e^{-1}}{4\sqrt{\pi}}.
\end{aligned} \tag{7}$$

(a) follows from Lemma 1, which holds because we condition on the event $E^f(t)$ here. (b) follows since $f_t(\mathbf{x}) \sim \mathcal{N}(\mu_{\text{fb}[t]}(\mathbf{x}), \beta_t^2 \sigma_{\text{fb}[t]}^2(\mathbf{x}))$ and makes use of the Gaussian anti-concentration inequality, i.e., $\mathbb{P}(Z > a) \geq \frac{e^{-a^2}}{4\sqrt{\pi}a}$ where Z follows a standard Gaussian distribution. □

The next Lemma shows that the probability that the selected input is unsaturated (i.e., desirable according to Definition 1) can be lower-bounded.

Lemma 4. *For any $\mathcal{F}_{t-1}$, conditioned on the event $E^f(t)$, we have that*

$$\mathbb{P}\left(\mathbf{x}_t \in \mathcal{X} \setminus S_t, |\mathcal{F}_{t-1}\right) \geq p - 1/t^2.$$

Proof. To begin with, we can lower-bounded the probability that the selected $\mathbf{x}_t$ is unsaturated as follows:

$$\mathbb{P}\left(\mathbf{x}_t \in \mathcal{X} \setminus S_t|\mathcal{F}_{t-1}\right) \geq \mathbb{P}\left(f_t(\mathbf{x}^*) > f_t(\mathbf{x}), \forall \mathbf{x} \in S_t|\mathcal{F}_{t-1}\right). \tag{8}$$

The inequality above holds because the event on the right hand sight implies the event on the left hand side. Specifically, because $\mathbf{x}^*$ is always unsaturated (Definition 1), therefore, as long as $f_t(\mathbf{x}^*) > f_t(\mathbf{x}), \forall \mathbf{x} \in S_t$, then the selected $\mathbf{x}_t$ is guaranteed to be unsaturated because it is selected as $\mathbf{x}_t = \arg\max_{\mathbf{x} \in \mathcal{X}} f_t(\mathbf{x})$.

Next, we assume that both events $E^f(t)$ and $E^{f_t}(t)$ holds, which allows us to derive an upper bound on $f_t(\mathbf{x})$ for all $\mathbf{x} \in S_t$:

$$f_t(\mathbf{x}) \overset{(a)}{\leq} f(\mathbf{x}) + c_t \sigma_{\text{fb}[t]}(\mathbf{x}) \overset{(b)}{\leq} f(\mathbf{x}) + \Delta(\mathbf{x}) = f(\mathbf{x}) + f(\mathbf{x}^*) - f(\mathbf{x}) = f(\mathbf{x}^*), \tag{9}$$

in which (a) results from Lemma 1 and Lemma 2 and (b) follows from Definition 1. As a result,equation (10) implies that when both both events $E^f(t)$ and $E^{f_t}(t)$ hold, we have that

$$\mathbb{P}\left(f_t(\mathbf{x}^*) > f_t(\mathbf{x}), \forall \mathbf{x} \in S_t|\mathcal{F}_{t-1}\right) \geq \mathbb{P}\left(f_t(\mathbf{x}^*) > f(\mathbf{x}^*)|\mathcal{F}_{t-1}\right). \tag{10}$$

Next, combining equations (9) and (11) and separately considering the cases where the event $E^{f_t}(t)$ is true or false, we have that

$$\begin{aligned}
\mathbb{P}\left(\mathbf{x}_t \in \mathcal{X} \setminus S_t|\mathcal{F}_{t-1}\right) &\geq \mathbb{P}\left(f_t(\mathbf{x}^*) > f_t(\mathbf{x}), \forall \mathbf{x} \in S_t|\mathcal{F}_{t-1}\right) \\
&\overset{(a)}{\geq} \mathbb{P}\left(f_t(\mathbf{x}^*) > f(\mathbf{x}^*)|\mathcal{F}_{t-1}\right) - \mathbb{P}\left(\overline{E^{f_t}(t)}|\mathcal{F}_{t-1}\right) \\
&\overset{(b)}{\geq} p - 1/t^2.
\end{aligned} \tag{11}$$

This completes the proof. □

We use $\sigma_{t-1}(\cdot)$ to represent the GP posterior standard deviation conditioned on all selected input queries from iterations 1 to $t-1$.

Lemma 5. *We have for all $t \geq 1$ and all $\mathbf{x} \in \mathcal{X}$ that*

$$\frac{\sigma_{\text{fb}[t]}(\mathbf{x})}{\sigma_{t-1}(\mathbf{x})} \leq e^C.$$

Proof. We use $\mathbf{y}_{1:\text{fb}[t]}$ to denote the output observations from iterations 1 to fb$[t]$, use $\mathbf{y}_{\text{fb}[t]+1:t-1}$ to represent the the output observations from iterations fb$[t]+1$ to $t-1$, and use $\mathbf{y}_A$ to denote the vector of observations at a set of inputs $A \subset \mathcal{X}$. We use $H(\cdot)$ to represent the entroy of a random variable.

To begin with, we establish the relationship between the following conditional information gain and the ratio of GP posterior standard deviations which we intend to upper-bound:

$$\begin{aligned}\mathbb{I}(f(\mathbf{x}); \mathbf{y}_{\text{fb}[t]+1:t-1}|\mathbf{y}_{1:\text{fb}[t]})) &= H(f(\mathbf{x})|\mathbf{y}_{1:\text{fb}[t]})) - H(f(\mathbf{x})|\mathbf{y}_{1:t-1})) \\ &= \frac{1}{2}\log(2\pi e\sigma^2_{\text{fb}[t]}(\mathbf{x})) - \frac{1}{2}\log(2\pi e\sigma^2_{t-1}(\mathbf{x})) \\ &= \log\frac{\sigma_{\text{fb}[t]}(\mathbf{x})}{\sigma_{t-1}(\mathbf{x})}.\end{aligned} \tag{12}$$

The first equality comes from the definition of conditional information gain and the second equality follows immediately from the entroy of Gaussian random variables. The equation above allows us to upper-bound the ratio of GP posterior standard deviations as follows:

$$\begin{aligned}\frac{\sigma_{\text{fb}[t]}(\mathbf{x})}{\sigma_{t-1}(\mathbf{x})} &= \exp(\mathbb{I}(f(\mathbf{x}); \mathbf{y}_{\text{fb}[t]+1:t-1}|\mathbf{y}_{1:\text{fb}[t]}))) \\ &\overset{(a)}{\leq} \exp(\mathbb{I}(f; \mathbf{y}_{\text{fb}[t]+1:t-1}|\mathbf{y}_{1:\text{fb}[t]}))) \\ &\overset{(b)}{\leq} \exp\big(\max_{A\subset\mathcal{X},|A|\leq B-1} \mathbb{I}(f; \mathbf{y}_A|\mathbf{y}_{1:\text{fb}[t]})\big) \overset{(c)}{\leq} e^C,\end{aligned} \tag{13}$$

in which (a) is because the information gain about f is larger than that of $f(\mathbf{x})$, (b) follow since the size of $\mathbf{y}_{\text{fb}[t]+1:t-1}$ is at most $B-1$ in our batch setting with a batch size of B, and (c) is a result of the definition of the constant C. This completes the proof. □

Next, we are ready to prove an upper bound on the expected instantaneous regret $r_t = f(\mathbf{x}^*) - f(\mathbf{x}_t)$.

Lemma 6. *For any $\mathcal{F}_{t-1}$, conditioned on the event $E^f(t)$, we have that*

$$\mathbb{E}\left[r_t|\mathcal{F}_{t-1}\right] \leq c_t e^C \left(1 + \frac{10}{p}\right) \mathbb{E}\left[\sigma_{t-1}(\mathbf{x}_t)|\mathcal{F}_{t-1}\right] + \frac{2B'}{t^2}.$$

Proof. To begin with, define

$$\overline{\mathbf{x}}_t \triangleq \arg\min_{\mathbf{x}\in\mathcal{X}\setminus S_t} \sigma_{\text{fb}[t]}(\mathbf{x}). \tag{14}$$

