A. Derivation of $\mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\mathcal{D}_t} | \mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]$ Term in (2)

By the definition of conditional mutual information,

$$
\begin{aligned}
&\mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\mathcal{D}_t} | \mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}] \\
&= \mathbb{H}[\mathbf{y}_{\mathcal{D}_t} | \mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}] - \mathbb{H}[\mathbf{y}_{\mathcal{D}_t} | \mathbf{f}_{\mathcal{D}}, \mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}] \\
&= \mathbb{H}[\mathbf{y}_{\mathcal{D}_t} | \mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}] - \mathbb{H}[\mathbf{y}_{\mathcal{D}_t} | \mathbf{f}_{\mathcal{D}_t}] \\
&= 0.5|\mathcal{D}_t| \log(2\pi e) + 0.5 \log|\sigma_n^2 I + \Sigma_{\mathcal{D}_t\mathcal{D}_t}| - 0.5|\mathcal{D}_t| \log(2\pi e) - 0.5 \log|\sigma_n^2 I| \\
&= 0.5 \log(|\sigma_n^2 I + \Sigma_{\mathcal{D}_t\mathcal{D}_t}||\sigma_n^2 I|^{-1}) \\
&= 0.5 \log(|\sigma_n^2 I + \Sigma_{\mathcal{D}_t\mathcal{D}_t}||\sigma_n^{-2} I|) \\
&= 0.5 \log|I + \sigma_n^{-2}\Sigma_{\mathcal{D}_t\mathcal{D}_t}|
\end{aligned}
$$

where the third equality is due to the definition of Gaussian entropy, that is, $\mathbb{H}[\mathbf{y}_{\mathcal{D}_t} | \mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}] \triangleq 0.5|\mathcal{D}_t| \log(2\pi e) + 0.5 \log|\sigma_n^2 I + \Sigma_{\mathcal{D}_t\mathcal{D}_t}|$ and $\mathbb{H}[\mathbf{y}_{\mathcal{D}_t} | \mathbf{f}_{\mathcal{D}_t}] \triangleq 0.5|\mathcal{D}_t| \log(2\pi e) + 0.5 \log|\sigma_n^2 I|$, the latter of which follows from $\epsilon = y_{\mathbf{x}} - f(\mathbf{x}) \sim \mathcal{N}(0, \sigma_n^2)$ for all $\mathbf{x} \in \mathcal{D}_t$ and hence $p(\mathbf{y}_{\mathcal{D}_t} | \mathbf{f}_{\mathcal{D}_t}) = \mathcal{N}(\mathbf{0}, \sigma_n^2 I)$.

B. Proof of Proposition 1

$$
\begin{aligned}
&\log|\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}| \\
&= -\log|\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}^{-1}| \\
&= -\log|U^\top U| \\
&= -\log|U^\top||U| \\
&= -\log|U|^2 \\
&= -2\log\prod_{n=1}^{N}|U_{nn}| \\
&= -2\sum_{n=1}^{N}\log|U_{nn}| \\
&= -\sum_{n=1}^{N}\log|U_{nn}|^2 \\
&= -\sum_{n=1}^{N}\log|U_{nn}^\top||U_{nn}| \\
&= -\sum_{n=1}^{N}\log|U_{nn}^\top U_{nn}| \\
&= \sum_{n=1}^{N}\log|U_{nn}^\top U_{nn}|^{-1} \\
&= \sum_{n=1}^{N}\log|(U_{nn}^\top U_{nn})^{-1}|
\end{aligned}
$$

where the first, third, fourth, eighth, ninth, and last equalities follow from the properties of the determinant, the second equality is due to the Cholesky factorization of $\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}^{-1}$, and the fifth equality follows from the property that the determinant of an upper triangular block matrix is a product of determinants of its diagonal blocks (i.e., $|U| = \prod_{n=1}^{N}|U_{nn}|$).

C. Proof of Proposition 4

From the definition of $D_{\text{KL}}(\Psi_{\mathcal{D}_t\mathcal{D}_t}, \overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t})$,

$$
\begin{aligned}
& D_{\text{KL}}(\Psi_{\mathcal{D}_t\mathcal{D}_t}, \overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}) \\
& = 0.5\left(\text{tr}(\Psi_{\mathcal{D}_t\mathcal{D}_t}\overline{\Psi}^{-1}_{\mathcal{D}_t\mathcal{D}_t}) - \log|\Psi_{\mathcal{D}_t\mathcal{D}_t}\overline{\Psi}^{-1}_{\mathcal{D}_t\mathcal{D}_t}| - |\mathcal{D}_t|\right) \\
& = 0.5\left(-\log|\Psi_{\mathcal{D}_t\mathcal{D}_t}\overline{\Psi}^{-1}_{\mathcal{D}_t\mathcal{D}_t}|\right) \\
& = 0.5\left(-\log|\Psi_{\mathcal{D}_t\mathcal{D}_t}| + \log|\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}|\right) \\
& = 0.5\log|\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}| - 0.5\log|\Psi_{\mathcal{D}_t\mathcal{D}_t}| \\
& = \tilde{\mathbb{I}}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\mathcal{D}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}] - \mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\mathcal{D}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]\ .
\end{aligned}
$$

The second equality is due to $\text{tr}(\Psi_{\mathcal{D}_t\mathcal{D}_t}\overline{\Psi}^{-1}_{\mathcal{D}_t\mathcal{D}_t}) = \text{tr}(\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}\overline{\Psi}^{-1}_{\mathcal{D}_t\mathcal{D}_t}) = \text{tr}(I) = |\mathcal{D}_t|$, which follows from the observations that the blocks within the B-block bands of $\Psi_{\mathcal{D}_t\mathcal{D}_t}$ and $\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}$ coincide and $\overline{\Psi}^{-1}_{\mathcal{D}_t\mathcal{D}_t}$ is B-block-banded (Proposition 3). It follows that

