Supplementary Material

On Lower Bounds for Standard and Robust Gaussian Process Bandit Optimization (ICML 2021)

Xu Cai and Jonathan Scarlett

A. Formal Statements of Existing Results

In this section, we provide a more detailed overview of existing regret bounds in the literature. While our main focus is on lower bounds, we also state several existing upper bounds for comparison purposes. A particularly well-known upper bound from (Srinivas et al., 2010) is expressed in terms of the *maximum information gain*, defined as follows:

$$\gamma_t = \max_{\mathbf{x}_1,\cdots,\mathbf{x}_t} \frac{1}{2}\log\det(\mathbf{I}_t + \sigma^{-2}\mathbf{K}_t), \tag{27}$$

where $\mathbf{K}_t$ is a $t \times t$ kernel matrix with (i,j)-th entry $k(\mathbf{x}_i, \mathbf{x}_j)$. It was established in (Srinivas et al., 2010) that $\gamma_T = O((\log T)^{d+1})$ for the SE kernel, and $\gamma_T = O(T^{\frac{d(d+1)}{2\nu+d(d+1)}}\log T)$ for the Matérn-ν kernel, and we outline recent improvements on these bounds in Section A.4.

A.1. Standard Setting

We first state a standard cumulative regret upper bound (Srinivas et al., 2010; Chowdhury & Gopalan, 2017) and its straightforward adaptation to simple regret (e.g., see (Bogunovic et al., 2020, App. C)).

Theorem 5. (Simple Regret Upper Bound – Standard Setting (Srinivas et al., 2010; Chowdhury & Gopalan, 2017)) *Fix $\epsilon > 0$, $B > 0$, $T \in \mathbb{Z}$, and $\delta \in (0,1)$, and suppose that*

$$\frac{T}{\beta_T \gamma_T} \geq \frac{C_1}{\epsilon^2}, \tag{28}$$

where $C_1 = 8/\log(1+\sigma^{-2})$ and $\beta_T = \left(B + \sigma\sqrt{2(\gamma_{T-1} + \log\frac{e}{\delta})}\right)^2$. Then, there exists an algorithm that, for any $f \in \mathcal{F}_k(B)$, returns $\mathbf{x}^{(T)}$ satisfying $r(\mathbf{x}^{(T)}) \leq \epsilon$ with probability at least $1-\delta$.

Theorem 6. (Cumulative Regret Upper Bound – Standard Setting (Srinivas et al., 2010; Chowdhury & Gopalan, 2017)) *Fix $B > 0$, $T \in \mathbb{Z}$, and $\delta \in (0,1)$, and let $C_1 = 8/\log(1+\sigma^{-2})$ and $\beta_T = \left(B + \sigma\sqrt{2(\gamma_{T-1} + \log\frac{e}{\delta})}\right)^2$. Then, there exists an algorithm such that, for any $f \in \mathcal{F}_k(B)$, we have $R_T \leq \sqrt{C_1 T \beta_T \gamma_T}$ with probability at least $1-\delta$.*

The following lower bounds were proved in (Scarlett et al., 2017).

Theorem 7. (Simple Regret Lower Bound – Standard Setting (Scarlett et al., 2017, Thm. 1)) *Fix $\epsilon \in \left(0, \frac{1}{2}\right)$, $B > 0$, and $T \in \mathbb{Z}$. Suppose there exists an algorithm that, for any $f \in \mathcal{F}_k(B)$, achieves average simple regret $\mathbb{E}[r(\mathbf{x}^{(T)})] \leq \epsilon$. Then, if $\frac{\epsilon}{B}$ is sufficiently small, we have the following:*

1. *For $k = k_{\text{SE}}$, it is necessary that*
$$T = \Omega\left(\frac{\sigma^2}{\epsilon^2}\left(\log\frac{B}{\epsilon}\right)^{d/2}\right). \tag{29}$$
2. *For $k = k_{\text{Matérn}}$, it is necessary that*
$$T = \Omega\left(\frac{\sigma^2}{\epsilon^2}\left(\frac{B}{\epsilon}\right)^{d/\nu}\right). \tag{30}$$

Here, the implied constants may depend on (d, l, ν).

Theorem 8. (Cumulative Regret Lower Bound – Standard Setting (Scarlett et al., 2017, Thm. 2)) *For fixed $T \in \mathbb{Z}$ and $B > 0$, given any algorithm, we have the following:*

1. *For $k = k_{SE}$, there exists $f \in \mathcal{F}_k(B)$ such that*

$$\mathbb{E}[R_T] = \Omega\left(\sqrt{T\sigma^2\left(\log\frac{B^2T}{\sigma^2}\right)^d}\right) \tag{31}$$

provided that $\frac{\sigma}{B} = O\left(\sqrt{T}\right)$ with a sufficiently small implied constant.

2. *For $k = k_{Matérn}$, there exists $f \in \mathcal{F}_k(B)$ such that*

$$\mathbb{E}[R_T] = \Omega\left(B^{\frac{d}{2\nu+d}}\sigma^{\frac{2\nu}{2\nu+d}}T^{\frac{\nu+d}{2\nu+d}}\right) \tag{32}$$

provided that $\frac{\sigma}{B} = O\left(\sqrt{T^{\frac{1}{2+d/\nu}}}\right)$ with a sufficiently small implied constant.

Here, the implied constants may depend on (d, l, ν).

Remark 2. While Theorems 7 and 8 are stated in terms of the average regret, it is also noted in (Scarlett et al., 2017, Sec. 5.4) that the same scaling laws hold for regret bounds that are required to hold with a fixed constant probability above $\frac{3}{4}$. However, even when this probability is taken to approach one, the scaling of the lower bound therein remains unchanged, i.e., the dependence on the error probability is not characterized. We provide refined bounds characterizing this dependence in Theorems 1 and 2.

