A Appendix

A.1 Proof of Theorem 2

. Our proof follows the following sketch. We show that the given problem can be written as a linear regression problem in the induced RKHS. The greedy SBQ algorithm to choose data points is then equivalent to forward greedy feature selection in the transformed space (Lemma 4). After the selection is made, the weight optimization obtained through the posterior calculation ensures orthogonal projection (Lemma 5) which means the posterior calculation is nothing but fitting of the least squares regression on the chosen set of features. Finally we draw upon research in discrete optimization to get approximation guarantees for greedy feature selection for least squares regression (Lemma 7) that we use to obtain the convergence rates.

We will require the following definition of the Maximum Mean Discrepancy (MMD). MMD is a divergence measure between two distributions p and q over a class of functions $\mathcal{F}$. We restrict our attention to cases when $\mathcal{F}$ is a Reproducing Kernel Hilbert Space (RKHS), which allows MMD evaluation based only on kernels, rather than explicit function evaluations.

$$\mathrm{MMD}_{\mathcal{F}}(p,q) := \sup_{f\in\mathcal{F}} \left| \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x} - \int f(\mathbf{x})q(\mathbf{x})d\mathbf{x} \right|$$

If the sup is reached at ϕ,

$$\begin{aligned}\mathrm{MMD}_{\mathcal{F}}(p,q)^2 &= \left| \int \phi(\mathbf{x})p(\mathbf{x})d\mathbf{x} - \int \phi(\mathbf{x})q(\mathbf{x})d\mathbf{x} \right|^2 \\ &= \|\mu_p - \mu_q\|^2_{\mathcal{F}},\end{aligned}$$

where μ_p and μ_q are the mean function mappings under p and q respectively.

We make use of the following lemma that establishes a connection between MMD and Bayesian Quadrature.

Lemma 4. *[12] Let q be the distribution established by weights w_i of the Bayesian Quadrature over the selected points. Then, the expected variance of the weighted sum in Bayesian Quadration* (2) *is equal to* $\mathrm{MMD}^2(p,q)$.

We can make this explicit in our notation. If $\mathcal{F}$ is an RKHS, we can write the MMD cost function using only the kernel function $K(\cdot,\cdot)$ associated with the RKHS [10] as:-

$$\begin{aligned}\mathrm{MMD}_{\mathcal{F}}(p,q)^2 &= \int_{\mathbf{x}\sim p}\int_{\mathbf{y}\sim p} K(\mathbf{x},\mathbf{y})p(\mathbf{x})p(\mathbf{y})d\mathbf{x}d\mathbf{y} \\ &\quad - 2 * \int_{\mathbf{x}\sim p}\int_{\mathbf{y}\sim q} K(\mathbf{x},\mathbf{y})p(\mathbf{x})q(\mathbf{y})d\mathbf{x}\mathbf{y} \\ &\quad + \int_{\mathbf{x}\sim q}\int_{\mathbf{y}\sim q} K(\mathbf{x},\mathbf{y})q(\mathbf{x})q(\mathbf{y})d\mathbf{x}\mathbf{y} \\ &= \|\mu_p - \sum_i w_i\phi(\mathbf{x}_i)\|^2_{\mathcal{F}},\end{aligned}$$

where, $\phi(\cdot)$ represents the feature mapping under the kernel function $K(\cdot,\cdot)$, and i ranges over the selected points that define our discrete distribution q. Recall that Bayesian Quadrature deviates from simple kernel herding by allowing for and optimizing over non-uniform weights w_i. We can formally show that the weight optimization obtained through the posterior calculation performs an orthogonal projection of μ_p onto the span of selected points to get μ_q in the induced kernel space.

Lemma 5. *The weights obtained w_i through the posterior evaluation of $Z(\mathsf{S}_n)$ guarantee that $\sum_i w_i\phi(\mathbf{x}_i)$ is the orthogonal projection of μ_p onto span($\phi(\mathbf{x}_i)$).*