$$
\begin{aligned}
& D_{\text{KL}}(\Psi_{\mathcal{D}_t\mathcal{D}_t}, \overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}) \\
& = \tilde{\mathbb{I}}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\mathcal{D}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}] - \mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\mathcal{D}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}] \\
& = 0.5\log|\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}| - \sum_{b=1}^{|\mathcal{D}_t|} 0.5\log\left(1 + \sigma_n^{-2}\Sigma^{b-1}_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}\right) \\
& \leq 0.5\log\left(\prod_{\mathbf{x}\in\mathcal{D}_t}\left(1 + \sigma_n^{-2}\Sigma_{\{\mathbf{x}\}\{\mathbf{x}\}}\right)\right) - \sum_{b=1}^{|\mathcal{D}_t|} 0.5\log\left(1 + \sigma_n^{-2}\Sigma^{b-1}_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}\right) \\
& = \sum_{b=1}^{|\mathcal{D}_t|} 0.5\log\left(1 + \sigma_n^{-2}\Sigma_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}\right) - \sum_{b=1}^{|\mathcal{D}_t|} 0.5\log\left(1 + \sigma_n^{-2}\Sigma^{b-1}_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}\right) \\
& \leq \sum_{b=1}^{|\mathcal{D}_t|} 0.5\log\left(1 + \sigma_n^{-2}\exp(2C)\Sigma^{b-1}_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}\right) - \sum_{b=1}^{|\mathcal{D}_t|} 0.5\log\left(1 + \sigma_n^{-2}\Sigma^{b-1}_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}\right) \\
& \leq \exp(2C)\sum_{b=1}^{|\mathcal{D}_t|} 0.5\log\left(1 + \sigma_n^{-2}\Sigma^{b-1}_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}\right) - \sum_{b=1}^{|\mathcal{D}_t|} 0.5\log\left(1 + \sigma_n^{-2}\Sigma^{b-1}_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}\right) \\
& = (\exp(2C) - 1)\sum_{b=1}^{|\mathcal{D}_t|} 0.5\log\left(1 + \sigma_n^{-2}\Sigma^{b-1}_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}\right) \\
& = (\exp(2C) - 1)\,\mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\mathcal{D}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]\ .
\end{aligned}
$$

The second and last equalities are due to Lemma 4 in Appendix F and $\Sigma^{b-1}_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}$ is defined in Definition 1 in Appendix F. The first inequality is due to Hadamard's inequality and the observation that the blocks within the B-block bands of $\Psi_{\mathcal{D}_t\mathcal{D}_t}$ and $\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}$ (and thus their diagonal elements) coincide. The second inequality is due to Lemma 2 in Appendix F. The third inequality is due to Bernoulli's inequality.

Remark. The first inequality can also be interpreted as bounding the approximated information gain for an arbitrary $\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}$ by the approximated information gain for the $\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}$ with the highest possible degree of our proposed Markov approximation, i.e., for $N = |\mathcal{D}_t|$ and $B = 0$. In this case, all inputs of the batch are assumed to have conditionally independent corresponding outputs such that the determinant of the approximated matrix reduces to the product of its diagonal elements which are equal to the diagonal elements of the original matrix. Thus, $|\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}| \leq \prod_{\mathbf{x}\in\mathcal{D}_t}\left(1 + \sigma_n^{-2}\Sigma_{\{\mathbf{x}\}\{\mathbf{x}\}}\right)$ which interestingly coincides with Hadamard's inequality. Note that we only consider $B \geq 1$ for our proposed algorithm (Proposition 2) since the case of $B = 0$ entails an issue similar to that discussed at the beginning of Section 3 of selecting the same input $|\mathcal{D}_t|$ times within a batch.

D. Minimal KL Distance of Approximated Matrix

For the approximation quality of $\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}$ (4), the following result shows that the *Kullback-Leibler* (KL) distance of $\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}$ from $\Psi_{\mathcal{D}_t\mathcal{D}_t}$ is the least among all $|\mathcal{D}_t| \times |\mathcal{D}_t|$ matrices with a B-block-banded inverse:

Proposition 5. *Let KL distance $D_{\text{KL}}(\Psi, \widetilde{\Psi}) \triangleq 0.5(\text{tr}(\Psi\widetilde{\Psi}^{-1}) - \log|\Psi\widetilde{\Psi}^{-1}| - |\mathcal{D}_t|)$ between two $|\mathcal{D}_t| \times |\mathcal{D}_t|$ symmetric positive definite matrices Ψ and $\widetilde{\Psi}$ measure the error of approximating Ψ with $\widetilde{\Psi}$. Then, $D_{\text{KL}}(\Psi_{\mathcal{D}_t\mathcal{D}_t}, \overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}) \leq$*

$D_{\mathrm{KL}}(\Psi_{\mathcal{D}_t\mathcal{D}_t}, \widetilde{\Psi})$ *for any matrix* $\widetilde{\Psi}$ *with a* B*-block-banded inverse.*

Proof.

$$\begin{aligned}
&D_{\mathrm{KL}}(\Psi_{\mathcal{D}_t\mathcal{D}_t}, \overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}) + D_{\mathrm{KL}}(\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}, \widetilde{\Psi}) \\
&= 0.5\Big(\mathrm{tr}(\Psi_{\mathcal{D}_t\mathcal{D}_t}\overline{\Psi}^{-1}_{\mathcal{D}_t\mathcal{D}_t}) - \log|\Psi_{\mathcal{D}_t\mathcal{D}_t}\overline{\Psi}^{-1}_{\mathcal{D}_t\mathcal{D}_t}| - |\mathcal{D}_t|\Big) + 0.5\Big(\mathrm{tr}(\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}\widetilde{\Psi}^{-1}) - \log|\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}\widetilde{\Psi}^{-1}| - |\mathcal{D}_t|\Big) \\
&= 0.5\left(\mathrm{tr}(\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}\widetilde{\Psi}^{-1}) - \log|\Psi_{\mathcal{D}_t\mathcal{D}_t}| - \log|\widetilde{\Psi}^{-1}| - |\mathcal{D}_t|\right) \\
&= 0.5\left(\mathrm{tr}(\Psi_{\mathcal{D}_t\mathcal{D}_t}\widetilde{\Psi}^{-1}) - \log|\Psi_{\mathcal{D}_t\mathcal{D}_t}\widetilde{\Psi}^{-1}| - |\mathcal{D}_t|\right) \\
&= D_{\mathrm{KL}}(\Psi_{\mathcal{D}_t\mathcal{D}_t}, \widetilde{\Psi})\ .
\end{aligned}$$

The second equality is due to $\mathrm{tr}(\Psi_{\mathcal{D}_t\mathcal{D}_t}\overline{\Psi}^{-1}_{\mathcal{D}_t\mathcal{D}_t}) = \mathrm{tr}(\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}\overline{\Psi}^{-1}_{\mathcal{D}_t\mathcal{D}_t}) = \mathrm{tr}(I) = |\mathcal{D}_t|$, which follows from the observations that the blocks within the B-block bands of $\Psi_{\mathcal{D}_t\mathcal{D}_t}$ and $\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}$ coincide and $\overline{\Psi}^{-1}_{\mathcal{D}_t\mathcal{D}_t}$ is B-block-banded (Proposition 3). The third equality follows from the first observation above and the definition that $\widetilde{\Psi}^{-1}$ is B-block-banded. Since $D_{\mathrm{KL}}(\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}, \widetilde{\Psi}) \geq 0$, $D_{\mathrm{KL}}(\Psi_{\mathcal{D}_t\mathcal{D}_t}, \overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}) \leq D_{\mathrm{KL}}(\Psi_{\mathcal{D}_t\mathcal{D}_t}, \widetilde{\Psi})$. □