That is, $\overline{\mathbf{x}}_t$ is the unsaturated input with the smallest GP posterior standard deviation. Note that given a $\mathcal{F}_{t-1}$, $\overline{\mathbf{x}}_t$ is deterministic. Next, we have that

$$\begin{aligned}\mathbb{E}[\sigma_{\text{fb}[t]}(\mathbf{x}_t)|\mathcal{F}_{t-1}] &\geq \mathbb{E}\left[\sigma_{\text{fb}[t]}(\mathbf{x}_t)|\mathcal{F}_{t-1}, \mathbf{x}_t \in \mathcal{X}\setminus S_t\right] \mathbb{P}\left(\mathbf{x}_t \in \mathcal{X}\setminus S_t|\mathcal{F}_{t-1}\right) \\ &\geq \sigma_{\text{fb}[t]}(\overline{\mathbf{x}}_t)(p - 1/t^2),\end{aligned} \tag{15}$$

where the second inequality makes use of Lemma 4, which holds here because we have also conditioned on the event $E^f(t)$ in Lemma 4. Next, conditioned on both $E^f(t)$ and $E^{f_t}(t)$, we have that

$$\begin{aligned}r_t = f(\mathbf{x}^*) - f(\mathbf{x}_t) &= f(\mathbf{x}^*) - f(\overline{\mathbf{x}}_t) + f(\overline{\mathbf{x}}_t) - f(\mathbf{x}_t) \\ &\overset{(a)}{\leq} \Delta(\overline{\mathbf{x}}_t) + f_t(\overline{\mathbf{x}}_t) + c_t\sigma_{\text{fb}[t]}(\overline{\mathbf{x}}_t) - f_t(\mathbf{x}_t) + c_t\sigma_{\text{fb}[t]}(\mathbf{x}_t) \\ &\overset{(b)}{\leq} c_t\sigma_{\text{fb}[t]}(\overline{\mathbf{x}}_t) + c_t\sigma_{\text{fb}[t]}(\overline{\mathbf{x}}_t) + c_t\sigma_{\text{fb}[t]}(\mathbf{x}_t) + f_t(\overline{\mathbf{x}}_t) - f_t(\mathbf{x}_t) \\ &\overset{(c)}{\leq} c_t\left(2\sigma_{\text{fb}[t]}(\overline{\mathbf{x}}_t) + \sigma_{\text{fb}[t]}(\mathbf{x}_t)\right),\end{aligned} \tag{16}$$

in which (a) makes use of Lemma 1 and Lemma 2, (b) follows since $\overline{\mathbf{x}}_t$ is unsaturated, and (c) follows from the way in which $\mathbf{x}_t$ is selected: $\mathbf{x}_t = \arg\max_{\mathbf{x}\in\mathcal{X}} f_t(\mathbf{x})$. Next, the expected value of

r_t can be upper-bounded as follows:

$$\begin{aligned}
\mathbb{E}\left[r_t|\mathcal{F}_{t-1}\right] &\leq \mathbb{E}\left[c_t\left(2\sigma_{\text{fb}[t]}(\overline{\mathbf{x}}_t)+\sigma_{\text{fb}[t]}(\mathbf{x}_t)\right)|\mathcal{F}_{t-1}\right]+2B'\mathbb{P}\left(\overline{E^{f_t}(t)}|\mathcal{F}_{t-1}\right)\\
&\overset{(a)}{\leq} \mathbb{E}\left[c_t\left(\frac{2}{p-1/t^2}\sigma_{\text{fb}[t]}(\mathbf{x}_t)+\sigma_{\text{fb}[t]}(\mathbf{x}_t)\right)|\mathcal{F}_{t-1}\right]+\frac{2B'}{t^2}\\
&= \mathbb{E}\left[c_t\left(1+\frac{2}{p-1/t^2}\right)\sigma_{\text{fb}[t]}(\mathbf{x}_t)|\mathcal{F}_{t-1}\right]+\frac{2B'}{t^2}\\
&\overset{(b)}{\leq} c_t\left(1+\frac{2}{p-1/t^2}\right)\mathbb{E}\left[e^C\sigma_{t-1}(\mathbf{x}_t)|\mathcal{F}_{t-1}\right]+\frac{2B'}{t^2}\\
&\overset{(c)}{\leq} c_te^C\left(1+\frac{10}{p}\right)\mathbb{E}\left[\sigma_{t-1}(\mathbf{x}_t)|\mathcal{F}_{t-1}\right]+\frac{2B'}{t^2},
\end{aligned} \tag{17}$$

where (a) follows from equation (16) and (b) makes use of Lemma 5. (c) follows since $2/(p-1/t^2) \leq 10/p$, which holds because (i) $p-1/t^2<0$ for $t<5$, (ii) $2/(p-1/t^2)\leq 10/p$ for $t=5$, and (iii) $2/(p-1/t^2)$ is decreasing as t increases when $t\geq 5$.

□

Definition 2. *Define $Y_0=0$, and for all $t=1,\ldots,T$,*

$$\begin{aligned}
\overline{r}_t &= r_t\mathbb{I}\{E^f(t)\},\\
X_t &= \overline{r}_t - c_te^C\left(1+\frac{10}{p}\right)\sigma_{t-1}(\mathbf{x}_t)-\frac{2B'}{t^2}\\
Y_t &= \sum_{s=1}^{t}X_s.
\end{aligned}$$

Lemma 7. *Conditioned on the event $E^f(t)$, $(Y_t : t=0,\ldots,T)$ is a super-martingale with respect to the filtration $\mathcal{F}_t$.*

Proof.

$$\begin{aligned}
\mathbb{E}[Y_t-Y_{t-1}|\mathcal{F}_{t-1}] &= \mathbb{E}[X_t|\mathcal{F}_{t-1}]\\
&= \mathbb{E}[\overline{r}_t - c_te^C\left(1+\frac{10}{p}\right)\sigma_{t-1}(\mathbf{x}_t)-\frac{2B'}{t^2}|\mathcal{F}_{t-1}]\\
&= \mathbb{E}[\overline{r}_t|\mathcal{F}_{t-1}]-\left(c_te^C\left(1+\frac{10}{p}\right)\mathbb{E}[\sigma_{t-1}(\mathbf{x}_t)|\mathcal{F}_{t-1}]+\frac{2B'}{t^2}\right)\leq 0.
\end{aligned} \tag{18}$$

If the event $E^f(t)$ holds, then $\overline{r}_t=r_t$ and the inequality follows from Lemma 6. If $E^f(t)$ does not hold, $\overline{r}_t=0$ and the inequality holds trivially. □

Lastly, we can apply the Azuma-Hoeffding's inequality to the martingale $(Y_t : t=0,\ldots,T)$ to derive the upper bound on the cumulative regret R_T.

Lemma 8. *Define $C_1 \triangleq \frac{2}{\log(1+\sigma^{-2})}$. With probability of $\geq 1-\delta$, we have that*

$$R_T \leq c_Te^C\left(1+\frac{10}{p}\right)\sqrt{C_1T\gamma_T}+\frac{B'\pi^2}{3}+\left[4B'+c_Te^C\left(1+\frac{10}{p}\right)K_0\right]\sqrt{2T\log(2/\delta)}. \tag{19}$$

Proof. To begin with, note that

$$\begin{aligned}
|Y_t-Y_{t-1}| = |X_t| &\leq |\overline{r}_t| + c_te^C\left(1+\frac{10}{p}\right)\sigma_{t-1}(\mathbf{x}_t)+\frac{2B'}{t^2}\\
&\leq 2B'+c_te^C\left(1+\frac{10}{p}\right)K_0+2B'\\
&= 4B'+c_te^C\left(1+\frac{10}{p}\right)K_0,
\end{aligned} \tag{20}$$

where we have made use of our assumption that $\Theta(\mathbf{x}, \mathbf{x}') \leq K_0$ in the second inequality. Next, applying the Azuma-Hoeffding's inequality to $(Y_t : t = 0, \ldots, T)$ with a probability of $\delta/2$, we have with probability $\geq 1 - \delta/2$ that

$$\begin{aligned}
\sum_{t=1}^{T} \overline{r}_t &\leq \sum_{t=1}^{T} c_t e^C \Big(1 + \frac{10}{p}\Big)\sigma_{t-1}(\mathbf{x}_t) + \sum_{t=1}^{T} \frac{2B'}{t^2} + \sqrt{2\log(2/\delta)\sum_{t=1}^{T}\Big(4B' + c_t e^C\big(1 + \frac{10}{p}\big)K_0\Big)^2} \\
&\overset{(a)}{\leq} c_T e^C \Big(1 + \frac{10}{p}\Big)\sum_{t=1}^{T}\sigma_{t-1}(\mathbf{x}_t) + \frac{B'\pi^2}{3} + \left[4B' + c_T e^C\left(1 + \frac{10}{p}\right)K_0\right]\sqrt{2T\log(2/\delta)} \\
&\overset{(b)}{\leq} c_T e^C \Big(1 + \frac{10}{p}\Big)\sqrt{C_1 T\gamma_T} + \frac{B'\pi^2}{3} + \left[4B' + c_T e^C\left(1 + \frac{10}{p}\right)K_0\right]\sqrt{2T\log(2/\delta)}.
\end{aligned} \tag{21}$$

(a) follows since c_t is increasing in t, and (b) follows from the proof of Lemma 5.4 in the work of [57]. Next, note that $\overline{r}_t = r_t, \forall t \geq 1$ with probability of $\geq 1 - \delta/2$ according to Lemma 1. Therefore, the upper bound derived in the equation above is an upper bound on $R_T = \sum_{t=1}^{T} r_t$ (with probability of $\geq 1 - \delta$), and the proof is completed.