Proof. Note that it suffices to show that the residual of the projection $\mu_p - \sum_j w_j\phi(\mathbf{x}_j)$ is orthogonal to $\phi(\mathbf{x}_i)$ for all i in $\mathcal{H}$. Recall that $w_i = \sum_j [\mathbf{K}^{-1}]_{ij}\mathbf{z}_j$, and $\mathbf{z}_i = \int k(\mathbf{x},\mathbf{x}_i)p(\mathbf{x})d(\mathbf{x})$. For an arbitrary index i,

$$\begin{aligned}&\langle \mu_p - \sum_j w_j\phi(\mathbf{x}_j), \phi(\mathbf{x}_i)\rangle_{\mathcal{H}} \\ &= \int k(\mathbf{x},\mathbf{x}_i)p(\mathbf{x})d(\mathbf{x}) - \langle \sum_j w_j\phi(\mathbf{x}_j), \phi(\mathbf{x}_i)\rangle_{\mathcal{H}} \\ &= \mathbf{z}_i - \langle \sum_j w_j\phi(\mathbf{x}_j), \phi(\mathbf{x}_i)\rangle_{\mathcal{H}} \\ &= \mathbf{z}_i - \sum_j w_j k(\mathbf{x}_j,\mathbf{x}_i) \\ &= \mathbf{z}_i - \sum_j\sum_t [\mathbf{K}^{-1}]_{tj}\mathbf{z}_t k(\mathbf{x}_j,\mathbf{x}_i) \\ &= \mathbf{z}_i - \sum_t \mathbf{z}_t \sum_j \mathbf{K}_{ji}[\mathbf{K}^{-1}]_{tj} \\ &= \mathbf{z}_i - \mathbf{z}_i,\end{aligned}$$

where the last equality follows by noting that $\sum_j \mathbf{K}_{ji}[\mathbf{K}^{-1}]_{tj}$ is inner product of row i of $\mathbf{K}$ and row t of $\mathbf{K}^{-1}$ which is 1 if $t = i$ and 0 otherwise. This completes the proof. □

Lemma 5 implies that given the selected points, the posterior evaluation is equivalent to the optimizing for $\mathbf{w}$ to minimize

$\text{MMD}(p,q)^2$. In other words, the weight optimization is a simple linear regression in the mapped space $\mathcal{F}$, and SBQ is equivalent to a greedy forward selection algorithm in $\mathcal{F}$.

We shall also make use of recent results in generalization of submodular functions. Let $\mathfrak{p}(\mathsf{S})$ be the power set of the set S.

Definition 6 (λ-weak submodular functions [6, 8])**.** *A set function $g : \mathfrak{p}([n]) \rightarrow \mathbb{R}$ is λ-weak submodular if $\exists \lambda > 0$ s.t. $\forall \mathsf{L}, \mathsf{S} \subset [n]$ $\mathsf{L} \cap \mathsf{S} = \emptyset$,*

$$\sum_{j \in \mathsf{S}} [g(\mathsf{L} \cup \{j\}) - g(\mathsf{L})] \geq \lambda \, [g(\mathsf{L} \cup \mathsf{S}) - g(\mathsf{L})]$$

Weak submodularity generalizes submodularity so that a greedy forward selection algorithm guarantees a $(1 - {}^{1}\!/_{e^{\lambda}})$ approximation for λ-weak submodular functions [8]. Standard submodular functions have a guarantee of $(1 - {}^{1}\!/_{e})$ [20]. Thus, submodular functions are 1-weak submodular. To provide guarantees for Algorithm 1, we show that the normalized set optimization function is $\frac{m}{M}$-weak submodular, where m, M depend on the spectrum of the kernel matrix.

Lemma 7. *[6] The linear regression function is $\frac{m}{M}$-weak submodular where m is the smallest $2r$ sparse eigenvalue and M is the largest $r+1$-sparse eigenvalues of the dot product matrix of the features.*

We note that Lemma 7 as proposed and proved by Das & Kempe [6] is for the euclidean space. However, their results directly translate to general RKHS as long as the RKHS is bounded, or the candidate atoms have bounded norm. Hence under additional assumptions of bounded norm, the proofs and results of Das & Kempe [6] directly translate to general RKHS. From Lemma 7 and recent results on weakly submodular functions, ([8, Corollary 1]), we get the following approximation guarantee for $g(\cdot)$ under the assumptions of Lemma 7.

$$g(\mathsf{S}_G) \geq \left(1 - \exp(-\frac{mk}{Mr})\right) g(\mathsf{S}^{\star}).$$

Setting $g(\mathsf{S}) = v(\phi) - v(\mathsf{S})$, we get the final result.