E. Pseudocode for DB-GP-UCB

Algorithm 1 DB-GP-UCB

Input: Objective function f, input domain $\mathcal{D}$, batch size $|\mathcal{D}_t|$, time horizon T, prior mean $m_{\mathbf{x}}$ and kernel $k_{\mathbf{xx}'}$, approximation parameters B and N

for $t = 1, \ldots, T$ **do**

Select acquisition function $a(\mathcal{D}_t) \triangleq \begin{cases} \mathbf{1}^\top \mu_{\mathcal{D}_t} + \sqrt{\alpha_t\, \mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\mathcal{D}_t}|\mathbf{y}_{\overline{\mathcal{D}}_{1:t\text{-}1}}]} & (2) \quad \text{if } B = N-1\ , \\ \sum_{n=1}^{N} \mathbf{1}^\top \mu_{\mathcal{D}_{tn}} + \sqrt{0.5\alpha_t \log|\Psi_{\mathcal{D}_{tn}\mathcal{D}_{tn}|\mathcal{D}^B_{tn}}|} & (5) \quad \text{otherwise} \end{cases}$

Select batch $\overline{\mathcal{D}}_t \triangleq \arg\max_{\mathcal{D}_t \subset \mathcal{D}} a(\mathcal{D}_t)$

Query batch $\overline{\mathcal{D}}_t$ to obtain $\mathbf{y}_{\overline{\mathcal{D}}_t} \triangleq (f(\mathbf{x}) + \epsilon)^\top_{\mathbf{x} \in \overline{\mathcal{D}}_t}$

end for

Output: Recommendation $\widetilde{\mathbf{x}} \triangleq \arg\max_{\mathbf{x} \in \mathcal{D}} \mu_{\{\mathbf{x}\}}$

F. Proof of Theorem 1

We first define a different notion of posterior variance:

Definition 1 (Updated Posterior Variance)**.** *Let* $\mathcal{D}_t \triangleq \{\mathbf{x}_1, \ldots, \mathbf{x}_{|\mathcal{D}_t|}\}$ *be the batch selected in iteration* t*. Assume an arbitrary ordering of the inputs in* $\mathcal{D}_t$*. Then, for* $0 \leq b-1 < |\mathcal{D}_t|$*,* $\Sigma^{b-1}_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}$ *is defined as the updated posterior variance at input* $\mathbf{x}_b$ *that is obtained by applying (1) conditioned on the previous inputs in the batch* $\mathcal{D}^{b-1}_t \triangleq \{\mathbf{x}_1, \ldots, \mathbf{x}_{b-1}\}$*. Note that performing this update is possible without querying* $\mathcal{D}^{b-1}_t$ *since* $\Sigma^{b-1}_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}$ *is independent of the outputs* $\mathbf{y}_{\mathcal{D}^{b-1}_t}$*. For* $b-1 = 0$*,* $\Sigma^{b-1}_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}$ *reduces to* $\Sigma_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}$*.*

The following lemmas are necessary for proving our main result here:

Lemma 1. *Let* $\delta \in (0,1)$ *be given and* $\beta_t \triangleq 2\log(|\mathcal{D}|\pi_t/\delta)$ *where* $\sum_{t=1}^{\infty} \pi_t^{-1} = 1$ *and* $\pi_t > 0$*. Then,*

$$\Pr\left(\forall \mathbf{x} \in \mathcal{D}\ \ \forall t \in \mathbb{N}\ \ |f(\mathbf{x}) - \mu_{\{\mathbf{x}\}}| \leq \beta_t^{1/2}\Sigma^{1/2}_{\{\mathbf{x}\}\{\mathbf{x}\}}\right) \geq 1-\delta\ .$$

Lemma 1 above corresponds to Lemma 5.1 in (Srinivas et al., 2010); see its proof therein. For example, $\pi_t = t^2\pi^2/6 > 0$ satisfies $\sum_{t=1}^{\infty} \pi_t^{-1} = 1$.

Lemma 2. *For* f *sampled from a known GP prior with known noise variance* σ_n^2*, the ratio of* $\Sigma_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}$ *to* $\Sigma^{b-1}_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}$ *for all* $\mathbf{x}_b \in \mathcal{D}_t$ *is bounded by*

$$\frac{\Sigma_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}}{\Sigma^{b-1}_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}} = \exp\left(2\,\mathbb{I}[\mathbf{f}_{\{\mathbf{x}_b\}}; \mathbf{y}_{\mathcal{D}^{b-1}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]\right) \leq \exp(2C)$$

where $\Sigma^{b-1}_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}$ *and* $\mathcal{D}^{b-1}_t$ *are previously defined in Definition 1, and for all* $\mathbf{x} \in \mathcal{D}$ *and* $t \in \mathbb{N}$,

$$C \geq \mathbb{I}[\mathbf{f}_{\{\mathbf{x}\}}; \mathbf{y}_{\mathcal{D}^{b-1}_t} | \mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]$$

is a suitable constant.

Lemma 2 above is a combination of Proposition 1 and equation 9 in (Desautels et al., 2014); see their proofs therein. The only difference is that we equivalently bound the ratio of variances instead of the ratio of standard deviations, thus leading to an additional factor of 2 in the argument of exp.

Remark. Since the upper bound $\exp(2C)$ will appear in our regret bounds, we need to choose C suitably. A straightforward choice $C \triangleq \gamma_{|\mathcal{D}_t|-1} = \max_{\mathcal{A}\subset\mathcal{D}, |\mathcal{A}|\leq|\mathcal{D}_t|-1} \mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\mathcal{A}}] \geq \max_{\mathcal{A}\subset\mathcal{D}, |\mathcal{A}|\leq|\mathcal{D}_t|-1} \mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\mathcal{A}} | \mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}] \geq \mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\mathcal{D}^{b-1}_t} | \mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}] \geq \mathbb{I}[\mathbf{f}_{\{\mathbf{x}\}}; \mathbf{y}_{\mathcal{D}^{b-1}_t} | \mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]$ (see equations 11, 12, and 13 in (Desautels et al., 2014)) is unfortunately unsatisfying from the perspective of asymptotic scaling since it grows at least as $\Omega(\log|\mathcal{D}_t|)$, thus implying that $\exp(2C)$ grows at least linearly in $|\mathcal{D}_t|$ and yielding a regret bound that is also at least linear in $|\mathcal{D}_t|$. The work of Desautels et al. (2014) shows that when initializing an algorithm suitably, one can obtain a constant C *independent* of the batch size $|\mathcal{D}_t|$. Refer to Section 4 in (Desautels et al., 2014) for a more detailed discussion.

