On Greedy Maximization of Entropy

Dravyansh Sharma CS1110214@CSE.IITD.AC.IN
IIT Delhi, New Delhi, India

Amit Deshpande AMITDESH@MICROSOFT.COM
Microsoft Research, Bangalore, India

Ashish Kapoor AKAPOOR@MICROSOFT.COM
Microsoft Research, Redmond, USA

Abstract

Submodular function maximization is one of the key problems that arise in many machine learning tasks. Greedy selection algorithms are the proven choice to solve such problems, where prior theoretical work guarantees $(1 - 1/e)$ approximation ratio. However, it has been empirically observed that greedy selection provides almost optimal solutions in practice. The main goal of this paper is to explore and answer why the greedy selection does significantly better than the theoretical guarantee of $(1 - 1/e)$. Applications include, but are not limited to, sensor selection tasks which use both entropy and mutual information as a maximization criteria. We give a theoretical justification for the nearly optimal approximation ratio via detailed analysis of the *curvature* of these objective functions for Gaussian RBF kernels.

1. Introduction

Consider a real-world scenario where the task is to sense a certain physical phenomenon of interest, e.g., temperature, in an area (Krause et al., 2008) with a limited number of sensors. Another scenario is selecting a subset of data points to be labeled from a large corpus, for the purposes of supervised learning (Settles, 2010). Similarly, the task could consist of determining what tests to run on a medical patient for diagnosing ailments (Kapoor & Horvitz, 2009). The key underlying question in all these scenarios is how to choose a subset of actions that would provide the most useful information pertaining to the task at hand.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

All of the above scenarios can be considered as a subset selection problem, where the goal is to determine which subset maximizes a given objective function defined over the subsets. Prior works have considered criteria such as mutual information and entropy, which make the objective function submodular. Traditionally, these submodular maximization problems are solved via greedy selection and often these prior works point to a result by Nemhauser et al. (Nemhauser et al., 1978) which guarantees that the greedy solution will be at least $(1 - 1/e)$ times the optimum. While there exist submodular functions for which the $(1-1/e)$ bound is tight, in several practical instances it has been observed that the greedy algorithm performs significantly better than $(1 - 1/e)$ times the optimum. For example, we reproduce the figure (see Figure 1) from Krause et al. (Krause et al., 2008), where the greedy method obtains an approximation ratio of over 0.95. While the greedy selection algorithms are popular in such subset selection problems, a better analysis explaining their empirical near-optimal performance is an unexplored direction to the best of our knowledge.

1.1. Our results

In this paper, we aim to answer why greedy selection results in nearly optimal solutions. We specifically focus on the popular kernels generated by Gaussian Radial Basis Functions (RBFs), and show that the greedy selection of points achieves an approximation ratio close to 1 that is much superior than the traditional guarantee of $(1 - 1/e)$. The key insight here is that the Gaussian RBF kernel matrices for *well-separated* points have a very dominant diagonal, making the submodular objective function close to linear (i.e., modular). Intuitively, it means that even though the objective function is submodular and it satisfies the diminishing returns property, the returns diminish only marginally even as we add more and more points.

Our main technical contribution is Theorem 5, where we

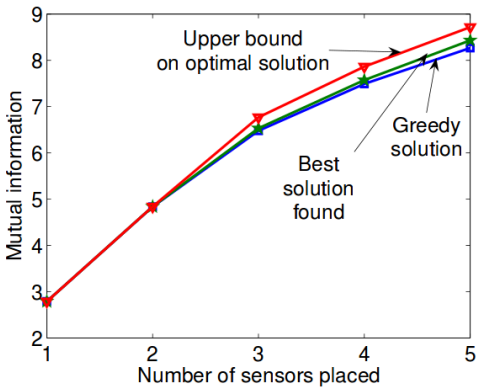

Figure 1. Plot depicting comparison of greedy algorithm with optimal solutions in (Krause et al., 2008). The data set consisted of 16 locations of the Intel Berkeley temperature data. As noted in the same work, the greedy algorithm is always within 95% of the optimal value, although the best known theoretical guarantee is about 63%.

bound the *curvature* of the subset selection criterion in terms of the bandwidth of the kernel, and the underlying dimensionality of the data points. An important consequence of Theorem 5 is that the curvature is nearly zero, which leads to an approximation ratio close to 1. A key challenge in our analysis is the non-monotonicity of these functions. Our work circumvents this using a monotone, submodular proxy to the objective function, that preserves the greedy selection choice at every step while having near-zero curvature. We provide empirical evidence that validates and highlights the key ideas in our result and their consequences.