The function class used in the proofs of the above results is illustrated in Figure 1. As discussed in (Scarlett et al., 2017), the upper and lower bounds are near-matching for the SE kernel, only differing in the constant multiplying d in the exponent. The gaps are more significant for the Matérn kernel when relying on the bounds on γ_T from (Srinivas et al., 2010); however, in Section A.4, we overview some recent improved upper bounds that significantly close these gaps.

A.2. Robust Setting – Corrupted Samples

In the robust setting with corrupted samples described in Section 4.6, the following results were proved in (Bogunovic et al., 2020).

Theorem 9. (Upper Bound – Corrupted Samples (Bogunovic et al., 2020, Thm. 5)) *In the setting of corrupted samples with corruption threshold C, RKHS norm bound B, and time horizon T, there exists an algorithm (assumed to have knowledge of C) that, with probability at least $1-\delta$, attains cumulative regret $R_T = \mathcal{O}\big(\big(B + C + \sqrt{\ln(1/\delta)}\big)\sqrt{\gamma_T T} + \gamma_T\sqrt{T}\big)$.*

Theorem 10. (Lower Bound – Corrupted Samples (Bogunovic et al., 2020, App. J)) *In the setting of corrupted samples with corruption threshold C, if the RKHS norm B exceeds some universal constant, then for any algorithm, there exists a function $f \in \mathcal{F}_k(B)$ that incurs $\Omega(C)$ cumulative regret with probability arbitrarily close to one for any time horizon $T \geq C$.*

Note that Theorem 9 concerns the case that C is known. Additional upper bounds for the case of unknown C are also given in (Bogunovic et al., 2020), but we focus on the known C case; this is justified by the fact that we are focusing on lower bounds, and any given lower bound is stronger when it also applies to algorithms knowing C. Having said this, in future work, it may be interesting to determine whether the case of unknown C is provably harder; this is partially addressed in (Bogunovic et al., 2021) in the linear bandit setting.

Although Theorem 10 shows that the linear dependence on C is unavoidable, characterizing the optimal *joint* dependence on C and T is very much an open problem, as highlighted in (Bogunovic et al., 2020). Letting $\overline{R}_T^{(0)}$ and $\underline{R}_T^{(0)}$ be generic upper and lower bounds on the cumulative regret in the uncorrupted setting, we see that Theorem 9 is of the form $O\big(C\overline{R}_T^{(0)}\big)$ (multiplicative dependence on C), whereas Theorem 10 implies a lower bound of $\Omega\big(\underline{R}_T^{(0)} + C\big)$ (additive dependence on C).

A similar gap briefly existed in the standard multi-armed bandit problem (Lykouris et al., 2018), but was closed in (Gupta et al., 2019), in which the additive dependence was shown to be tight. However, the techniques for attaining a matching upper bound do not appear to extend easily to the RKHS setting. In Section 4.6, we show that, at least for deterministic algorithms, a fully additive dependence is in fact impossible.

A.3. Robust Setting – Corrupted Final Point

In the robust setting with a corrupted final point described in Section 4.7, the following results were proved in (Bogunovic et al., 2018a).

Theorem 11. (Upper Bound – Corrupted Final Point (Bogunovic et al., 2018a, Thm. 1)) *Fix $\xi > 0$, $\epsilon > 0$, $B > 0$, $T \in \mathbb{Z}$, $\delta \in (0, 1)$, and a distance function $\mathrm{dist}(\mathbf{x}, \mathbf{x}')$, and suppose that*

$$\frac{T}{\beta_T \gamma_T} \geq \frac{C_1}{\epsilon^2}, \tag{33}$$

where $C_1 = 8/\log(1 + \sigma^{-2})$ and $\beta_T = \big(B + \sigma\sqrt{2(\gamma_{T-1} + \log\frac{e}{\delta})}\big)^2$. Then, there exists an algorithm that, for any $f \in \mathcal{F}_k(B)$, returns $\mathbf{x}^{(T)}$ satisfying $r_\xi(\mathbf{x}^{(T)}) \leq \epsilon$ with probability at least $1 - \delta$.

Theorem 12. (Lower Bound – Corrupted Final Point (Bogunovic et al., 2018a, Thm. 2)) *Fix $\xi \in \big(0, \frac{1}{2}\big)$, $\epsilon \in \big(0, \frac{1}{2}\big)$, $B > 0$, and $T \in \mathbb{Z}$, and set $\mathrm{dist}(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_2$. Suppose that there exists an algorithm that, for any $f \in \mathcal{F}_k(B)$, reports a point $\mathbf{x}^{(T)}$ achieving ξ-regret $r_\xi(\mathbf{x}^{(T)}) \leq \epsilon$ with probability at least $1 - \delta$. Then, provided that $\frac{\epsilon}{B}$ and δ are sufficiently small, we have the following:*

1. *For $k = k_{\mathrm{SE}}$, it is necessary that $T = \Omega\big(\frac{\sigma^2}{\epsilon^2}\big(\log\frac{B}{\epsilon}\big)^{d/2}\big)$.*
2. *For $k = k_{\text{Matérn}}$, it is necessary that $T = \Omega\big(\frac{\sigma^2}{\epsilon^2}\big(\frac{B}{\epsilon}\big)^{d/\nu}\big)$.*

Here, the implied constants may depend on (ξ, d, l, ν).