Lemma 3. *For all* $t \in \mathbb{N}$ *and* $\mathbf{x}_b \in \mathcal{D}_t$,

$$\Sigma^{b-1}_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}} \leq 0.5 C_0 \log\left(1 + \sigma_n^{-2} \Sigma^{b-1}_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}\right)$$

where $C_0 \triangleq 2/\log(1+\sigma_n^{-2})$.

Lemma 3 above corresponds to an intermediate step of Lemma 5.4 in (Srinivas et al., 2010); see its proof therein.

Lemma 4. *The information gain for a batch* $\mathcal{D}_t$ *chosen in any iteration* t *can be expressed in terms of the updated posterior variances of the individual inputs* $\mathbf{x}_b \in \mathcal{D}_t, b \in \{1, \ldots, |\mathcal{D}_t|\}$ *of the batch* $\mathcal{D}_t$. *That is, for all* $t \in \mathbb{N}$,

$$\mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\mathcal{D}_t} | \mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}] = 0.5 \sum_{b=1}^{|\mathcal{D}_t|} \log\left(1 + \sigma_n^{-2} \Sigma^{b-1}_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}\right).$$

Lemma 4 above corresponds to Lemma 5.3 in (Srinivas et al., 2010) (the only difference being that we equivalently sum over $1, \ldots, |\mathcal{D}_t|$ instead of $1, \ldots, T$); see its proof therein.

Lemma 5. *Let* $\delta \in (0,1)$ *be given,* $C_0 \triangleq 2/\log(1+\sigma_n^{-2})$, *and* $\alpha_t \triangleq C_0 |\mathcal{D}_t| \exp(2C) \beta_t$ *where* β_t *and* $\exp(2C)$ *are previously defined in Lemmas 1 and 2, respectively. Then,*

$$\Pr\left(\forall \mathcal{D}_t \subset \mathcal{D}\ \ \forall t \in \mathbb{N}\ \ \sum_{\mathbf{x}\in\mathcal{D}_t} |f(\mathbf{x}) - \mu_{\{\mathbf{x}\}}| \leq \sqrt{\alpha_t\, \mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\mathcal{D}_t} | \mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]}\right) \geq 1 - \delta\,.$$

Proof. For all $\mathcal{D}_t \subset \mathcal{D}$ and $t \in \mathbb{N}$,

$$\begin{aligned}
&\sum_{\mathbf{x}\in\mathcal{D}_t} \beta_t\, \Sigma_{\{\mathbf{x}\}\{\mathbf{x}\}} \\
&= \beta_t \sum_{b=1}^{|\mathcal{D}_t|} \Sigma_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}} \\
&\leq \beta_t \sum_{b=1}^{|\mathcal{D}_t|} \exp(2C)\, \Sigma^{b-1}_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}} \\
&\leq 0.5 C_0 \exp(2C) \beta_t \sum_{b=1}^{|\mathcal{D}_t|} \log\left(1 + \sigma_n^{-2} \Sigma^{b-1}_{\{\mathbf{x}_b\}\{\mathbf{x}_b\}}\right) \\
&= C_0 \exp(2C) \beta_t\, \mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\mathcal{D}_t} | \mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}] \\
&= |\mathcal{D}_t|^{-1} \alpha_t\, \mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\mathcal{D}_t} | \mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]
\end{aligned}$$

where the first inequality is due to Lemma 2, the second inequality is due to Lemma 3, and the second equality is due to Lemma 4. Thus,

$$\sum_{\mathbf{x}\in\mathcal{D}_t} \beta_t^{1/2}\Sigma_{\{\mathbf{x}\}\{\mathbf{x}\}}^{1/2} \le \sqrt{|\mathcal{D}_t| \sum_{\mathbf{x}\in\mathcal{D}_t} \beta_t \Sigma_{\{\mathbf{x}\}\{\mathbf{x}\}}} \le \sqrt{\alpha_t\, \mathbb{I}[\mathbf{f}_\mathcal{D};\mathbf{y}_{\mathcal{D}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]}$$

where the first inequality is due to the Cauchy-Schwarz inequality. It follows that

$$\begin{aligned}
&\Pr\left(\forall\mathcal{D}_t\subset\mathcal{D}\;\;\forall t\in\mathbb{N}\;\;\sum_{\mathbf{x}\in\mathcal{D}_t}|f(\mathbf{x})-\mu_{\{\mathbf{x}\}}|\le\sqrt{\alpha_t\,\mathbb{I}[\mathbf{f}_\mathcal{D};\mathbf{y}_{\mathcal{D}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]}\right)\\
&\ge\Pr\left(\forall\mathcal{D}_t\subset\mathcal{D}\;\;\forall t\in\mathbb{N}\;\;\sum_{\mathbf{x}\in\mathcal{D}_t}|f(\mathbf{x})-\mu_{\{\mathbf{x}\}}|\le\sum_{\mathbf{x}\in\mathcal{D}_t}\beta_t^{1/2}\Sigma_{\{\mathbf{x}\}\{\mathbf{x}\}}^{1/2}\right)\\
&\ge\Pr\left(\forall\mathbf{x}\in\mathcal{D}\;\;\forall t\in\mathbb{N}\;\;|f(\mathbf{x})-\mu_{\{\mathbf{x}\}}|\le\beta_t^{1/2}\Sigma_{\{\mathbf{x}\}\{\mathbf{x}\}}^{1/2}\right)\\
&\ge 1-\delta
\end{aligned}$$

where the first two inequalities are due to the property that for logical propositions A and B, $[A \implies B] \implies [\Pr(A) \le \Pr(B)]$, and the last inequality is due to Lemma 1. □

Lemma 6. *Let $\nu_t \ge \tilde{\mathbb{I}}[\mathbf{f}_\mathcal{D};\mathbf{y}_{\mathcal{D}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]-\mathbb{I}[\mathbf{f}_\mathcal{D};\mathbf{y}_{\mathcal{D}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]$ be an upper bound on the approximation error of $\overline{\Psi}_{\mathcal{D}_t\mathcal{D}_t}$. Then, for all $t\in\mathbb{N}$,*

$$\sum_{n=1}^{N}\sqrt{0.5\log|\Psi_{\mathcal{D}_{tn}\mathcal{D}_{tn}|\mathcal{D}_{tn}^B}|}\le\sqrt{N(\mathbb{I}[\mathbf{f}_\mathcal{D};\mathbf{y}_{\mathcal{D}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]+\nu_t)}\,.$$

Proof.