□

Note that $c_T = \mathcal{O}(\log^2 T)$. From Lemma 8, we have that

$$R_T = \mathcal{O}\left(e^C(\log^2 T)\sqrt{T}\left(1 + \sqrt{\gamma_T}\right)\right) = \widetilde{\mathcal{O}}\left(e^C\sqrt{T}\left(1 + \sqrt{\gamma_T}\right)\right). \tag{22}$$

D Proof of Theorem 2

In this section, the main technical challenge is to rigorously account for the mismatch between the kernel with which we assume the objective function f is sampled (i.e., the exact NTK Θ) and the kernel with which the acquisition function is sampled (i.e., the empirical NTK $\widetilde{\Theta}$). For ease of exposition, we use k and $\widetilde{k}$ (instead of Θ and $\widetilde{\Theta}$) to represent the exact and empirical NTK in the proof in this section. Similarly, we also use $\widetilde{\ }$ to indicate that a term is associated with the empirical NTK $\widetilde{k}$. For example, we use $\widetilde{\mu}_{\text{fb}[t]}(\cdot)$ and $\widetilde{\sigma}^2_{\text{fb}[t]}(\cdot)$ to represent the GP posterior mean and variance calculated using the empirical NTK $\widetilde{k}$.

For simplicity, we assume that the event in Proposition 1 holds throughout the entire proof, which happens with probability of $\geq 1 - \delta/4$. That is, the approximation error between exact and empirical NTKs is bounded:

$$|\widetilde{k}(\mathbf{x}, \mathbf{x}') - k(\mathbf{x}, \mathbf{x}')| = \Big|\langle \nabla_\theta f(\mathbf{x}, \widetilde{\theta}), \nabla_\theta f(\mathbf{x}', \widetilde{\theta})\rangle - \Theta(\mathbf{x}, \mathbf{x}')\Big| \leq (L+1)\varepsilon, \qquad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}. \tag{23}$$

To begin with, we use the following lemma to bound the difference between the GP posterior standard deviations calculated using the exact and empirical NTKs. Here, to simplify the derivations and results, we assume that $(L+1)\varepsilon \leq 1$ and $\sigma^2 \leq 1$. Note that these assumptions are not essential to the proof but are only used get cleaner expressions. Here, again for ease of expositions, we define $\hat{K}_0 \triangleq \max\{1, K_0\}$, and $\hat{K}_0^2 \triangleq \max\{1, K_0^2\}$.

Lemma 9. *We have* $\forall t \geq 1, \forall \mathbf{x} \in \mathcal{X}$ *that*

$$|\sigma_{fb[t]}(\mathbf{x}) - \widetilde{\sigma}_{fb[t]}(\mathbf{x})| \leq \sqrt{(L+1)\varepsilon\left(1 + \frac{4\hat{K}_0^2 t^2}{\sigma^4}\right)}.$$

Proof. Denote by $\boldsymbol{K}_t$ the $\text{fb}[t] \times \text{fb}[t]$-dimensional gram matrix of exact NTK covariance values calculated using all $\text{fb}[t]$ observations up to iteration $\text{fb}[t]$, and use $\widetilde{\boldsymbol{K}}_t$ to represent the corresponding

gram matrix calculated using the empirical NTK $\widetilde{k}$. If we define $A = (\boldsymbol{K}_t + \sigma^2)^{-1}$ or $A = (\widetilde{\boldsymbol{K}}_t + \sigma^2)^{-1}$, then for both values of A, we have that

$$\|A\|_2 = \sqrt{\max[\text{eig}(A^\top A)]} = \sqrt{\max[\text{eig}(A)^2]} \leq \frac{1}{\sigma^2}. \tag{24}$$

This allows us to derive the following equation:

$$\begin{aligned}\left\|(\boldsymbol{K}_t + \sigma^2)^{-1} - (\widetilde{\boldsymbol{K}}_t + \sigma^2)^{-1}\right\|_2 &\leq \left\|(\boldsymbol{K}_t + \sigma^2)^{-1}\right\|_2 \left\|(\widetilde{\boldsymbol{K}}_t + \sigma^2)^{-1}\right\|_2 \left\|\boldsymbol{K}_t - \widetilde{\boldsymbol{K}}_t\right\|_2 \\ &\leq \frac{1}{\sigma^2} \times \frac{1}{\sigma^2} \times t(L+1)\varepsilon = \frac{t(L+1)\varepsilon}{\sigma^4}.\end{aligned} \tag{25}$$

Define the $\text{fb}[t]$-dimensional vectors $\boldsymbol{k}_t(\mathbf{x}) = [k(\mathbf{x}, \mathbf{x}_\tau)]_{\tau=1,\ldots,\text{fb}[t]}$ and $\widetilde{\boldsymbol{k}}_t(\mathbf{x}) = [\widetilde{k}(\mathbf{x}, \mathbf{x}_\tau)]_{\tau=1,\ldots,\text{fb}[t]}$. Then making use of the approximation guarantee from equation (24), we define $\widetilde{\boldsymbol{k}}_t(\mathbf{x}) = \boldsymbol{k}_t(\mathbf{x}) + (L+1)\varepsilon\boldsymbol{\nu}(\mathbf{x})$, where $\boldsymbol{\nu}(\mathbf{x})$ is an $\text{fb}[t]-$dimensional vector where every element satisfies $|\boldsymbol{\nu}(\mathbf{x})_i| \leq 1, \forall i \in [\text{fb}[t]]$. Now we can use these definitions to derive the following upper bound.

$$\begin{aligned}
&|\sigma^2_{\text{fb}[t]}(\mathbf{x}) - \widetilde{\sigma}^2_{\text{fb}[t]}(\mathbf{x})| = |k(\mathbf{x},\mathbf{x}) - \boldsymbol{k}_t(\mathbf{x})^\top (\boldsymbol{K}_t + \sigma^2 I)^{-1}\boldsymbol{k}_t(\mathbf{x}) \\
&\qquad - \widetilde{k}(\mathbf{x},\mathbf{x}) + \widetilde{\boldsymbol{k}}_t(\mathbf{x})^\top (\widetilde{\boldsymbol{K}}_t + \sigma^2 I)^{-1}\widetilde{\boldsymbol{k}}_t(\mathbf{x})| \\
&\leq |k(\mathbf{x},\mathbf{x}) - \widetilde{k}(\mathbf{x},\mathbf{x})| + |\boldsymbol{k}_t(\mathbf{x})^\top (\boldsymbol{K}_t + \sigma^2 I)^{-1}\boldsymbol{k}_t(\mathbf{x}) - \widetilde{\boldsymbol{k}}_t(\mathbf{x})^\top (\widetilde{\boldsymbol{K}}_t + \sigma^2 I)^{-1}\widetilde{\boldsymbol{k}}_t(\mathbf{x})| \\
&\leq (L+1)\varepsilon + \Big|\boldsymbol{k}_t(\mathbf{x})^\top \left((\boldsymbol{K}_t + \sigma^2 I)^{-1} - (\widetilde{\boldsymbol{K}}_t + \sigma^2 I)^{-1}\right)\boldsymbol{k}_t(\mathbf{x}) \\
&- 2(L+1)\varepsilon\boldsymbol{\nu}(\mathbf{x})^\top (\widetilde{\boldsymbol{K}}_t + \sigma^2 I)^{-1}\boldsymbol{k}_t(\mathbf{x}) - (L+1)^2\varepsilon^2\boldsymbol{\nu}(\mathbf{x})^\top (\widetilde{\boldsymbol{K}}_t + \sigma^2 I)^{-1}\boldsymbol{\nu}(\mathbf{x})\Big| \\
&\leq (L+1)\varepsilon + \left\|\boldsymbol{k}_t(\mathbf{x})\right\|_2 \left\|(\boldsymbol{K}_t + \sigma^2)^{-1} - (\widetilde{\boldsymbol{K}}_t + \sigma^2)^{-1}\right\|_2 \left\|\boldsymbol{k}_t(\mathbf{x})\right\|_2 + \\
&2(L+1)\varepsilon\left\|\boldsymbol{\nu}(\mathbf{x})\right\|_2 \left\|(\widetilde{\boldsymbol{K}}_t + \sigma^2)^{-1}\right\|_2 \left\|\boldsymbol{k}_t(\mathbf{x})\right\|_2 + (L+1)^2\varepsilon^2 \left\|\boldsymbol{\nu}(\mathbf{x})\right\|_2 \left\|(\widetilde{\boldsymbol{K}}_t + \sigma^2 I)^{-1}\right\|_2 \left\|\boldsymbol{\nu}(\mathbf{x}\right\|_2) \\
&\leq (L+1)\varepsilon + K_0\sqrt{t}\frac{t(L+1)\varepsilon}{\sigma^4}K_0\sqrt{t} + 2(L+1)\varepsilon\sqrt{t}\frac{1}{\sigma^2}K_0\sqrt{t} + (L+1)^2\varepsilon^2\sqrt{t}\frac{1}{\sigma^2}\sqrt{t} \\
&= (L+1)\varepsilon + K_0^2\frac{t^2(L+1)\varepsilon}{\sigma^4} + 2K_0(L+1)\varepsilon\frac{t}{\sigma^2} + (L+1)^2\varepsilon^2\frac{t}{\sigma^2} \\
&\leq (L+1)\varepsilon + 4\hat{K}_0^2\frac{t^2(L+1)\varepsilon}{\sigma^4} \\
&\leq (L+1)\varepsilon\left(1 + \frac{4\hat{K}_0^2 t^2}{\sigma^4}\right).
\end{aligned} \tag{26}$$

Elementary calculation tells us that for $a, b, c > 0$, if $a^2 - b^2 \leq c^2$, then $a \leq \sqrt{b^2 + c^2} \leq b + c$, which leads to $a - b \leq c$. As a result, the equation above tells us that $|\sigma_{\text{fb}[t]}(\mathbf{x}) - \widetilde{\sigma}_{\text{fb}[t]}(\mathbf{x})| \leq \sqrt{(L+1)\varepsilon\left(1 + \frac{4\hat{K}_0^2 t^2}{\sigma^4}\right)}$. □

The next Lemma gives an upper bound on the difference between the GP posterior means calculated using the exact and empirical NTKs.