1.2. Related work

Gaussian process (GP) model (Rasmussen & Williams, 2006) is a non-parametric generalization of linear regression, where the prediction error minimization reduces to a problem of maximum entropy sampling. The model is of fundamental significance to the problem of sensor placements, and is known to outperform classical models based on geometric assumptions, which turn out to be too strong in practice (Krause et al., 2008). Greedy algorithm is known to work well for the problem of maximum entropy sampling (Cressie, 1991) (Shewry & Wynn, 1987). However, for the problem of sensor placement, (Krause et al., 2008) propose a modified criterion of maximizing the mutual information of selected sensor locations, for which again a greedy procedure gives good approximation.

Both the entropy and the mutual information maximization problems are known to be NP-hard (Ko et al., 1995; Krause et al., 2008). However, the greedy selection gives an efficient, polynomial time algorithm with at least $(1-1/e)$ factor approximation to the optimum for both these objective functions, and more generally, for any non-negative, monotone, submodular function (Nemhauser et al., 1978). Conforti and Conuejols (Conforti & Cornuejols, 1984) gave a tighter analysis to prove that the greedy selection actually gives $(1-e^{-c})/c$ factor approximation when the objective function is monotone and has *curvature* c, which means that the approximation ratio tends to 1 as c tends to 0. Intuitively, it means that the approximation ratio of greedy selection tends to 1 as the function gets closer to being linear, as one would expect. This was the state of the art until recently, when Sviridenko, Vondrak and Ward showed that both a modified continuous greedy and local search give almost $(1-c/e)$ factor approximation, and this is essentially the best possible in the value oracle model (Sviridenko et al., 2015). However, both these algorithms, modified continuous greedy as well as local search, are computationally quite expensive and not as practical as the usual greedy selection. Note that *curvature* is defined only for monotone, submodular functions, whereas entropy and mutual information are submodular but not necessarily monotone.

Krause et al. (Krause et al., 2008) get around non-monotonicity of mutual information by showing that it is *approximately monotone* over small sets, under some reasonable assumptions on the underlying kernel as well as the discretization of the underlying space. This allows them to show an almost $(1-1/e)$ approximation. This guarantee holds for any non-negative, monotone, submodular function, and no better analysis is known even for special cases such as Gaussian RBF kernels. In this work, we prove near-optimal approximation guarantee for the maximum entropy sampling problem on Gaussian RBF kernels, and also establish *exact monotonicity* of mutual information over small subsets.

Finally we note that several fundamental problems in disparate domains can be effectively formulated as sensor selection problems. Among the more prominent ones are the problem of state estimation in linear dynamical systems (Shamaiah et al., 2010), target localization and tracking (Wang et al., 2004), (Wang et al., 2005), (Isler & Bajcsy, 2005), graphical models (Krause & Guestrin, 2012), coverage problems and mission assignment schemes (Rowaihy et al., 2007). Information-theoretic approaches, including both entropy and mutual-information based methods, have been widely acknowledged as prominent heuristics for sensor placement and other active learning problems.

1.3. Outline

The outline of our paper is as follows. In Section 2 we describe submodular functions and their key properties such as monotonicity and curvature. In Section 3 we describe

the algorithms or pseudo-codes for maximum entropy sampling as well as mutual information maximization. In Section 4 we provide a better analysis of greedy for the maximum entropy sampling on Gaussian RBF kernels. The key ideas here are

- A lower bound on the smallest eigenvalue of a Gaussian RBF kernel matrix, depending only on the intrinsic dimensionality of the data points and their minimum inter-point separation but *independent of the total number of points*.
- An upper bound on the Euclidean length of any row of a Gaussian RBF kernel matrix, with a similar dependence as above.