For the SE kernel, (33) holds with $T = O^*\big(\frac{1}{\epsilon^2}\big(\log\frac{1}{\epsilon}\big)^{2d}\big)$ for constant B and σ^2 (Bogunovic et al., 2018a), where $O^*(\cdot)$ hides dimension-independent log factors. Thus, the upper and lower bounds nearly match. A detailed treatment of the Matérn kernel is deferred to Appendix A.4.

The function class used in the proof of Theorem 12 is illustrated in Figure 3. In contrast to the standard setting, where the difficulty of the class used (depicted in Figure 1) was in narrowing down the main (positive) peak, here the difficulty is in *avoiding* any point that can be perturbed into a *negative* valley.

While such an approach suffices for proving Theorem 12, it has the significant drawback that an algorithm that returns a *completely random* point (even with $T = 0$) has a fairly high chance of being within ϵ of optimal. That is, Theorem 12 only gives a hardness result for the case that the algorithm is required to succeed with probability sufficiently close to one (i.e., δ is sufficiently small).

This problem is exacerbated further in higher dimensions and/or for smaller values of ξ. To see this, note that the "bad" region (gray area in Figure 3) has volume proportional to ξ^d, which may be very close to zero (whereas the volume of the domain $[0, 1]^d$ is one for any d). In light of this limitation, it would be preferable to have a hardness result associated with any algorithm that succeeds with a universal constant probability, rather than only those that succeed with very high probability depending on ξ and d. In Section 4.7, we present a refined bound that addresses this exact issue.

A.4. Further Existing Upper Bounds for the Matérn Kernel

When comparing the lower bounds from (Scarlett et al., 2017) with the upper bounds from (Srinivas et al., 2010), the gaps are relatively small for the SE kernel, e.g., $O^*\big(\sqrt{T(\log T)^{2d}}\big)$ (Srinivas et al., 2010) vs. $\Omega\big(\sqrt{T(\log T)^{d/2}}\big)$ (Scarlett et al., 2017) for the cumulative regret.[8] In contrast, the gaps are more significant for the Matérn kernel, e.g., $O^*\big(T^{\frac{1}{2}\cdot\frac{2\nu+3d(d+1)}{2\nu+d(d+1)}}\big)$ (Srinivas et al., 2010) vs. $O^*\big(T^{\frac{\nu+d}{2\nu+d}}\big)$ (Scarlett et al., 2017), with the former in fact failing to be sub-linear in T unless $2\nu - d(d+1) > 0$. In the following, we outline some more recent results and observations that significantly close these gaps for the Matérn kernel. We focus our discussion on the cumulative regret (with constant values of B and σ^2), but except where stated otherwise, similar observations apply in the case of simple regret.

Recently, (Janz et al., 2020) gave the first practical algorithm (referred to as π-GP-UCB) to provably attain sublinear regret for all $\nu > 1$ and $d \geq 1$ in the RKHS setting. Specifically, the regret bound is $O^*\big(T^{\frac{d(2d+3)+2\nu}{d(2d+4)+4\nu}}\big)$. It was also pointed out

[8] The published version of (Scarlett et al., 2017) mistakenly omits the division by two in the exponent; see https://arxiv.org/abs/1706.00090v3 for a corrected version.

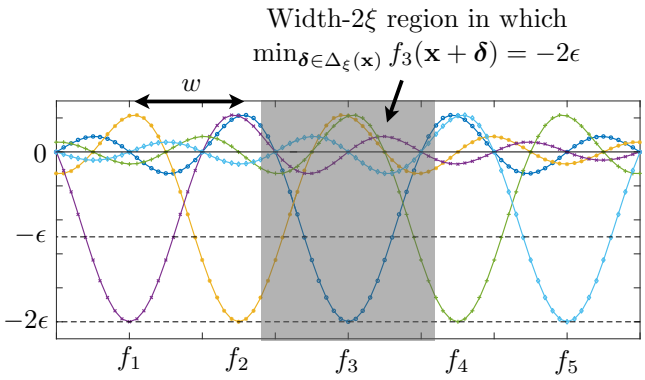

Figure 3. Illustration of functions $f_1, \dots, f_5$ equal to a common function shifted by various multiples of a given parameter w. In the ξ-stable setting, there is a wide region (shown in gray for the dark blue curve f_3) within which the perturbed function value equals -2ϵ.

in (Janz et al., 2020) that the SupKernelUCB algorithm of (Valko et al., 2013) can be extended to continuous domains via a fine discretization; the dependence on the number of arms N is logarithmic, so even a very fine discretization of the space with $\frac{1}{\mathrm{poly}(T)}$ spacing only amounts to a $\log N = O(d \log T)$ term. With this approach, one attains a cumulative regret of $R_T = O^*(\sqrt{T\gamma_T})$ (assuming d to be constant); in contrast with the result in (Srinivas et al., 2010), this yields $R_T = o(T)$ whenever $\gamma_T = o(T)$ (or more precisely, whenever γ_T is sufficiently sublinear to overcome any hidden logarithmic factors). Despite its significant theoretical value, SupKernelUCB has been observed to perform poorly in practice (Calandriello et al., 2019).