Lemma 10. *With probability of* $\geq 1 - \delta/4$*, we have* $\forall t \geq 1, \forall \mathbf{x} \in \mathcal{X}$ *that*

$$|\mu_{fb[t]}(\mathbf{x}) - \widetilde{\mu}_{fb[t]}(\mathbf{x})| \leq 2\hat{K}_0\frac{t^2(L+1)\varepsilon}{\sigma^4}\left(B' + \sigma\sqrt{2\log(4T/\delta)}\right).$$

Proof. Define $\mathbf{y}_t = [y_\tau]_{\tau=1,\ldots,\text{fb}[t]}$. We have that $y_\tau = f(\mathbf{x}_\tau) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Standard Gaussian concentration tells us that $|\epsilon| \leq z\sigma$ with probability of $\geq 1 - \exp(-z^2/2)$. Substituting $z = \sqrt{2\log(4T/\delta)}$ and making use of the assumption that $|f(\mathbf{x})| \leq B', \forall \mathbf{x} \in \mathcal{X}$, we have that $|y_\tau| \leq B' + \sigma\sqrt{2\log(4T/\delta)}$ with probability of $\geq 1 - \delta/(4T)$. Now taking a union bound over all T

iterations, we have that $|y_\tau| \leq B' + \sigma\sqrt{2\log(4T/\delta)}, \forall \tau = 1, \ldots, T$ with probability of $\geq 1 - \delta/4$. This further implies that $\|\mathbf{y}_t\|_2 = \sqrt{\sum_{\tau=1}^{\text{fb}[t]} y_\tau^2} \leq \sqrt{t}\left(B' + \sigma\sqrt{2\log(4T/\delta)}\right)$. Now we are ready to bound the term in question:

$$
\begin{aligned}
&|\mu_{\text{fb}[t]}(\mathbf{x}) - \widetilde{\mu}_{\text{fb}[t]}(\mathbf{x})| = |\boldsymbol{k}_t(\mathbf{x})^\top (K_t + \sigma^2 I)^{-1}\mathbf{y}_t - \widetilde{\boldsymbol{k}}_t(\mathbf{x})^\top (\widetilde{K}_t + \sigma^2 I)^{-1}\mathbf{y}_t| \\
&= |\boldsymbol{k}_t(\mathbf{x})^\top (K_t + \sigma^2 I)^{-1}\mathbf{y}_t - \boldsymbol{k}_t(\mathbf{x})^\top (\widetilde{K}_t + \sigma^2 I)^{-1}\mathbf{y}_t - (L+1)\varepsilon \boldsymbol{\nu}(\mathbf{x})^\top (\widetilde{K}_t + \sigma^2 I)^{-1}\mathbf{y}_t| \\
&\leq \left\|\boldsymbol{k}_t(\mathbf{x})\right\|_2 \left\|(K_t + \sigma^2 I)^{-1} - (\widetilde{K}_t + \sigma^2 I)^{-1}\right\| \|\mathbf{y}_t\|_2 + (L+1)\varepsilon \left\|\boldsymbol{\nu}(\mathbf{x})\right\|_2 \left\|(\widetilde{K}_t + \sigma^2 I)^{-1}\right\|_2 \|\mathbf{y}_t\|_2 \\
&\leq K_0\sqrt{t}\frac{t(L+1)\varepsilon}{\sigma^4}\sqrt{t}\left(B' + \sigma\sqrt{2\log(4T/\delta)}\right) + (L+1)\varepsilon\sqrt{t}\frac{1}{\sigma^2}\sqrt{t}\left(B' + \sigma\sqrt{2\log(4T/\delta)}\right) \\
&\leq 2\hat{K}_0\frac{t^2(L+1)\varepsilon}{\sigma^4}\left(B' + \sigma\sqrt{2\log(4T/\delta)}\right).
\end{aligned}
\tag{27}
$$

□

Next, similar to the proof in Appendix C, here we also need a lemma showing the concentration of the function f. Define $\beta_t = 2\log(2\pi^2 t^2|\mathcal{X}|/(3\delta))$, and $c_t = \beta_t(1 + \sqrt{2\log(|\mathcal{X}|t^2)})$. Note that the value of β_t defined here is slightly different due to the use of different error probabilities (i.e., we have used an error probability of $\delta/2$ in the proof in Appendix C yet $\delta/4$ in this section).

Lemma 11. *$|\mu_{fb[t]}(\mathbf{x}) - f(\mathbf{x})| \leq \beta_t \sigma_{fb[t]}(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}$, with probability of $\geq 1 - \delta/4, \forall t \geq 1$.*

The proof of Lemma 11 is the same as that of Lemma 1. The next Lemma proves the concentration of the objective function f around the GP posterior mean calculated using the empirical NTK $\widetilde{k}$, which consists of an additional error term $\epsilon_{m,t}$ due to the use of the empirical NTK compared with Lemma 11 above.

Lemma 12. *Define*

$$
\epsilon_{m,t} \triangleq 2\hat{K}_0\frac{t^2(L+1)\varepsilon}{\sigma^4}\left(B' + \sigma\sqrt{2\log(4T/\delta)}\right) + \beta_t\sqrt{(L+1)\varepsilon\left(1 + \frac{4\hat{K}_0^2 t^2}{\sigma^4}\right)}.
$$

Define $E^{\widetilde{f}}(t)$ as the event that $|\widetilde{\mu}_{fb[t]}(\mathbf{x}) - f(\mathbf{x})| \leq \beta_t \widetilde{\sigma}_{fb[t]}(\mathbf{x}) + \epsilon_{m,t}, \forall \mathbf{x} \in \mathcal{X}$. We have that $\mathbb{P}(E^{\widetilde{f}}(t)) \geq 1 - \delta/2, \forall t \geq 1$.

Proof.

$$
\begin{aligned}
&|\widetilde{\mu}_{\text{fb}[t]}(\mathbf{x}) - f(\mathbf{x})| \leq |\widetilde{\mu}_{\text{fb}[t]}(\mathbf{x}) - \mu_{\text{fb}[t]}(\mathbf{x})| + |\mu_{\text{fb}[t]}(\mathbf{x}) - f(\mathbf{x})| \\
&\overset{(a)}{\leq} 2\hat{K}_0\frac{t^2(L+1)\varepsilon}{\sigma^4}\left(B' + \sigma\sqrt{2\log(4T/\delta)}\right) + \beta_t\sigma_{\text{fb}[t]}(\mathbf{x}) \\
&\overset{(b)}{\leq} 2\hat{K}_0\frac{t^2(L+1)\varepsilon}{\sigma^4}\left(B' + \sigma\sqrt{2\log(4T/\delta)}\right) + \beta_t\left(\widetilde{\sigma}_{\text{fb}[t]}(\mathbf{x}) + \sqrt{(L+1)\varepsilon\left(1 + \frac{4\hat{K}_0^2 t^2}{\sigma^4}\right)}\right) \\
&= \beta_t\widetilde{\sigma}_{\text{fb}[t]}(\mathbf{x}) + 2\hat{K}_0\frac{t^2(L+1)\varepsilon}{\sigma^4}\left(B' + \sigma\sqrt{2\log(4T/\delta)}\right) + \beta_t\sqrt{(L+1)\varepsilon\left(1 + \frac{4\hat{K}_0^2 t^2}{\sigma^4}\right)} \\
&= \beta_t\widetilde{\sigma}_{\text{fb}[t]}(\mathbf{x}) + \epsilon_{m,t}
\end{aligned}
\tag{28}
$$

(a) follows from Lemma 10 and Lemma 11 and hence holds with probability of $\geq 1 - \delta/4 - \delta/4 = 1 - \delta/2$, and (b) results from Lemma 9. □

Denote by $\widetilde{f}_t$ the sampled function in iteration t using the empirical NTK, i.e., $\widetilde{f}_t \sim \mathcal{GP}(\widetilde{\mu}_{\text{fb}[t]}(\cdot), \beta_t^2\widetilde{\sigma}^2_{\text{fb}[t]}(\cdot,\cdot))$.