$$\begin{aligned}
&\sum_{n=1}^{N}\sqrt{0.5\log|\Psi_{\mathcal{D}_{tn}\mathcal{D}_{tn}|\mathcal{D}_{tn}^B}|}\\
&\le\sqrt{N\sum_{n=1}^{N}0.5\log|\Psi_{\mathcal{D}_{tn}\mathcal{D}_{tn}|\mathcal{D}_{tn}^B}|}\\
&=\sqrt{N\tilde{\mathbb{I}}[\mathbf{f}_\mathcal{D};\mathbf{y}_{\mathcal{D}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]}\\
&=\sqrt{N(\mathbb{I}[\mathbf{f}_\mathcal{D};\mathbf{y}_{\mathcal{D}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]+\tilde{\mathbb{I}}[\mathbf{f}_\mathcal{D};\mathbf{y}_{\mathcal{D}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]-\mathbb{I}[\mathbf{f}_\mathcal{D};\mathbf{y}_{\mathcal{D}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}])}\\
&\le\sqrt{N(\mathbb{I}[\mathbf{f}_\mathcal{D};\mathbf{y}_{\mathcal{D}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]+\nu_t)}
\end{aligned}$$

where the first inequality is due to the Cauchy-Schwarz inequality. □

Lemma 7. *Let $t\in\mathbb{N}$ be given. If*

$$\sum_{\mathbf{x}\in\mathcal{D}_t}|f(\mathbf{x})-\mu_{\{\mathbf{x}\}}|\le\sqrt{\alpha_t\,\mathbb{I}[\mathbf{f}_\mathcal{D};\mathbf{y}_{\mathcal{D}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]} \tag{6}$$

for all $\mathcal{D}_t \subset \mathcal{D}$, then $\sum_{\mathbf{x}\in\overline{\mathcal{D}}_t} r_\mathbf{x} \le 2\sqrt{\alpha_t\, N(\mathbb{I}[\mathbf{f}_\mathcal{D};\mathbf{y}_{\overline{\mathcal{D}}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]+\nu_t)}$ and $\min_{\mathbf{x}\in\overline{\mathcal{D}}_t} r_\mathbf{x} \le 2\sqrt{|\overline{\mathcal{D}}_t|^{-2}\alpha_t\, N(\mathbb{I}[\mathbf{f}_\mathcal{D};\mathbf{y}_{\overline{\mathcal{D}}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]+\nu_t)}$.

Proof.

$$\begin{aligned}
&\sum_{\mathbf{x}\in\overline{\mathcal{D}}_t} r_{\mathbf{x}} \\
&= \sum_{\mathbf{x}\in\overline{\mathcal{D}}_t} (f(\mathbf{x}^*) - f(\mathbf{x})) \\
&= \sum_{\mathbf{x}\in\overline{\mathcal{D}}_t} f(\mathbf{x}^*) - \sum_{\mathbf{x}\in\overline{\mathcal{D}}_t} f(\mathbf{x}) \\
&\le \left(\sum_{\mathbf{x}\in\overline{\mathcal{D}}_t} \mu_{\{\mathbf{x}\}} + \sqrt{\alpha_t\, N(\mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\overline{\mathcal{D}}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}] + \nu_t)}\right) - \sum_{\mathbf{x}\in\overline{\mathcal{D}}_t} f(\mathbf{x}) \\
&= \sqrt{\alpha_t\, N(\mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\overline{\mathcal{D}}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}] + \nu_t)} + \left(\sum_{\mathbf{x}\in\overline{\mathcal{D}}_t} \mu_{\{\mathbf{x}\}} - \sum_{\mathbf{x}\in\overline{\mathcal{D}}_t} f(\mathbf{x})\right) \\
&= \sqrt{\alpha_t\, N(\mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\overline{\mathcal{D}}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}] + \nu_t)} + \sum_{\mathbf{x}\in\overline{\mathcal{D}}_t} \left(\mu_{\{\mathbf{x}\}} - f(\mathbf{x})\right) \\
&\le \sqrt{\alpha_t\, N(\mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\overline{\mathcal{D}}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}] + \nu_t)} + \sqrt{\alpha_t\, N(\mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\overline{\mathcal{D}}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}] + \nu_t)} \\
&= 2\sqrt{\alpha_t\, N(\mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\overline{\mathcal{D}}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}] + \nu_t)}\ .
\end{aligned} \tag{7}$$

The first equality in (7) is by definition (Section 2). The first inequality in (7) is due to

$$\begin{aligned}
&\sum_{\mathbf{x}\in\overline{\mathcal{D}}_t} f(\mathbf{x}^*) \\
&= \sum_{\mathbf{x}\in\mathcal{D}_t^*} f(\mathbf{x}) \\
&\le \sum_{\mathbf{x}\in\mathcal{D}_t^*} \mu_{\{\mathbf{x}\}} + \sqrt{\alpha_t\, \mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\mathcal{D}_t^*}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]} \\
&\le \sum_{\mathbf{x}\in\mathcal{D}_t^*} \mu_{\{\mathbf{x}\}} + \sqrt{\alpha_t\, \tilde{\mathbb{I}}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\mathcal{D}_t^*}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}]} \\
&= \sum_{\mathbf{x}\in\mathcal{D}_t^*} \mu_{\{\mathbf{x}\}} + \sqrt{\alpha_t \sum_{n=1}^{N} 0.5\log|\Psi_{\mathcal{D}_{tn}^*\mathcal{D}_{tn}^*|\mathcal{D}_{tn}^{*B}}|} \\
&\le \sum_{\mathbf{x}\in\mathcal{D}_t^*} \mu_{\{\mathbf{x}\}} + \sqrt{\alpha_t} \sum_{n=1}^{N} \sqrt{0.5\log|\Psi_{\mathcal{D}_{tn}^*\mathcal{D}_{tn}^*|\mathcal{D}_{tn}^{*B}}|} \\
&\le \sum_{\mathbf{x}\in\overline{\mathcal{D}}_t} \mu_{\{\mathbf{x}\}} + \sqrt{\alpha_t} \sum_{n=1}^{N} \sqrt{0.5\log|\Psi_{\overline{\mathcal{D}}_{tn}\overline{\mathcal{D}}_{tn}|\overline{\mathcal{D}}_{tn}^{B}}|} \\
&\le \sum_{\mathbf{x}\in\overline{\mathcal{D}}_t} \mu_{\{\mathbf{x}\}} + \sqrt{\alpha_t\, N(\mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\overline{\mathcal{D}}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}] + \nu_t)}
\end{aligned} \tag{8}$$

where, in (8), the first inequality is due to (6), the second inequality is due to Proposition 4 (see the paragraph after this proposition in particular), the third inequality is due to the simple observation that $\sum_{n=1}^{N}\sqrt{a_n} \ge \sqrt{\sum_{n=1}^{N} a_n}$, the fourth inequality follows from the definition of $\overline{\mathcal{D}}_t$ in (5) and, with a slight abuse of notation, $\mathcal{D}_t^*$ is defined as a batch of $|\overline{\mathcal{D}}_t|$ inputs $\mathbf{x}^*$, and the last inequality is due to Lemma 6. The last inequality in (7) follows from (6) and an argument equivalent to the one in (8) (i.e., by substituting $\mathcal{D}_t^*$ by $\overline{\mathcal{D}}_t$).