Another recent work (Shekhar & Javidi, 2020) gave an algorithm whose regret bounds further improve on those of (Janz et al., 2020). The algorithm is again impractical due to the large constant factors, though a practical heuristic is also given. Unlike the other works outlined above, the simple regret and cumulative regret behave very differently, due to the exploratory nature of the algorithm. Defining $\mathcal{I}_0 = (0, 1]$, $\mathcal{I}_1 = \left(1, \frac{d(d+1)}{2}\right]$, $\mathcal{I}_2 = \left(\frac{d(d+1)}{2}, \frac{d^2+5d+12}{4}\right]$, and $\mathcal{I}_3 = (0, \infty) \setminus (\mathcal{I}_0 \cup \mathcal{I}_1 \cup \mathcal{I}_2)$, the simple regret bounds of (Shekhar & Javidi, 2020) are summarized as follows:

- For $\nu \in \mathcal{I}_0 \cup \mathcal{I}_1$ (i.e., $\nu \leq \frac{d(d+1)}{2}$), one has $r(\mathbf{x}^{(T)}) = O^*\left(T^{\frac{-\nu}{2\nu+d}}\right)$, which matches the lower bound of (Scarlett et al., 2017);
- For $\nu \in \mathcal{I}_2$ (which is only possible for $d \leq 5$), one has $r(\mathbf{x}^{(T)}) = O^*\left(T^{\frac{-1}{d+2}}\right)$;
- For $\nu \in \mathcal{I}_3$, one has $r(\mathbf{x}^{(T)}) = O^*\left(T^{-\frac{1}{2}+\frac{d(d+3)}{4\nu+d(d+5)}}\right)$.

In addition, the cumulative regret bounds of (Shekhar & Javidi, 2020) are summarized as follows:

- For $\nu \in \mathcal{I}_0$ (i.e., $\nu \leq 1$), one has $R_T = O^*(T^{\frac{\nu+d}{2\nu+d}})$, which matches the lower bound of (Scarlett et al., 2017).
- For $\nu \in \mathcal{I}_1 \cup \mathcal{I}_2$, one has $R_T = O^*(T^{\frac{d+1}{d+2}})$;
- For $\nu \in \mathcal{I}_3$, one has $R_T = O^*\left(T^{\frac{1}{2}+\frac{d(d+3)}{4\nu+d(d+5)}}\right)$.

Note that all of the preceding bounds are high-probability bounds (e.g., holding with probability 0.99).

Finally, in a very recent work (Vakili et al., 2021), improved bounds on the information gain γ_T were given, in particular yielding $\gamma_T = O^*\left(T^{\frac{d}{2\nu+d}}\right)$ for the Matérn kernel with $\nu > \frac{1}{2}$. When combined with the above-mentioned $R_T = O^*(\sqrt{T\gamma_T})$ upper bound, this in fact yields matching upper and lower regret bounds for the Matérn kernel, up to logarithmic factors.

With the above outline in place, we are now in a position to explain the Matérn kernel upper bounds shown in Table 1:

- For the standard setting, we first apply the above-mentioned upper bound $R_T = O^*(\sqrt{T\gamma_T})$ for SupKernelUCB, which also has a multiplicative $\sqrt{\log \frac{1}{\delta}}$ dependence on the target error probability δ (Valko et al., 2013). Substituting $\gamma_T = O^*\big(T^{\frac{d}{2\nu+d}}\big)$ gives the desired bound, $R_T = O^*\big(T^{\frac{\nu+d}{2\nu+d}}\sqrt{\log \frac{1}{\delta}}\big)$.

- For the setting of corrupted samples, we substitute $\gamma_T = O^*\big(T^{\frac{d}{2\nu+d}}\big)$ and $\delta = \Theta(1)$ into Theorem 9. Notice that the term containing C is only multiplied by $\sqrt{\gamma_T}$, rather than γ_T.

- For the setting of a corrupted final point, we first substitute $\gamma_T = O^*\big(T^{\frac{d}{2\nu+d}}\big)$ into Theorem 11. We then note that $\beta_T = O(\max\{\gamma_T, \log \frac{1}{\delta}\})$, and treat two cases separately depending on which term attains the maximum. The case that $\beta_T = O(\gamma_T)$ leads to the term $O^*\big(\big(\frac{1}{\epsilon}\big)^{\frac{2(2\nu+d)}{2\nu-d}}\big)$ (and requires $d < 2\nu$ to obtain a non-vacuous statement in (33)), and the case that $\beta_T = O\big(\log \frac{1}{\delta}\big)$ leads to the term $O^*\big(\big(\frac{\log \frac{1}{\delta}}{\epsilon^2}\big)^{1+\frac{d}{2\nu}}\big)$.

Finally, we briefly note that we could similarly slightly improve the SE kernel upper bounds containing δ in Table 1 by treating two cases separately similarly to last item above; however, for the SE kernel this only amounts to minor differences. Specifically, the standard cumulative regret from (Chowdhury & Gopalan, 2017) is $O^*\big(\sqrt{T(\log T)^{2d} + T(\log T)^d \log \frac{1}{\delta}}\big)$, and the time to ϵ-optimality from (Bogunovic et al., 2018a) is $O^*\big(\frac{1}{\epsilon^2}\big(\log \frac{1}{\epsilon}\big)^{2d} + \frac{1}{\epsilon^2}\big(\log \frac{1}{\epsilon}\big)^d \log \frac{1}{\delta}\big)$.

A.5. Other Settings

The following works are more distinct from ours, so we only give a brief outline:

- The noiseless setting with RKHS functions was studied in (Bull, 2011; Lyu et al., 2019; Vakili et al., 2020). In particular, (Bull, 2011) characterized the minimax-optimal simple regret for the Matérn kernel, showing that $O\big(\big(\frac{1}{\epsilon}\big)^{d/\nu}\big)$ is the optimal time to ϵ-optimality. The papers (Lyu et al., 2019; Vakili et al., 2020) also considered the cumulative regret, which can be considerably smaller compared to the noisy setting (e.g., (Vakili et al., 2020) shows that $O(1)$ cumulative regret is possible when $\nu > d$).