Lemma 13. *Define $E^{\widetilde{f}_t}(t)$ as the event that $|\widetilde{\mu}_{fb[t]}(\mathbf{x}) - \widetilde{f}_t(\mathbf{x})| \leq \beta_t \sqrt{2\log(|\mathcal{X}|t^2)}\widetilde{\sigma}_{fb[t]}(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}$. We have that $\mathbb{P}(E^{\widetilde{f}_t}(t)) \geq 1 - 1/t^2, \forall t \geq 1$.*

Lemma 13 is the counterpart to Lemma 2 in Appendix C and can be proved using the same techniques. Of note, conditioned on both events $E^{\widetilde{f}}(t)$ and $E^{\widetilde{f}_t}(t)$, we have that

$$\begin{aligned} |f(\mathbf{x}) - \widetilde{f}_t(\mathbf{x})| &\leq |f(\mathbf{x}) - \widetilde{\mu}_{\text{fb}[t]}(\mathbf{x})| + |\widetilde{\mu}_{\text{fb}[t]}(\mathbf{x}) - \widetilde{f}_t(\mathbf{x})| \\ &\leq \beta_t \widetilde{\sigma}_{\text{fb}[t]}(\mathbf{x}) + \epsilon_{m,t} + \beta_t \sqrt{2\log(|\mathcal{X}|t^2)}\widetilde{\sigma}_{\text{fb}[t]}(\mathbf{x}) \\ &= c_t \widetilde{\sigma}_{\text{fb}[t]}(\mathbf{x}) + \epsilon_{m,t}. \end{aligned} \tag{29}$$

Next, we similarly define the set of saturated inputs.

Definition 3. *Define the set of saturated inputs in iteration t as*

$$S_t = \{\mathbf{x} \in \mathcal{X} : \Delta(\mathbf{x}) > c_t \widetilde{\sigma}_{fb[t]}(\mathbf{x}) + 2\epsilon_{m,t}\},$$

in which $\Delta(\mathbf{x}) = f(\mathbf{x}^*) - f(\mathbf{x})$.

Again, $\mathbf{x}^*$ is always unsaturated.

Lemma 14. *For any $\mathcal{F}_{t-1}$, conditioned on the events $E^{\widetilde{f}}(t)$, we have that $\forall \mathbf{x} \in \mathcal{X}$,*

$$\mathbb{P}\left(\widetilde{f}_t(\mathbf{x}) + \epsilon_{m,t} > f(\mathbf{x}) | \mathcal{F}_{t-1}\right) \geq p, \tag{30}$$

in which $p = \frac{1}{4e\sqrt{\pi}}$.

Proof.

$$\begin{aligned} \mathbb{P}\left(\widetilde{f}_t(\mathbf{x}) + \epsilon_{m.t} > f(\mathbf{x}) | \mathcal{F}_{t-1}\right) &= \mathbb{P}\left(\frac{\widetilde{f}_t(\mathbf{x}) - \widetilde{\mu}_{\text{fb}[t]}(\mathbf{x}) + \epsilon_{m.t}}{\beta_t \widetilde{\sigma}_{\text{fb}[t]}(\mathbf{x})} > \frac{f(\mathbf{x}) - \widetilde{\mu}_{\text{fb}[t]}(\mathbf{x})}{\beta_t \widetilde{\sigma}_{\text{fb}[t]}(\mathbf{x})} \Big| \mathcal{F}_{t-1}\right) \\ &\geq \mathbb{P}\left(\frac{\widetilde{f}_t(\mathbf{x}) - \widetilde{\mu}_{\text{fb}[t]}(\mathbf{x}) + \epsilon_{m.t}}{\beta_t \widetilde{\sigma}_{\text{fb}[t]}(\mathbf{x})} > \frac{|f(\mathbf{x}) - \widetilde{\mu}_{\text{fb}[t]}(\mathbf{x})|}{\beta_t \widetilde{\sigma}_{\text{fb}[t]}(\mathbf{x})} \Big| \mathcal{F}_{t-1}\right) \\ &= \mathbb{P}\left(\frac{\widetilde{f}_t(\mathbf{x}) - \widetilde{\mu}_{\text{fb}[t]}(\mathbf{x})}{\beta_t \widetilde{\sigma}_{\text{fb}[t]}(\mathbf{x})} > \frac{|f(\mathbf{x}) - \widetilde{\mu}_{\text{fb}[t]}(\mathbf{x})| - \epsilon_{m.t}}{\beta_t \widetilde{\sigma}_{\text{fb}[t]}(\mathbf{x})} \Big| \mathcal{F}_{t-1}\right) \\ &\overset{(a)}{\geq} \mathbb{P}\left(\frac{\widetilde{f}_t(\mathbf{x}) - \widetilde{\mu}_{\text{fb}[t]}(\mathbf{x})}{\beta_t \widetilde{\sigma}_{\text{fb}[t]}(\mathbf{x})} > 1 \Big| \mathcal{F}_{t-1}\right) \\ &\overset{(b)}{\geq} \frac{\exp(-1)}{4\sqrt{\pi}}. \end{aligned} \tag{31}$$

(a) follows from Lemma 12, and (b) follows because $\widetilde{f}_t(\mathbf{x}) \sim \mathcal{N}(\widetilde{\mu}_{\text{fb}[t]}(\mathbf{x}), \beta_t^2 \widetilde{\sigma}^2_{\text{fb}[t]}(\mathbf{x}))$ and makes use of the Gaussian anti-concentration inequality.

□

Next, we again prove a lower bound on the probability that the selected input is unsaturated.

Lemma 15. *For any $\mathcal{F}_{t-1}$, conditioned on the event $E^{\widetilde{f}}(t)$, we have that*

$$\mathbb{P}\left(\mathbf{x}_t \in \mathcal{X} \setminus S_t, | \mathcal{F}_{t-1}\right) \geq p - 1/t^2.$$

Proof. The proof here follows similar steps as the proof of Lemma 4. To begin with, we have the following relationship.

$$\mathbb{P}\left(\mathbf{x}_t \in \mathcal{X} \setminus S_t | \mathcal{F}_{t-1}\right) \geq \mathbb{P}\left(\widetilde{f}_t(\mathbf{x}^*) > \widetilde{f}_t(\mathbf{x}), \forall \mathbf{x} \in S_t | \mathcal{F}_{t-1}\right), \tag{32}$$

The validity of this equation can be justified in a similar way as equation (9) in the proof of Lemma 4, i.e., the event on the right hand sight implies the event on the left hand side.

Next, we assume that both events $E^{\widetilde{f}}(t)$ and $E^{\widetilde{f}}(t)$ are true, which allows us to derive an upper bound on $\widetilde{f}_t(\mathbf{x})$ for all $\mathbf{x} \in S_t$:

$$\widetilde{f}_t(\mathbf{x}) \overset{(a)}{\leq} f(\mathbf{x}) + c_t\widetilde{\sigma}_{\text{fb}[t]}(\mathbf{x}) + \epsilon_{m,t} \overset{(b)}{\leq} f(\mathbf{x}) + \Delta(\mathbf{x}) - \epsilon_{m,t} = f(\mathbf{x}^*) - \epsilon_{m,t}, \tag{33}$$

in which (a) follows from Lemmas 12 and 13, and (b) is a result of Definition 3.

Therefore, (34) implies that when both both events $E^{\widetilde{f}}(t)$ and $E^{\widetilde{f}_t}(t)$ hold,

$$\mathbb{P}\left(\widetilde{f}_t(\mathbf{x}^*) > \widetilde{f}_t(\mathbf{x}), \forall \mathbf{x} \in S_t | \mathcal{F}_{t-1}\right) \geq \mathbb{P}\left(\widetilde{f}_t(\mathbf{x}^*) > f(\mathbf{x}^*) - \epsilon_{m,t} | \mathcal{F}_{t-1}\right). \tag{34}$$

Next, we can show that

$$\begin{aligned}
\mathbb{P}\left(\mathbf{x}_t \in \mathcal{X} \setminus S_t | \mathcal{F}_{t-1}\right) &\geq \mathbb{P}\left(\widetilde{f}_t(\mathbf{x}^*) > \widetilde{f}_t(\mathbf{x}), \forall \mathbf{x} \in S_t | \mathcal{F}_{t-1}\right) \\
&\geq \mathbb{P}\left(\widetilde{f}_t(\mathbf{x}^*) > f(\mathbf{x}^*) - \epsilon_{m,t} | \mathcal{F}_{t-1}\right) - \mathbb{P}\left(\overline{E^{\widetilde{f}_t}(t)} | \mathcal{F}_{t-1}\right) \\
&\geq p - 1/t^2,
\end{aligned} \tag{35}$$

where the last inequality makes use of Lemma 14. □

The next Lemma derives an upper bound on the expected instantaneous regret $r_t = f(\mathbf{x}^*) - f(\mathbf{x}_t)$.

Lemma 16. *For any $\mathcal{F}_{t-1}$, conditioned on the event $E^{\widetilde{f}}(t)$, we have that*

$$\mathbb{E}\left[r_t | \mathcal{F}_{t-1}\right] \leq c_t e^C \left(1 + \frac{10}{p}\right) \mathbb{E}\left[\sigma_{t-1}(\mathbf{x}_t) | \mathcal{F}_{t-1}\right] + \epsilon'_{m,t} + \frac{2B'}{t^2},$$

where

$$\epsilon'_{m,t} \triangleq 3c_t \sqrt{(L+1)\varepsilon \left(1 + \frac{4\hat{K}_0^2 t^2}{\sigma^4}\right)} + 4\epsilon_{m,t}. \tag{36}$$

Proof. Define

$$\overline{\mathbf{x}}_t \triangleq \arg\min_{\mathbf{x} \in \mathcal{X} \setminus S_t} \sigma_{\text{fb}[t]}(\mathbf{x}). \tag{37}$$

Note that given a $\mathcal{F}_{t-1}$, $\overline{\mathbf{x}}_t$ is deterministic. Next, this definiton also leads to:

$$\begin{aligned}
\mathbb{E}[\sigma_{\text{fb}[t]}(\mathbf{x}_t) | \mathcal{F}_{t-1}] &\geq \mathbb{E}\left[\sigma_{\text{fb}[t]}(\mathbf{x}_t) | \mathcal{F}_{t-1}, \mathbf{x}_t \in \mathcal{X} \setminus S_t\right] \mathbb{P}\left(\mathbf{x}_t \in \mathcal{X} \setminus S_t | \mathcal{F}_{t-1}\right) \\
&\geq \sigma_{\text{fb}[t]}(\overline{\mathbf{x}}_t)(p - 1/t^2),
\end{aligned} \tag{38}$$

where the last inequality makes use of Lemma 15.