From (7),

$$\min_{\mathbf{x}\in\overline{\mathcal{D}}_t} r_{\mathbf{x}} \le \frac{1}{|\overline{\mathcal{D}}_t|}\sum_{\mathbf{x}\in\overline{\mathcal{D}}_t} r_{\mathbf{x}} \le 2\sqrt{|\overline{\mathcal{D}}_t|^{-2}\alpha_t\, N(\mathbb{I}[\mathbf{f}_{\mathcal{D}}; \mathbf{y}_{\overline{\mathcal{D}}_t}|\mathbf{y}_{\mathcal{D}_{1:t\text{-}1}}] + \nu_t)}\ .$$

□

Main Proof.

$$
\begin{aligned}
& R'_T \\
&= \sum_{t=1}^{T} \sum_{\mathbf{x}\in\overline{\mathcal{D}}_t} r_{\mathbf{x}} \\
&\leq \sum_{t=1}^{T} 2\sqrt{\alpha_t\ N(\mathbb{I}[\mathbf{f}_{\mathcal{D}};\mathbf{y}_{\overline{\mathcal{D}}_t}|\mathbf{y}_{\overline{\mathcal{D}}_{1:t\text{-}1}}]+\nu_t)} \\
&\leq 2\sqrt{T\sum_{t=1}^{T}\alpha_t\ N(\mathbb{I}[\mathbf{f}_{\mathcal{D}};\mathbf{y}_{\overline{\mathcal{D}}_t}|\mathbf{y}_{\overline{\mathcal{D}}_{1:t\text{-}1}}]+\nu_t)} \\
&\leq 2\sqrt{T\alpha_T\ N\left(\sum_{t=1}^{T}\mathbb{I}[\mathbf{f}_{\mathcal{D}};\mathbf{y}_{\overline{\mathcal{D}}_t}|\mathbf{y}_{\overline{\mathcal{D}}_{1:t\text{-}1}}]+\sum_{t=1}^{T}\nu_t\right)} \\
&= 2\sqrt{T\alpha_T\ N\left(\mathbb{I}[\mathbf{f}_{\mathcal{D}};\mathbf{y}_{\overline{\mathcal{D}}_{1:T}}]+\bar{\nu}_T\right)} \\
&\leq 2\sqrt{T\alpha_T\ N\left(\gamma_T+\bar{\nu}_T\right)} \\
&= \sqrt{C_2 T|\overline{\mathcal{D}}_T|\exp(2C)\beta_T\ N\left(\gamma_T+\bar{\nu}_T\right)}
\end{aligned}
$$

holds with probability $1-\delta$ where the first equality is by definition (Section 2), the first inequality follows from Lemmas 5 and 7, the second inequality is due to the Cauchy-Schwarz inequality, the third inequality is due to the non-decreasing α_t with increasing t, the second equality follows from the chain rule for mutual information and the definition of $\bar{\nu}_T \triangleq \sum_{t=1}^{T}\nu_t$, the fourth inequality is by definition (Theorem 1), and the third equality is due to the definition of α_t in Lemma 5, $|\overline{\mathcal{D}}_1| = \ldots = |\overline{\mathcal{D}}_T|$ and the definition that $C_2 \triangleq 4C_0 = 8/\log(1+\sigma_n^{-2})$.

Analogous reasoning leads to the result that

$$
R_T = \sum_{t=1}^{T}\min_{\mathbf{x}\in\overline{\mathcal{D}}_t} r_{\mathbf{x}} \leq 2\sqrt{T|\overline{\mathcal{D}}_T|^{-2}\alpha_T\ N(\gamma_T+\bar{\nu}_T)} = \sqrt{C_2 T|\overline{\mathcal{D}}_T|^{-1}\exp(2C)\beta_T\ N(\gamma_T+\bar{\nu}_T)}
$$

holds with probability $1-\delta$, where the first equality is by definition (Section 2).

G. Comparison of Regret Bounds

Table 1. Bounds on R_T ($\beta_T \triangleq 2\log(|\mathcal{D}|T^2\pi^2/(6\delta))$, $C_1 \triangleq 4/\log(1+\sigma_n^{-2})$, $C_2 \triangleq 2C_1$, $C_3 \triangleq 9C_1$). Note that $|\mathcal{D}_T| = 1$ in γ_T for GP-UCB and $H_{\text{DPP}} \triangleq \sum_{t=1}^{T} H(DPP(K_t))$ with $H(DPP(K))$ denoting the entropy of a $(|\mathcal{D}_t|-1)$-DPP with kernel K (see (Kathuria et al., 2016) for details on their proposed kernels). Also, note that for DB-GP-UCB and GB-BUCB, we assume the use of the initialization strategy proposed by Desautels et al. (2014); otherwise, the factor C' is replaced by $\sqrt{\exp(2C)}$.

BO Algorithm	Bound on R_T
DB-GP-UCB (5)	$C'\sqrt{C_2 T|\mathcal{D}_T|^{-1}\beta_T\ N(\gamma_T+\bar{\nu}_T)}$
GP-UCB-PE (Contal et al., 2013)	$\sqrt{C_1 T|\mathcal{D}_T|^{-1}\beta_T\gamma_T}$
GP-BUCB (Desautels et al., 2014)	$C'\sqrt{C_2 T|\mathcal{D}_T|^{-1}\beta_{T|\mathcal{D}_T|}\gamma_T}$
GP-UCB (Srinivas et al., 2010)	$\sqrt{C_2 T\beta_T\gamma_T}$
UCB-DPP-SAMPLE (Kathuria et al., 2016)	$\sqrt{2C_3 T|\mathcal{D}_T|\beta_T\left[\gamma_T - H_{\text{DPP}} + |\mathcal{D}_T|\log(|\mathcal{D}|)\right]}$

Table 2. Bounds on maximum mutual information γ_T (Srinivas et al., 2010; Kathuria et al., 2016) and values of C' (Desautels et al., 2014) for different commonly-used kernels ($\alpha \triangleq d(d+1)/(2\nu + d(d+1)) \leq 1$ with ν being the Matérn parameter).

Kernel	γ_T	C'
Linear	$d\log(T\lvert\mathcal{D}_T\rvert)$	$\exp(2/e)$
RBF	$(\log(T\lvert\mathcal{D}_T\rvert))^d$	$\exp((2d/e)^d)$
Matérn	$(T\lvert\mathcal{D}_T\rvert)^\alpha \log(T\lvert\mathcal{D}_T\rvert)$	e

H. Synthetic Benchmark Objective Functions and Real-World pH Field

Table 3. Synthetic benchmark objective functions.