- A counterpart to the non-Bayesian RKHS setting is the Bayesian setting, in which f is assumed to be randomly drawn from a Gaussian process. Regret bounds for such settings have also been extensively studied, and the regret can again be much smaller in the noiseless setting (Grünewälder et al., 2010; De Freitas et al., 2012; Kawaguchi et al., 2015; Grill et al., 2018). In the noisy setting, the upper bounds of (Srinivas et al., 2010) are dictated by $\sqrt{T\gamma_T}$, in contrast with the higher $\sqrt{T\gamma_T^2}$ term in the RKHS setting. Further improvements over (Srinivas et al., 2010) are given in (Shekhar & Javidi, 2018; Scarlett, 2018; Wang et al., 2020), including broad scenarios in which the order-optimal scaling of the cumulative regret is roughly $\sqrt{T}$.

- Follow-up works to (Bogunovic et al., 2018a) include (i) investigations of mixed strategies (Sessa et al., 2020) with an average function value requirement, and (ii) distributional robustness in place of the simple perturbation model (Kirschner et al., 2020; Nguyen et al., 2020). In addition, earlier works considered related notions of *input noise* in Bayesian optimization (Nogueira et al., 2016; Beland & Nair, 2017; Dai Nguyen et al., 2017). Other robustness notions in GP optimization include outliers (Martinez-Cantin et al., 2018), batch settings with failed experiments (Bogunovic et al., 2018b), and heavy-tailed noise (Chowdhury & Gopalan, 2019). The latter of these provides lower bounds adapted from the techniques of (Scarlett et al., 2017), indicating that our alternative techniques based on Lemma 1 could also be of interest in such a setting.

B. Proof of Theorem 3 (Corrupted Samples)

The analysis re-uses some aspects of the proof of Theorem 2 for the standard setting, given in Section 4. We consider the function class $\mathcal{F} = \{f_1, \dots, f_M\}$ from Section 4.1, with a parameter $\epsilon > 0$ to be chosen later. As mentioned in Section 2.2, for a deterministic algorithm, the adversary knows which action is played each round. Hence, we can consider an adversary that trivially pushes the function value $f(\mathbf{x}_t)$ down to zero, at a cost of $|c_t(\mathbf{x}_t)| = f(\mathbf{x}_t)$, until its budget does not allow doing so. When the remaining budget does not allow pushing the function value down to zero, the adversary pushes the value as close to zero as possible, after which its budget is exhausted and no further corruptions occur.

Since we are considering the noiseless setting, the player simply observes $y_1 = y_2 = \ldots = y_t = 0$ for any time t before the adversary exhausts its budget. In particular, if the adversary still has not exhausted its budget by time T, then we have $y_1 = y_2 = \ldots = y_T = 0$. Intuitively, in this case, since the player has not observed any function values, it cannot know where the function peak is. Since any sample away from the peak incurs regret $\Theta(\epsilon)$ by construction, this leads to $\Omega(T\epsilon)$ regret. We note that ideas with similar intuition have also appeared in simpler robust bandit problems, including the finite-arm setting (Lykouris et al., 2018, Sec. 5) and the case of linear rewards (Bogunovic et al., 2021).

To make this intuition precise, we first introduce the following terminology: For a given set of sampled points $\mathbf{x}_1, \ldots, \mathbf{x}_T$ up to time T, define a given function $\tilde{f} \in \mathcal{F}$ to be *corruptible after time* T if $\sum_{t=1}^{T} |\tilde{f}(\mathbf{x}_t)| < C$, and *non-corruptible after time* T otherwise. Then, we have the following lemma.

Lemma 5. *Suppose that $T = \frac{\alpha C M}{\epsilon}$ for some sufficiently small constant $\alpha > 0$. Then, under the preceding setup, for any set of sampled points $\mathbf{x}_1, \ldots, \mathbf{x}_T$, there exists a set of $\frac{M}{2}$ functions among $\{f_j\}_{j=1}^{M}$ that are corruptible after time T, i.e., $\sum_{t=1}^{T} |f_j(\mathbf{x}_t)| < C$.*

Proof. We will upper bound the average of $\sum_{t=1}^{T} |f_j(\mathbf{x}_t)|$ over all values $j \in \{1, \ldots, M\}$, and then use Markov's inequality to establish the existence of the $\frac{M}{2}$ values of j in the lemma statement. First, for any *fixed* time index t and point $\mathbf{x}_t$, the second part of Lemma 3 implies that

$$\sum_{m=1}^{M} |f_m(\mathbf{x}_t)| = O(\epsilon). \tag{34}$$

Renaming m to j, summing both sides over $t = 1, \ldots, T$, and dividing both sides by M, it follows that

$$\frac{1}{M} \sum_{j=1}^{M} \Bigg(\sum_{t=1}^{T} |f_j(\mathbf{x}_t)| \Bigg) = O\Big(\frac{T\epsilon}{M}\Big) = O(\alpha C) \leq \frac{C}{4}, \tag{35}$$

where the first equality uses the assumption $T = \frac{\alpha C M}{\epsilon}$, and the second equality uses the assumption that α is sufficiently small.

As hinted above, interpreting the left-hand side of (35) as an average of $\sum_{t=1}^{T} |f_m(\mathbf{x}_t)|$ with respect to j, Markov's inequality implies that we can only have $\sum_{t=1}^{T} |f_j(\mathbf{x}_t)| \geq C$ for at most a $\frac{1}{4}$ fraction of the j values in $\{1, \ldots, M\}$. Thus, at least $\frac{M}{2}$ of the functions give $\sum_{t=1}^{T} |f_j(\mathbf{x}_t)| < C$, as desired.