Next, conditioned on both $E^{\widetilde{f}}(t)$ and $E^{\widetilde{f}_t}(t)$, we have that

$$\begin{aligned}
r_t &= f(\mathbf{x}^*) - f(\mathbf{x}_t) = f(\mathbf{x}^*) - f(\overline{\mathbf{x}}_t) + f(\overline{\mathbf{x}}_t) - f(\mathbf{x}_t) \\
&\overset{(a)}{\leq} \Delta(\overline{\mathbf{x}}_t) + \widetilde{f}_t(\overline{\mathbf{x}}_t) + c_t\widetilde{\sigma}_{\text{fb}[t]}(\overline{\mathbf{x}}_t) + \epsilon_{m,t} - \widetilde{f}_t(\mathbf{x}_t) + c_t\widetilde{\sigma}_{\text{fb}[t]}(\mathbf{x}_t) + \epsilon_{m,t} \\
&\overset{(b)}{\leq} c_t\widetilde{\sigma}_{\text{fb}[t]}(\overline{\mathbf{x}}_t) + 2\epsilon_{m,t} + c_t\widetilde{\sigma}_{\text{fb}[t]}(\overline{\mathbf{x}}_t) + c_t\widetilde{\sigma}_{\text{fb}[t]}(\mathbf{x}_t) + 2\epsilon_{m,t} + \widetilde{f}_t(\overline{\mathbf{x}}_t) - \widetilde{f}_t(\mathbf{x}_t) \\
&\overset{(c)}{\leq} c_t\left(2\widetilde{\sigma}_{\text{fb}[t]}(\overline{\mathbf{x}}_t) + \widetilde{\sigma}_{\text{fb}[t]}(\mathbf{x}_t)\right) + 4\epsilon_{m,t} \\
&\overset{(d)}{\leq} c_t\left(2\sigma_{\text{fb}[t]}(\overline{\mathbf{x}}_t) + \sigma_{\text{fb}[t]}(\mathbf{x}_t)\right) + 3c_t\sqrt{(L+1)\varepsilon\left(1 + \frac{4\hat{K}_0^2 t^2}{\sigma^4}\right)} + 4\epsilon_{m,t} \\
&= c_t\left(2\sigma_{\text{fb}[t]}(\overline{\mathbf{x}}_t) + \sigma_{\text{fb}[t]}(\mathbf{x}_t)\right) + \epsilon'_{m,t}.
\end{aligned} \tag{39}$$

(a) follows from Lemmas 12 and 13, (b) follows from the definition of unsaturated inputs (Definition 3), (c) results from the way in which $\mathbf{x}_t$ is selected: $\mathbf{x}_t = \arg\max_{\mathbf{x}\in\mathcal{X}} \widetilde{f}_t(\mathbf{x})$, (d) makes use of Lemma 9.

Next, we can upper-bound the expected instantaneous regret:

$$\begin{aligned}
\mathbb{E}\left[r_t|\mathcal{F}_{t-1}\right] &\leq \mathbb{E}\left[c_t\left(2\sigma_{\text{fb}[t]}(\overline{\mathbf{x}}_t)+\sigma_{\text{fb}[t]}(\mathbf{x}_t)\right)+\epsilon'_{m,t}|\mathcal{F}_{t-1}\right]+2B'\mathbb{P}\left(\overline{E^{\widetilde{f}_t}(t)}|\mathcal{F}_{t-1}\right)\\
&\overset{(a)}{\leq} \mathbb{E}\left[c_t\left(\frac{2}{p-1/t^2}\sigma_{\text{fb}[t]}(\mathbf{x}_t)+\sigma_{\text{fb}[t]}(\mathbf{x}_t)\right)+\epsilon'_{m,t}|\mathcal{F}_{t-1}\right]+\frac{2B'}{t^2}\\
&= \mathbb{E}\left[c_t\left(1+\frac{2}{p-1/t^2}\right)\sigma_{\text{fb}[t]}(\mathbf{x}_t)+\epsilon'_{m,t}|\mathcal{F}_{t-1}\right]+\frac{2B'}{t^2} \qquad (40)\\
&\overset{(b)}{\leq} c_t\left(1+\frac{2}{p-1/t^2}\right)\mathbb{E}\left[e^C\sigma_{t-1}(\mathbf{x}_t)|\mathcal{F}_{t-1}\right]+\epsilon'_{m,t}+\frac{2B'}{t^2}\\
&\overset{(c)}{\leq} c_te^C\left(1+\frac{10}{p}\right)\mathbb{E}\left[\sigma_{t-1}(\mathbf{x}_t)|\mathcal{F}_{t-1}\right]+\epsilon'_{m,t}+\frac{2B'}{t^2}.
\end{aligned}$$

(a) follows from equation (39), (b) makes use of Lemma 5, and (c) follows since $2/(p-1/t^2) \leq 10/p$. This completes the proof. □

We similarly define the following stochastic process, which will be shown to be a super-martingale in the subsequent Lemma.

Definition 4. *Define $Y_0 = 0$, and for all $t = 1, \ldots, T$,*

$$\overline{r}_t = r_t\mathbb{I}\{E^{\widetilde{f}}(t)\},$$

$$X_t = \overline{r}_t - c_te^C\left(1+\frac{10}{p}\right)\sigma_{t-1}(\mathbf{x}_t) - \epsilon'_{m,t} - \frac{2B'}{t^2}$$

$$Y_t = \sum_{s=1}^{t} X_s.$$

Lemma 17. *Conditioned on the event $E^f(t)$, $(Y_t : t = 0, \ldots, T)$ is a super-martingale with respect to the filtration $\mathcal{F}_t$.*

The proof of Lemma 17 above follows closely the proof of Lemma 7 and is hence omitted.

Lemma 18. *Define $C_1 \triangleq \frac{2}{\log(1+\sigma^{-2})}$. With probability of $\geq 1-\delta$,*

$$R_T \leq c_Te^C\Big(1+\frac{10}{p}\Big)\sqrt{C_1T\gamma_T}+T\epsilon'_{m,T}+\frac{B'\pi^2}{3}+\Big(4B'+c_Te^C\big(1+\frac{10}{p}\big)K_0+\epsilon'_{m,T}\Big)\sqrt{2T\log(4/\delta)}. \qquad (41)$$

Proof. To begin with, we have that

$$\begin{aligned}
|Y_t - Y_{t-1}| = |X_t| &\leq |\overline{r}_t| + c_te^C\left(1+\frac{10}{p}\right)\sigma_{t-1}(\mathbf{x}_t)+\epsilon'_{m,t}+\frac{2B'}{t^2}\\
&\leq 2B' + c_te^C\left(1+\frac{10}{p}\right)K_0+\epsilon'_{m,t}+2B' \qquad (42)\\
&= 4B' + c_te^C\left(1+\frac{10}{p}\right)K_0+\epsilon'_{m,t}.
\end{aligned}$$

Using the Azuma-Hoeffding's inequality with an error probability of $\delta/4$, we have that

$$\begin{aligned}
\sum_{t=1}^{T} \overline{r}_t &\leq \sum_{t=1}^{T} c_t e^C \left(1+\frac{10}{p}\right) \sigma_{t-1}(\mathbf{x}_t) + \sum_{t=1}^{T} \epsilon'_{m,t} + \sum_{t=1}^{T} \frac{2B'}{t^2} + \\
&\qquad \sqrt{2\log(4/\delta) \sum_{t=1}^{T} \left(4B' + c_t e^C \left(1+\frac{10}{p}\right) K_0 + \epsilon'_{m,t}\right)^2} \\
&\overset{(a)}{\leq} c_T e^C \left(1+\frac{10}{p}\right) \sum_{t=1}^{T} \sigma_{t-1}(\mathbf{x}_t) + \sum_{t=1}^{T} \epsilon'_{m,t} + \frac{B'\pi^2}{3} + \\
&\qquad \left(4B' + c_T e^C \left(1+\frac{10}{p}\right) K_0 + \epsilon'_{m,T}\right) \sqrt{2T\log(4/\delta)} \\
&\overset{(b)}{\leq} c_T e^C \left(1+\frac{10}{p}\right) \sqrt{C_1 T \gamma_T} + \sum_{t=1}^{T} \epsilon'_{m,t} + \frac{B'\pi^2}{3} + \\
&\qquad \left(4B' + c_T e^C \left(1+\frac{10}{p}\right) K_0 + \epsilon'_{m,T}\right) \sqrt{2T\log(4/\delta)} \\
&\leq c_T e^C \left(1+\frac{10}{p}\right) \sqrt{C_1 T \gamma_T} + T\epsilon'_{m,T} + \frac{B'\pi^2}{3} + \\
&\qquad \left(4B' + c_T e^C \left(1+\frac{10}{p}\right) K_0 + \epsilon'_{m,T}\right) \sqrt{2T\log(4/\delta)}.
\end{aligned} \tag{43}$$