Name	Function	$\mathcal{D}$
Branin-Hoo	$f(\mathbf{x}) = a(x_2 - bx_1^2 + cx_1 - r)^2 + s(1-t)\cos(x_1) + s$ where $a = 1, b = 5.1/(4\pi^2), c = 5/\pi, r = 6, s = 10$, and $t = 1/(8\pi)$.	$[-5, 15]^2$
gSobol	$f(\mathbf{x}) = \prod_{i=1}^{d} \frac{\lvert 4x_i - 2\rvert + a_i}{1 + a_i}$ where $d = 2$ and $a_i = 1$ for $i = 1, \ldots, d$.	$[-5, 5]^2$
Mixture of cosines	$f(\mathbf{x}) = 1 - \sum_{i=1}^{2}(g(x_i) - r(x_i))$ where $g(x_i) = (1.6x_i - 0.5)^2$ and $r(x_i) = 0.3\cos(3\pi(1.6x_i - 0.5))$.	$[-1, 1]^2$

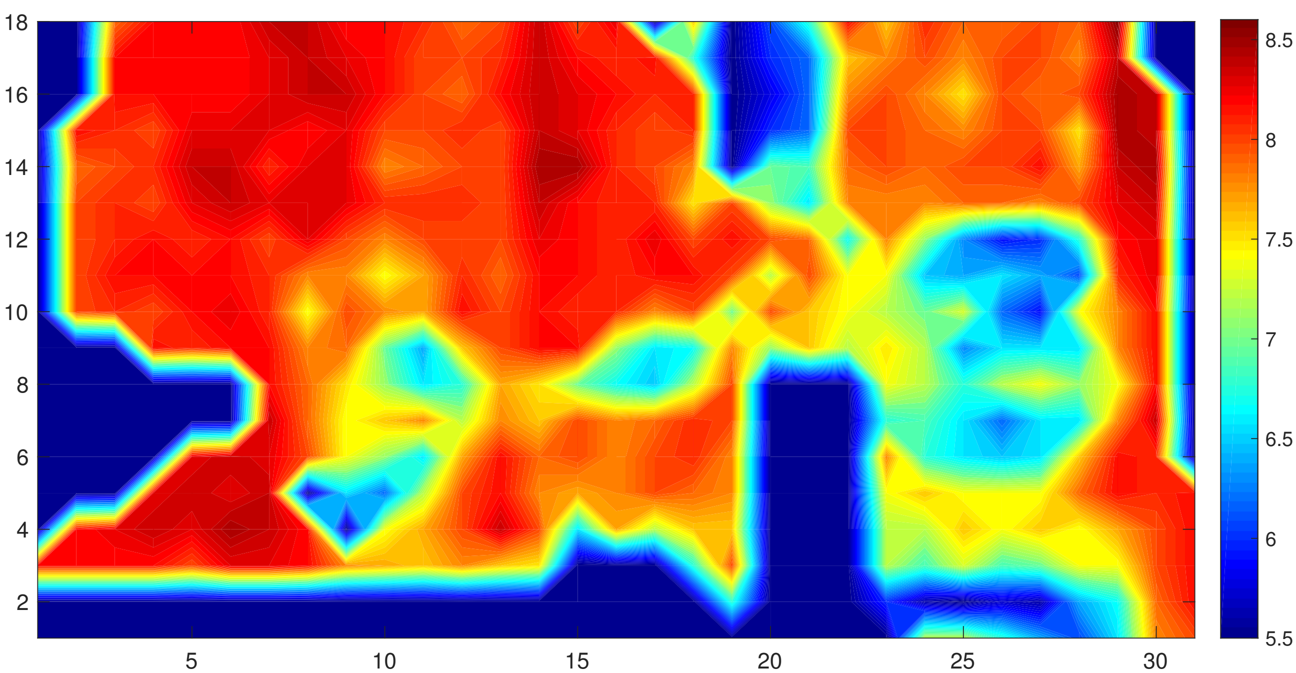

Figure 3. Real-world pH field of Broom's Barn farm (Webster & Oliver, 2007).

I. Details on the Implementations of Batch BO Algorithms

Table 4. Details on the available implementations of the batch BO algorithms for comparison with DB-GP-UCB in our experiments.

BO Algorithm	Language	URL of Source Code
GP-BUCB	MATLAB	`http://www.gatsby.ucl.ac.uk/~tdesautels/`
SM-UCB	MATLAB	`http://www.gatsby.ucl.ac.uk/~tdesautels/`
GP-UCB-PE	MATLAB	`http://econtal.perso.math.cnrs.fr/software/`
q-EI	R	`http://cran.r-project.org/web/packages/DiceOptim/`
BBO-LP	Python	`http://sheffieldml.github.io/GPyOpt/`

J. Analysis of the Trade-Off between the Approximation Quality vs. Time Efficiency of DB-GP-UCB

We now analyze the trade-off between the approximation quality vs. time efficiency of DB-GP-UCB by varying the Markov order B and number N of functions in DCOP. The mixture of cosines function (Anderson et al., 2000) is used as the objective function f and modeled as a sample of a GP. A large batch size $|\mathcal{D}_T| = 16$ is used as it allows us to compare a multitude of different configurations of $[N, B] \in \{[16, 14], [16, 12], \ldots, [16, 0], [8, 6], [8, 4], [8, 2], [8, 0], [4, 2], [4, 0], [2, 0]\}$. The acquisition function in our batch variant of GP-UCB (2) is used as the performance metric to evaluate the approximation quality of the batch $\overline{\mathcal{D}}_T$ (i.e., by plugging $\overline{\mathcal{D}}_T$ into (2) to compute the value of the acquisition function) produced by our DB-GP-UCB algorithm (5) for each configuration of $[N, B]$.

Fig. 4 (top) shows results of the normalized values of the acquisition function in (2) achieved by plugging in the batch $\overline{\mathcal{D}}_T$ produced by DP-GP-UCB (5) for different configurations of $[N, B]$ such that the optimal value of (2) (i.e., achieved in the case of $N = 1$) is normalized to 1. Fig. 4 (bottom) shows the corresponding time complexity of DP-GP-UCB plotted in $\log_{|\mathcal{D}|}$-scale, thus displaying the values of $(B + 1)|\overline{\mathcal{D}}_T|/N$. It can be observed that the approximation quality improves near-linearly with an increasing Markov order B at the expense of higher computational cost (i.e., exponential in B).