□

Since we are considering deterministic algorithms, we have that when observing $y_1 = 0$, $y_2 = 0$, and so on, any algorithm can only follow a fixed corresponding sequence $\mathbf{x}_1$, $\mathbf{x}_2$, and so on (until a non-zero y_t value is observed). However, Lemma 5 implies that no matter which such fixed sequence is chosen, there exist $\frac{M}{2}$ functions under which the adversary is able to continue corrupting $y_1 = y_2 = \ldots = y_T = 0$ up until the final point T. Since the function class is such as that any given point is ϵ-optimal for at most one function, it follows that the algorithm can only attain $R_T = o(T\epsilon)$ for at most one of these $\frac{M}{2}$ functions; all of the others must incur $R_T = \Omega(T\epsilon)$.

To make the $\Omega(T\epsilon)$ lower bound explicit, we need to select ϵ as a function of T. However, such a choice must be consistent with two assumptions already made in the above analysis: (i) $T = \frac{\alpha C M}{\epsilon}$ for some sufficiently small $\alpha = \Theta(1)$ in Lemma 5, and (ii) the function class in Section 4.1 is such that M satisfies (15) (SE kernel) or (16) (Matérn) kernel. Handling the two kernels separately, we proceed as follows:

- For the SE kernel, substituting (15) into $T = \frac{\alpha C M}{\epsilon}$ yields $T = \Theta\big(\frac{C}{\epsilon}\big(\log\frac{1}{\epsilon}\big)^{d/2}\big)$, and using the assumptions $\Theta(1) \leq C \leq T^{1-\Omega(1)}$ and $d = \Theta(1)$, an inversion of this expression (detailed in Section B.1 below) gives $\epsilon = \Theta\big(\frac{C}{T}(\log T)^{d/2}\big)$. Hence, we have $R_T = \Omega(T\epsilon) = \Omega\big(C(\log T)^{d/2}\big)$.

- For the Matérn kernel, substituting (16) into $T = \frac{\alpha C M}{\epsilon}$ yields $T = \Theta\big(\frac{C}{\epsilon}\big(\frac{1}{\epsilon}\big)^{d/\nu}\big)$, and inverting this gives $\epsilon = \Theta\big(\big(\frac{C}{T}\big)^{\frac{1}{1+d/\nu}}\big) = \Theta\big(\big(\frac{C}{T}\big)^{\frac{\nu}{d+\nu}}\big)$. Hence, we have $R_T = \Omega(T\epsilon) = \Omega\big(C^{\frac{\nu}{d+\nu}} T^{\frac{d}{d+\nu}}\big)$.

This completes the proof of Theorem 3.

We remark that since this proof considers the noiseless setting (i..e, $\sigma^2 = 0$), it may be interesting to establish whether the arguments can be refined for the noisy setting in order to obtain an improved lower bound on R_T.

B.1. Final Inversion Step for the SE Kernel

We first note that the scaling $T = \Theta\big(\frac{C}{\epsilon}\big(\log\frac{1}{\epsilon}\big)^{d/2}\big)$ is equivalent to

$$\frac{T}{C} = \Theta\bigg(\frac{1}{\epsilon}\Big(\log\frac{1}{\epsilon}\Big)^{d/2}\bigg). \tag{36}$$

We proceed by taking the logarithm on both sides. For the left-hand side, the assumption $\Theta(1) \le C \le T^{1-\Omega(1)}$ implies $\log\frac{T}{C} = \Theta(\log T)$. In addition, the assumption $d = \Theta(1)$ implies that the logarithm of the right-hand side of (36) behaves as $\Theta\big(\log\frac{1}{\epsilon} + \frac{d}{2}\log\log\frac{1}{\epsilon}\big) = \Theta\big(\log\frac{1}{\epsilon}\big)$ (note that $\epsilon \in (0,1)$, so $\frac{1}{\epsilon} > 1$). Hence, overall, taking the logarithm on both sides of (36) gives $\log\frac{1}{\epsilon} = \Theta(\log T)$. Substituting this finding into (36) gives $\frac{T}{C} = \Theta\big(\frac{1}{\epsilon}(\log T)^{d/2}\big)$, and re-arranging gives $\epsilon = \Theta\big(\frac{C}{T}(\log T)^{d/2}\big)$ as claimed.

C. Proof of Theorem 4 (Corrupted Final Point)

We continue the proof following the intuition provided for the idealized function class in Section 4.7.

C.1. Details for the Matérn Kernel

We seek to provide a function class that captures the essential properties of the idealized version, while ensuring that every function in the class has RKHS norm at most B under the Matérn kernel. Recall that Theorem 4 assumes that $\frac{\epsilon}{B}$ is sufficiently small.

To construct a given function f_m of the form in Figure 2, we will decompose it as

$$f_m(\mathbf{x}) = -c(\mathbf{x}) + b(\mathbf{x}) - s_m(\mathbf{x}), \tag{37}$$

where $c(\cdot)$ is a constant function equaling -2ϵ across the whole domain $[0,1]^d$, $b(\cdot)$ approximates the indicator function (scaled by 2ϵ) of being within a ball ($d \ge 2$) or interval ($d = 1$) of diameter roughly 3ξ at the center of the domain (see below for details), and $s_m(\mathbf{x})$ is the narrow spike whose location is determined by $m \in \{1,\dots,M\}$ (whereas $s_0(\mathbf{x}) = 0$ for all $\mathbf{x}$). We proceed by showing that suitable functions can be constructed having RKHS norm at most $\frac{B}{3}$ each, so that the triangle inequality applied to (37) yields $\|f_m\|_k \le B$.

For convenience, we first work with auxiliary functions centered at the origin, before shifting them to be centered at a suitable point in $[0,1]^d$.