(a) follows since c_t is increasing in t, and (b) follows from the proof of Lemma 5.4 in the work of [57]. Next, note that $\overline{r}_t = r_t, \forall t \geq 1$ with probability of $\geq 1-\delta/2$ according to Lemma 12. Also recall that throughout the entire proof in this section, we have conditioned on the event in Proposition 1, which also holds with probability of $\geq 1-\delta/4$. Therefore, also taking into account the error probability of $\delta/4$ from the Azuma-Hoeffding's inequality, the upper bound derived above is an upper bound on the cumulative regret $R_T = \sum_{t=1}^{T} r_t$ with probability of $\geq 1-\delta/2-\delta/4-\delta/4 = 1-\delta$. □

Now let's analyze the asymptotic scaling of the regret upper bound derived above. Firstly, note that $c_T = \mathcal{O}(\log^2 T)$. Next, recall that we have in the main text that $(L+1)\varepsilon = C_{\text{ntk}}(L+1)L^{3/2}\log^{1/4}(4L|\mathcal{X}|^2/\delta)m^{-1/4}$. This allows us to analyze the scaling of $\epsilon'_{m,T}$.

$$\begin{aligned}
\epsilon'_{m,T} &= 3c_T \sqrt{(L+1)\varepsilon \left(1+\frac{4\hat{K}_0^2 T^2}{\sigma^4}\right)} + 4\Big(2\hat{K}_0 \frac{T^2(L+1)\varepsilon}{\sigma^4} \left(B' + \sigma\sqrt{2\log(4T/\delta)}\right) + \\
&\qquad \beta_T \sqrt{(L+1)\varepsilon \left(1+\frac{4\hat{K}_0^2 T^2}{\sigma^4}\right)}\Big) \\
&= \widetilde{\mathcal{O}}\left((\log T)^2 \sqrt{(L+1)\varepsilon} T + T^2(L+1)\varepsilon + \log T \sqrt{(L+1)\varepsilon} T\right) \\
&= \widetilde{\mathcal{O}}\left(T^2\sqrt{(L+1)\varepsilon}\right) \\
&= \widetilde{\mathcal{O}}\left(T^2 m^{-1/8}(L+1)^{5/4}\right)
\end{aligned} \tag{44}$$

This allows us to analyze the asymptotic scaling of our regret upper bound (ignoring all log factors)

$$\begin{aligned}
R_T &= \widetilde{\mathcal{O}}\left(e^C\sqrt{T\gamma_T} + T\epsilon'_{m,T} + (e^C + \epsilon'_{m,T})\sqrt{T}\right) \\
&= \widetilde{\mathcal{O}}\left(e^C\sqrt{T}(\sqrt{\gamma_T}+1) + T\epsilon'_{m,T} + \sqrt{T}\epsilon'_{m,T}\right) \\
&= \widetilde{\mathcal{O}}\left(e^C\sqrt{T}(\sqrt{\gamma_T}+1) + T^3 m^{-1/8}(L+1)^{5/4}\right).
\end{aligned} \tag{45}$$

E Extension to Continuous Input Domains

To extend our theoretical results to cases where the input domain $\mathcal{X}$ is continuous, we can follow the techniques discussed in Section 3.1 of the work of [39]. We assume that $\mathcal{X} \subset [0,1]^d$. To begin with, we need to additionally assume that the objective function f is Lipschitz continuous with a Lipschitz constant $L > 0$. Next, we can construct a finite sub-domain $\widetilde{\mathcal{X}}$ of the continuous domain $\mathcal{X}$, where $\widetilde{X}$ has equal spacing of $\frac{1}{\sqrt{T}}$ in each dimension. As a result, the finite sub-domain $\widetilde{\mathcal{X}}$ contains $T^{d/2}$ points, i.e., $|\widetilde{\mathcal{X}}| = T^{d/2}$. Then, we can simply run our algorithms (Algo. 1 and Algo. 2) on this finite sub-domain $\widetilde{X}$.

As a consequence, for **Theorem 1**, we only need to make two changes to our theoretical results. Firstly, we need to modify β_t to be $\beta_t = 2\log(\pi^2 t^2 |\widetilde{\mathcal{X}}|/(3\delta)) = 2\log(\pi^2 t^2 T^{d/2}/(3\delta))$, which will only introduce an additional dependence on $\mathcal{O}(d\log T)$ into c_T (Appendix C) and hence *an additional multiplicative factor of* $\mathcal{O}(d\log T) = \widetilde{\mathcal{O}}(d)$ into the regret upper bound in Theorem 1. Secondly, due to the Lipschitz continuity of f and the fact that every input $\mathbf{x} \in \mathcal{X}$ has a neighbor in the finite sub-domain $\widetilde{\mathcal{X}}$ whose distance to it is less than $\frac{d}{\sqrt{T}}$, we have that $f(\mathbf{x}^*) \leq \max_{\widetilde{\mathbf{x}} \in \widetilde{\mathcal{X}}} f(\widetilde{\mathbf{x}}) + \mathcal{O}(\frac{Ld}{\sqrt{T}})$. As a result, this will introduce *an additional additive term of* $\mathcal{O}(T \times \frac{Ld}{\sqrt{T}}) = \mathcal{O}(Ld\sqrt{T})$ to the final upper bound on the cumulative regret.

For **Proposition 1**, an additional multiplicative factor of $d\log T$ will be introduced into the condition on m. For **Theorem 2**, to begin with, same as the analysis of Theorem 1 in the paragraph above, *an additional additive factor of* $\mathcal{O}(Ld\sqrt{T})$ will be introduced, and *an additional multiplicative factor of* $\mathcal{O}(d\log T) = \widetilde{\mathcal{O}}(d)$ will be introduced into the first term in Theorem 2. Moreover, as a result of the additional factor of $d\log T$ in the condition on m (Proposition 1), an additional multiplicative factor of $d\log T$ will also be introduced into the approximation quality of $(L+1)\varepsilon$ (Sec. 4.2). As a result, in the proof of Theorem 2 (Appendix D), *an additional multiplicative factor of* $\sqrt{d}$ will be introduced into the term $\epsilon'_{m,T}$ (see (45)) and hence into the second term in the upper bound in Theorem 2.

Of note, the modified results discussed above do not affect the scaling of our theoretical results (Theorem 1 and Theorem 2) in T (we ignore all dependencies on $\log T$), because the only additional term depending on T for both theorems is an additive term of $\widetilde{\mathcal{O}}(\sqrt{T})$.

F More Experimental Details

In all experiments, for simplicity, we set $\beta_t = 1, \forall t \geq 1$, which is consistent with many previous papers on BO which have found the theoretical values of β_t to be overly conservative [57]. For fair comparisons, in every experiment, all methods under comparison use the same set of initial inputs which are selected by random search. We use the ERF activation function in the synthetic experiment (Sec. 5.1) because the synthetic function we have adopted is very smooth. In all real-world experiments (Secs. 5.2, 5.3 and 5.4), we use the ReLU activation function since it has been found to be very effective in modeling complicated real-world functions.

For all methods under comparison, when maximizing the acquisition function to choose an input query, if the domain is discrete, we simply evaluate the acquisition function value at every input in the domain and then choose the input that maximizes it. When the domain is continuous, we firstly use random search to randomly sample $10,000$ inputs in the domain to evaluate their acquisition function values, and then use L-BFGS-B with 100 random restarts to refine the search. When the domain is mixed (i.e., consisting of both continuous and discrete inputs), we treat it as a continuous domain and after finding the input the maximizes the acquisition function, we round the discrete inputs to the nearest integer. For Neural UCB [69], we treat the UCB value calculated for each arm (input) as the acquisition function; for Neural TS, we treat the reward sampled for each arm (input) as the acquisition function [66].

We implement the training of the surrogate model $f_t^i(\mathbf{x};\theta)$ for both Algos. 1 and 2 based on the implementations from the work of [24], and we adopt all their default parameter settings (refer to the implementations of [24] for the specific parameter settings, available at `https://github.com/bobby-he/bayesian-ntk`) and only vary the architecture of the NN surrogate model (e.g., the depth and width of the NN, we replace the NN with a CNN for our experiments in Sec. 5.4) as we

have mentioned in the main text. For both Neural UCB and Neural TS, we adopt the implementations from the work of [66], use all their default parameter settings, and only modify the architecture of their NN surrogate model for a fair comparison with our methods. As we have mentioned in the main text, to apply Neural UCB and Neural TS for problems with continuous domains, we adapt their implementations such that we optimize their acquisition functions in the same way as our methods (i.e., through a combination of random search and L-BFGS-B as discussed above). Our experiments are performed using a computing cluster where each machine has an NVIDIA A100 GPU and 96 CPUs.