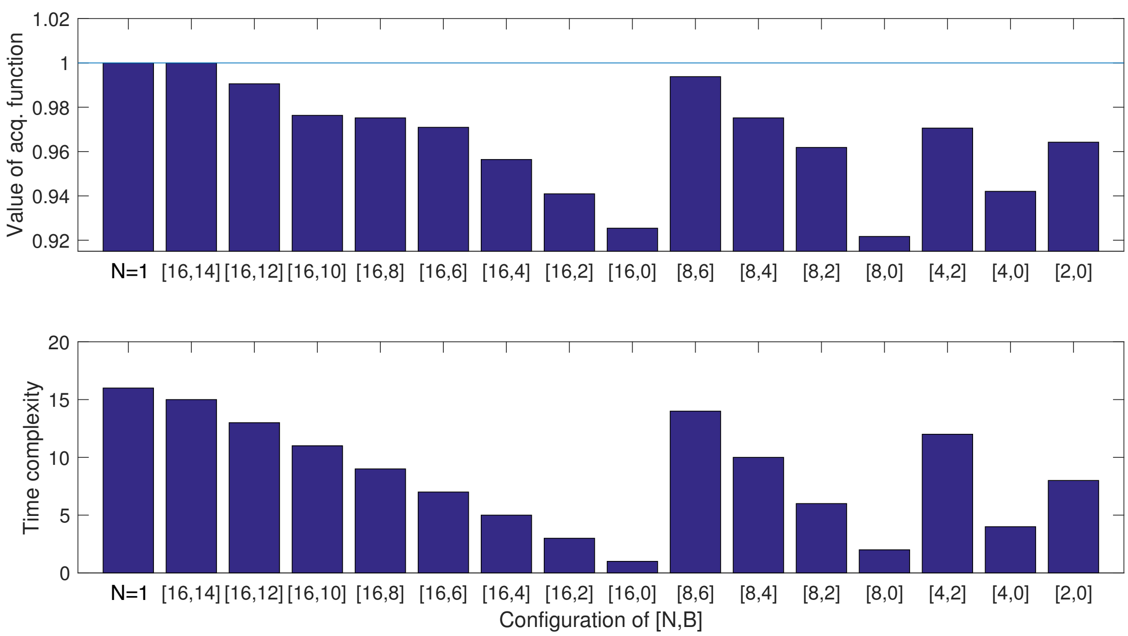

Figure 4. (Top) Mean of the normalized value of the acquisition function in (2) (over 64 experiments of randomly selected noisy observations of size 5) achieved by plugging in the batch $\overline{\mathcal{D}}_T$ (of size 16) produced by our DP-GP-UCB algorithm (5) for different configurations of $[N, B]$ (including the case of $N = 1$ yielding the optimal value of (2)); note that the horizontal line is set at the optimal baseline of $y = 1$ for easy comparison and the y-axis starts at $y = 0.915$. (Bottom) Time complexity of DP-GP-UCB for different configurations of $[N, B]$ plotted in $\log_{|\mathcal{D}|}$-scale.

K. Replication of Regret Graphs including Error Bars

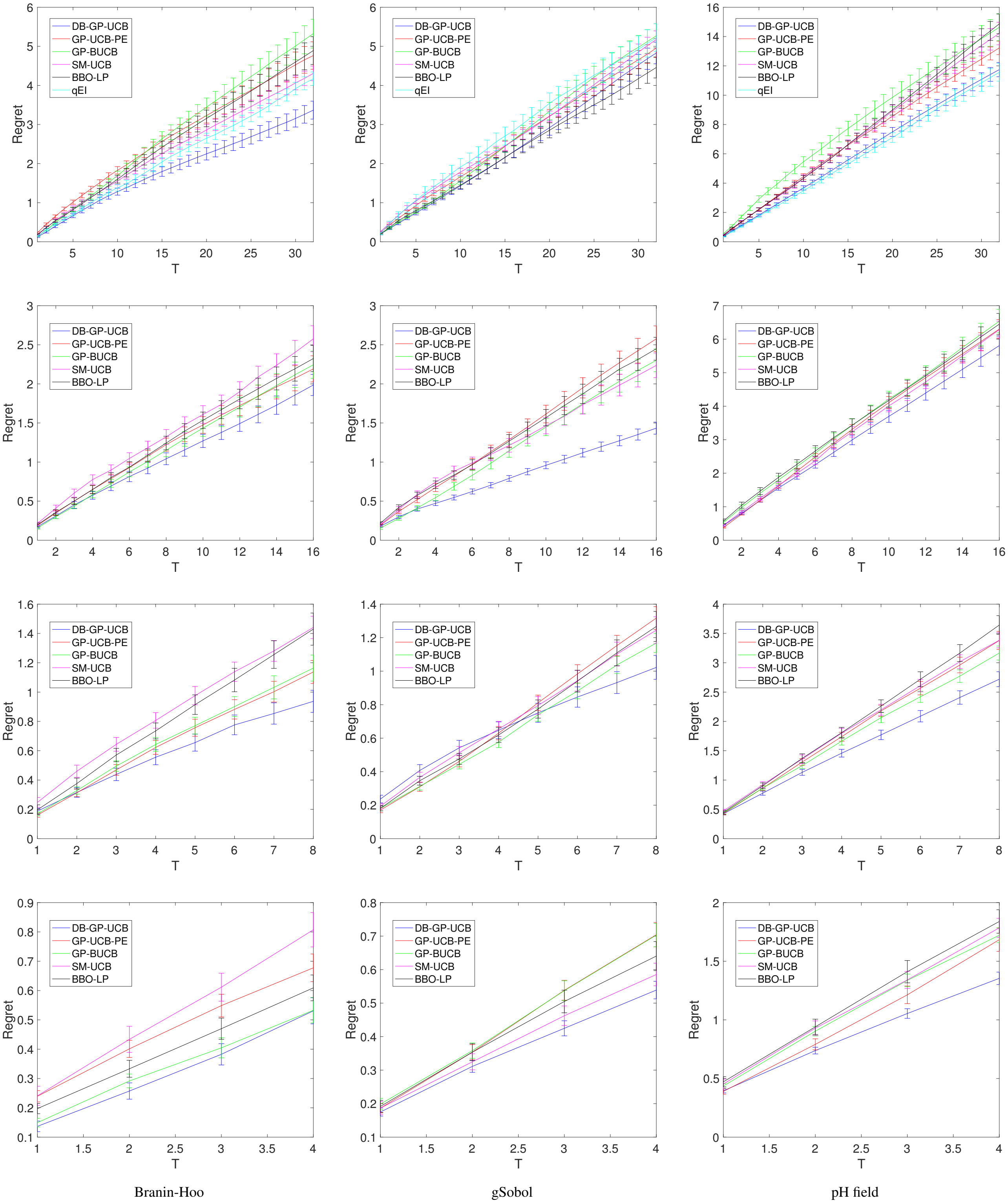

Figure 5. Cumulative regret incurred by tested algorithms with varying batch sizes $|\mathcal{D}_T| = 2, 4, 8, 16$ (rows from top to bottom) using a fixed budget of $T|\mathcal{D}_T| = 64$ function evaluations for the Branin-Hoo function, gSobol function, and real-world pH field. The error bars denote the standard error.