Lemma 6. *Let k be the Matérn-ν kernel, let $r > 0$ and $0 < w_0 \le \frac{r}{2}$ be fixed constants, and let $\epsilon > 0$ and $B > 0$ be such that $\frac{\epsilon}{B}$ is sufficiently small. There exists a function $b_0(\mathbf{x})$ on $\mathbb{R}^d$ satisfying (i) $b_0(\mathbf{x}) = 2\epsilon$ whenever $\|\mathbf{x}\|_2 \le r - w_0$; (ii) $b_0(\mathbf{x}) = 0$ whenever $\|\mathbf{x}\|_2 \ge r + w_0$; (iii) $b_0(\mathbf{x}) \in [0, 2\epsilon]$ whenever $r - w_0 \le \|\mathbf{x}\|_2 \le r + w_0$; and (iv) $\|b_0\|_k \le \frac{B}{3}$.*

Proof. Define the auxiliary "ball" function with radius $r > 0$ as

$$\mathrm{ball}(\mathbf{x}) = \mathbb{1}\{\|\mathbf{x}\|^2 \le r\}, \tag{38}$$

and for fixed $w_0 > 0$, let $g_0(\mathbf{x}) = h\big(\frac{\mathbf{x}}{w_0}\big)$ be a scaled version of the bump function $h(\cdot)$ from Lemma 4, and define

$$\tilde{b}_0(\mathbf{x}) = (g_0 \star \mathrm{ball})(\mathbf{x}), \tag{39}$$

where $\star$ denotes the convolution operation. By the definition of convolution and the fact that $g_0(\mathbf{x})$ is non-zero only for $\|\mathbf{x}\|_2 \le w_0$, we have the following:

- For $\mathbf{x}$ satisfying $\|\mathbf{x}\|_2 \ge r + w_0$, we have $\tilde{b}_0(\mathbf{x}) = 0$;

- For $\mathbf{x}$ satisfying $\|\mathbf{x}\|_2 \leq r - w_0$, we have $\tilde{b}_0(\mathbf{x}) = \int_{\mathbb{R}^d} g_0(\mathbf{x})\mathrm{d}\mathbf{x}$, which is a constant depending on w_0.
- For $\mathbf{x}$ satisfying $r - w_0 \leq \|\mathbf{x}\|_2 \leq r + w_0$, we have that $\tilde{b}_0(\mathbf{x})$ equals some value in between the two constants given in the previous two dot points.

We proceed by showing that $\|\tilde{b}_0\|_k < \infty$ under the Matérn kernel. We know from the proof of Lemma 4 that $\|g_0\|_k$ is a finite constant depending on w_0. As for $\mathrm{ball}(\cdot)$, it suffices for our purposes to note that its Fourier transform is bounded in absolute value (point-wise) by a constant depending on r, which is seen by writing

$$\Big| \int_{\mathbb{R}^d} \mathrm{ball}(\mathbf{x}) e^{\iota\langle \mathbf{x}, \boldsymbol{\xi}\rangle} \mathrm{d}\mathbf{x} \Big| \leq \int_{\|\mathbf{x}\|_2 \leq r} \mathrm{d}\mathbf{x} < \infty. \tag{40}$$

Then, using the formula for RKHS norm in Lemma 7, and using capital letters to denote the Fourier transforms of the respective spatial functions, we have

$$\|\tilde{b}_0\|_k = \int \frac{|G_0(\boldsymbol{\xi})|^2 \cdot |\mathrm{BALL}(\boldsymbol{\xi})|^2}{K(\boldsymbol{\xi})} d\boldsymbol{\xi} \tag{41}$$

$$\leq O(1) \cdot \int \frac{|G_0(\boldsymbol{\xi})|^2}{K(\boldsymbol{\xi})} d\boldsymbol{\xi} = O(1) \cdot \|g_0\|_k < \infty, \tag{42}$$

where (41) uses the fact that convolution in the spatial domain corresponds to multiplication in the Fourier domain, and (42) uses (40).

Finally, for any fixed r and w_0, we define $b_0(\mathbf{x})$ to be a constant times $\tilde{b}_0(\mathbf{x})$, with the constant chosen so that the maximum function value is 2ϵ. Since $\|\tilde{b}_0\|_k = O(1)$ and we scale by $O(\epsilon)$, it follows that $\|b_0\|_k = O(\epsilon)$, and thus $\|b_0\|_k \leq \frac{B}{3}$ due to the assumption that $\frac{\epsilon}{B}$ is sufficiently small. The remaining properties of $b_0(\cdot)$ in the lemma statement are directly inherited from those of $\tilde{b}_0(\cdot)$ above. □

We now construct the functions in (37) as follows for some arbitrarily small constant $\eta > 0$:

- For $c(\mathbf{x})$, let $r = \sqrt{d} + \eta$ and $w_0 = \eta$, so that $c(\mathbf{x}) = 1$ for all $\mathbf{x} \in [0,1]^d$;
- Let $b(\mathbf{x})$ be a shifted version (to be centered at $\left(\frac{1}{2}, \ldots, \frac{1}{2}\right)$) of the ball function $b_0(\mathbf{x})$ with $r = (3-\eta)\xi$ and $w_0 = \eta\xi$.
- For $m = 1, \ldots, M$, let $s_m(\mathbf{x})$ be a shifted version of the spike formed in Lemma 4, with RKHS norm $\frac{B}{3}$ in place of B.

While the radius of the "plain" region in Figure 2 (i.e., the region where points may have function value zero even after a worst-case perturbation) is not exactly ξ due to the "leeway" introduced by η, it is arbitrarily close when η is sufficiently small (e.g., 0.99ξ).