F.1 Synthetic Experiments

In the synthetic experiment, the objective function f is sampled from a GP with an SE kernel using a lengthscale of 0.1. The domain of f is a uniform grid of size $1,000$ in the interval of $[0, 1]$.

F.2 Real-world Experiments on Automated ML

In this section, we give more details on the three hyperparameter tuning experiments in Sec. 5.2.

Hyperparamter Tuning of Random Forest. Here we tune 6 categorical hyperparameters of random forest:

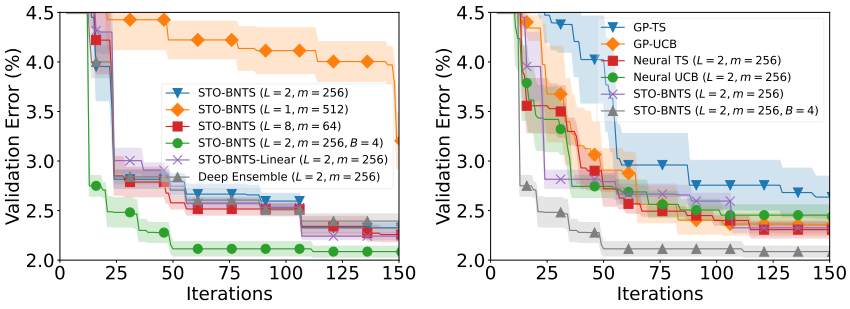
- the maximum depth of any individual tree (integer within $[1, 10]$),
- the minimum number of samples required to split an internal node (integer within $[2, 10]$),
- the minimum number of samples required to be at a leaf node (integer within $[1, 10]$),
- the maximum number of features to consider when looking for the best split (integer within $[1, 8]$),
- the criterion to measure the quality of a split (binary, "entropy" or "gini"),
- whether bootstrap samples are used when building trees (binary, True or False).

We use the publicly available diabetes prediction dataset which can be accessed from `https://www.kaggle.com/uciml/pima-indians-diabetes-database` and has the CC0 License. This dataset does not contains personally identifiable information or offensive content. It consists of 768 data instances, each containing 8 input features. We use 70% of the dataset as the training set and the remaining 30% as the validation set. We use random search to choose 5 initial inputs as the set of initialization, which is shared among all methods under comparison.

Hyperparameter Tuning of XGBoost. The MNIST dataset is publicly available and associated with the GNU General Public License, and can be obtained from the Keras Package[1]. It does not contain personally identifiable information or offensive content. In this experiment, we use the MNIST dataset to tune 9 hyperparameters of XGBoost [6]:

- gamma which represents the minimum loss reduction required to make a further partition on a leaf node of the tree (continuous, $[0, 10]$),
- the learning rate (continuous, $[10^{-6}, 1]$),
- the maximum depth of any individual tree (integer, $[1, 15]$),
- which booster to use (binary, "dart" or "gbtree"),
- the grow policy which controls the way new nodes are added to the tree (binary, "depthwise" or "lossguide"),
- the objective (bianry, "multi:softprob" or "multi:softmax"),
- the tree construction method (binary, "exact" or "hist"),
- alpha which is the L1 regularization term on the weights (continuous, $[0, 10]$), and
- lambda which is the L2 regularization term on the weights (continuous, $[0, 10]$).

[1] `https://keras.io/`



(a) (b)

Figure 1: Validation errors for hyperparameter tuning of CNN. $B = 1$ unless specified otherwise.

Hyperparameter Tuning of Convolutional Neural Networks. Here we use the MNIST dataset to tune 9 hyperparameters of convolutional neural networks (CNN). The CNN consists of one convolutional layer, followed by a max pooling layer and subsequently a fully connected layer. The 9 hyperparameters are:

- the learning rate (continuous, $[10^{-4}, 0.1]$),
- the weight decay (continuous, $[10^{-6}, 10^{-2}]$),
- the batch size (integer, $[64, 512]$),
- the max pooling size (integer, $[3, 5]$),
- the number of neurons in the convolutional layer (integer, $4, 16$),
- the size of the convolutional kernel (integer, $[3, 5]$),
- the number of neurons in the fully connected layer (integer, $[4, 16]$),
- which activation function to use (binary, ReLU or Tanh),
- which optimization method to use (binary, ADAM or RMSprop).

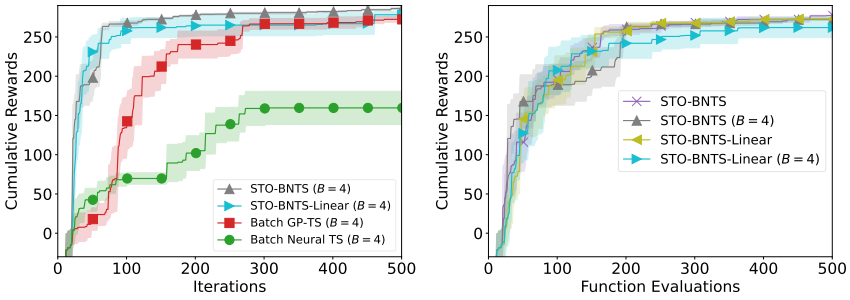
F.3 Real-world Experiments on Reinforcement Learning

The lunar lander task involves tuning 12 parameters of a heuristic controller which is used to control the LunarLander-v2 environment from OpenAI Gym [4]. The heuristic controller is provided by OpenAI Gym and can be found at `https://github.com/openai/gym/blob/8a96440084a6b9be66b2216b984a1c170e4a061c/gym/envs/box2d/lunar_lander.py#L447`. OpenAI Gym[2] is open-sourced and under the MIT License. The (14-dimensional) robots pushing and (20-dimensional) rover trajectory planning tasks were firstly introduced by the work of [62] where the detailed experimental settings can be found. Both tasks are publicly available at `https://github.com/zi-w/Ensemble-Bayesian-Optimization` and are under the MIT license. Due to the large number of iterations (500) of these three experiments (which is necessary as a result of the high dimensionality of the input spaces), standard GP-UCB and GP-TS become too computationally costly to run. Therefore, we applied random Fourier features approximations [11, 12] to the GP using a large number $1,000$ of random features, with which GP-UCB and GP-TS perform well and are still computationally feasible to run.

Using the Lunar-Lander experiment, we have also compared our STO-BNTS and STO-BNTS-Linear with batch versions of GP-TS and Neural TS. The results are shown in Fig. 5 (a), in which all methods use the same batch size of $B = 4$. The figure shows that our STO-BNTS and STO-BNTS-Linear are still able to significantly outperform the other baseline methods when the same batch size is used.

We also use the Lunar-Lander experiment to show an alternative visualization of the performances of our algorithms with batch evaluations in Fig. 5 (b). Specifically, the horizontal axis in Fig. 5 (b) is the number of function evaluations, in contrast to iterations in Fig. 3a. Same as Fig. 3a, this figure also shows the benefit of batch evaluations, because compared with the sequential algorithms ($B = 1$,

[2] `https://github.com/openai/gym`



(a) (b)

Figure 2: (a) Comparison of different algorithms with the same batch size of $B = 4$, including batch versions of our STO-BNTS and STO-BNTS-Linear, as well as batch GP-TS and batch Neural TS. (b) An alternative visualization of the performance of our algorithms with batch evaluations, using the 12-D Lunar-Lander experiment (Fig. 3a). The horizontal axis here is the number of function evaluations, in contrast to iterations in Fig. 3a.

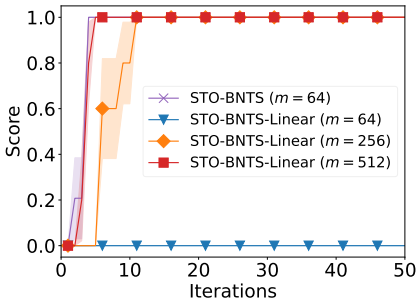


Figure 3: Results of STO-BNTS-Linear in the experiments on optimization over images (Sec. 5.4).

purple and yellow curves), our algorithms with a batch size of $B = 4$ only suffer slight degradations of the per-function evaluation performances.

F.4 Optimization over Images

In this experiment, to construct the score function (Fig. 3d), we firstly use the training set of the MNIST dataset (consisting of $60,000$ images) to train a CNN, and then use the trained CNN to predict the class probabilities for the 10 different classes using the testing set of $10,000$ images. Next, we use the predicted probability of class 0 for the $10,000$ testing images as the score function. As a result of our construction of the score function, similar images in general have similar score values since they share similar representations from the CNN, and images of 0 overall have much larger scores than images from the other classes. This can simulate the real-world scenario of image recommender system, in which the user may prefer a certain type of images and hence give higher ratings to them.

For all three CNN-based methods in Fig. 3d, we use a CNN with one convolutional layer (with convolutional kernels size of 3), followed by a max pooling layer (with a pooling size of 3), and then followed by a fully connected layer. We have used the ReLU activation function. The width of both the convolutional and fully connected layers are $m = 64$. Fig. 6 plots the results for STO-BNTS-Linear in this experiment, which shows that its performance can be dramatically improved if we increase the width m of the NN surrogate model (Sec. 5.4).