Using the assumption that $\frac{\epsilon}{B}$ is sufficiently small but ξ is constant, the choice of w in Lemma 4 means that we can assume w to be much smaller than ξ (e.g., a 0.1 fraction or less). As a result, a standard packing argument (Duchi, Sec. 13.2.3) reveals that we can "pack" at least $M = \left(\frac{c_0'\xi}{w}\right)^d$ bump functions into the sphere of radius roughly ξ (for some absolute constant c_0'), while ensuring that the supports of these functions are non-overlapping. Since ξ is assumed to be constant, this choice of M matches (14) up to a possible change in the value of c_0', and we conclude that the scaling (16) applies with a possibly different choice of c_3.

With this fact in place, we can proceed in the same way as Section 4.5. The "base function" and "auxiliary" function are chosen as $f(\mathbf{x}) = f_0(\mathbf{x})$ and $f'(\mathbf{x}) = f_{m'}(\mathbf{x})$ (for some $m' = 1, \ldots, M$), so that their difference is $s_{m'}(\mathbf{x})$ (since $s_0(\mathbf{x}) = 0$). Using (26) with the substitution $v_{m'}^{m'} \leftarrow 4\epsilon$ (i.e., the maximal value of $s_{m'}(\mathbf{x})$) and $N_{m'}(\tau)$ redefined to be the number of samples within the support of $s_{m'}(\mathbf{x})$, we obtain

$$\mathbb{E}_m[N_{m'}(\tau)] \cdot (4\epsilon)^2 \geq \frac{\sigma^2}{2} \log \frac{1}{2.4\delta}, \tag{43}$$

and summing over $m' = 1, \ldots, M$ gives

$$T \geq \frac{\sigma^2 M}{32\epsilon^2} \log \frac{1}{2.4\delta}. \tag{44}$$

Substituting the scaling on M in (16) (which we established also holds here) completes the proof.

C.2. Overview of Details for the SE Kernel

For the SE kernel, we follow the same argument as the Matérn kernel, but wherever the bump function from Lemma 4 is used, we replace it by the "approximate bump" function from Section 4.1. This creates a few more technical nuisances, but the argument is essentially the same, so we only outline the differences:

- In the analog of Lemma 6, the function value is not exactly constant in the "inner sphere", and is not exactly zero outside the "outer sphere", but it is arbitrarily close (e.g., at least 0.99 times the maximum in the former case, and below 0.01 times the maximum in the former case).
- In specifying the M functions, we no longer have disjoint supports, but we instead place the centers on a uniform grid, as was done in (Scarlett et al., 2017; Bogunovic et al., 2018a) and used in Section 4.1. Inside the "plain" sphere of radius 0.99ξ (see Figure 2), we can fit a cube of side-length $\frac{0.99\xi}{\sqrt{d}}$, and hence, the uniform grid still leads to M of the form $M = \big(\frac{c_0'\xi}{w}\big)^d$, albeit with a smaller constant c_0' depending on d.
- Once the function class with the grid-like structure is established, instead of following the simplified steps for the Matérn kernel in Section 4.5, we follow the slightly more involved (but still simple) steps from Section 4.2.

D. Further Auxiliary Lemmas

The following lemma states a well-known expression for the RKHS norm in terms of the Fourier transforms of the function and kernel.

Lemma 7. (Aronszajn, 1950, Sec. 1.5) *Consider an RKHS $\mathcal{H}$ for functions on $\mathbb{R}^d$, corresponding to a kernel of the form $k(\mathbf{x},\mathbf{x}') = k(r_{\mathbf{x},\mathbf{x}'})$ with $r_{\mathbf{x},\mathbf{x}'} = \mathbf{x}-\mathbf{x}'$, and let $K(\xi)$ be the d-dimensional Fourier transform of $k(\cdot)$. Then for any $f \in \mathcal{H}$ with Fourier transform $F(\xi)$, we have*

$$\|f\|_{\mathcal{H}} = \int \frac{|F(\xi)|^2}{K(\xi)} d\xi. \tag{45}$$

In addition, if $\mathcal{H}(D)$ is an RKHS on a compact subset $D \subseteq \mathbb{R}^d$ with the same kernel as $\mathcal{H}$, then we have for any $f \in \mathcal{H}(D)$ that

$$\|f\|_{\mathcal{H}(D)} = \inf_g \|g\|_{\mathcal{H}(\mathbb{R}^d)}, \tag{46}$$

where the infimum is over all functions $g \in \mathcal{H}(\mathbb{R}^d)$ that agree with f when restricted to D.

While the following lemma is not used in this paper, it is stated because it is a key component of the analysis in (Scarlett et al., 2017), and can thus be contrasted with the key lemma of our analysis (Lemma 1).

Lemma 8. (Auer et al., 1995) *For any function $a(\mathbf{y})$ taking values in a bounded range $[0, A]$, we have*

$$\big|\mathbb{E}_m[a(\mathbf{y})] - \mathbb{E}_0[a(\mathbf{y})]\big| \le A\, d_{\mathrm{TV}}(P_0, P_m) \tag{47}$$

$$\le A\sqrt{D(P_0\|P_m)}, \tag{48}$$

where $d_{\mathrm{TV}}(P_0,P_m) = \frac{1}{2}\int_{\mathbb{R}^T} |P_0(\mathbf{y}) - P_m(\mathbf{y})|\,\mathrm{d}\mathbf{y}$ is the total variation distance.

To simplify the final expression in Lemma 8, the divergence term therein can be further bounded using the following.

Lemma 9. (Scarlett et al., 2017, Eq. (44)) *Under the definitions in Section 4.1, we have*

$$D(P_0\|P_m) \le \sum_{j=1}^{M} \mathbb{E}_0[N_j]\overline{D}_m^j. \tag{49}$$