In Section 5 we prove *exact monotonicity* of mutual information over small subsets in Gaussian RBF kernels, improving upon the approximate monotonicity of (Krause et al., 2008).

2. Submodular functions and their properties

2.1. Submodularity, monotonicity, and curvature

We index a given set of n points by $[n] = \{1, 2, \ldots, n\}$ and denote the set of all subsets of $[n]$ by $2^{[n]}$.

Definition 1. A function $f : 2^{[n]} \to \mathbb{R}$ is *submodular* if $f(S \cup \{i\}) - f(S) \geq f(T \cup \{i\}) - f(T)$, for all $S \subseteq T$ and $i \notin T$.

In other words, submodular functions exhibit the property of diminishing returns.

Given a matrix $X \in \mathbb{R}^{n \times n}$, we use $X[S, T]$ to denote its $|S| \times |T|$ submatrix with row indices in $S \subseteq [n]$ and column indices in $T \subseteq [n]$. We denote the complement $[n] \setminus S$ by $\bar{S}$, and we abbreviate $X[S, \{i\}]$ as $X[S, i]$ and $X[S, [n] \setminus \{i\}]$ as $X[S, \bar{i}]$, respectively, for convenience.

Maximizing entropy and mutual information in Gaussian processes are known to be equivalent to maximizing certain functions of submatrices of given RBF kernels (Krause et al., 2008), so we directly define them by their corresponding linear algebraic problems.

Proposition 1. Given any symmetric, positive semidefinite matrix $X \in \mathbb{R}^{n \times n}$ the *entropy* $f(S) = \log\det(X[S, S])$ and the *mutual information* $F(S) = \log\det(X[S, S]) + \log\det\left(X[\bar{S}, \bar{S}]\right)$ are both submodular functions, where $X[S, S]$ denotes the $|S| \times |S|$ principal submatrix of X with row and column indices in $S \subseteq [n]$.

Proof. See Krause et al. (Krause et al., 2008). □

Definition 2. A submodular function $f : 2^{[n]} \to \mathbb{R}$ is *monotone* if, whenever $S \subseteq T$, we have $f(S) \leq f(T)$.

Now we show that if X has its smallest eigenvalue at least 1 then the entropy $\log\det(X[S, S])$ is monotone.

Proposition 2. Given any symmetric $X \in \mathbb{R}^{n \times n}$ with $\lambda_{\min}(X) \geq 1$, the function $f(S) = \log\det(X[S, S])$ is monotone.

Proof. For monotonicity, it suffices to show that $f(S) \leq f(S \cup \{i\})$, for all S and $i \notin S$. As in Proposition 1

$$\begin{aligned}
&f(S \cup \{i\}) - f(S) \\
&= \log\left(\frac{\det(X[S \cup \{i\}, S \cup \{i\}])}{\det(X[S, S])}\right) \\
&= -\log\left(X[S \cup \{i\}, S \cup \{i\}]^{-1}\right)_{ii} \quad \text{by Cramer's rule} \\
&\geq \log\lambda_{\min}(X[S \cup \{i\}, S \cup \{i\}]) \\
&\geq \log\lambda_{\min}(X) \\
&\geq 0 \qquad\qquad \text{using } \lambda_{\min}(X) \geq 1.
\end{aligned}$$

See (Strang, 2009) for Cramer's rule and the elementary fact that the smallest eigenvalue of a principal submatrix is at least the smallest eigenvalue of the bigger matrix. □

Thus, it is easy to make entropy function monotone just by scaling the matrix up so that its minimum eigenvalue is at least 1. For monotone, submodular functions, one can define their *curvature* as follows.

Definition 3. The *curvature* $c(f) \in [0, 1]$ of a monotone, submodular function $f : 2^{[n]} \to \mathbb{R}$ is defined as

$$c(f) = 1 - \min_{S \subsetneq [n], i \notin S} \frac{f(S \cup \{i\}) - f(S)}{f(\{i\}) - f(\emptyset)},$$

or equivalently, by submodularity, we can define

$$c(f) = 1 - \min_{i \in [n]} \frac{f([n]) - f([n] \setminus \{i\})}{f(\{i\}) - f(\emptyset)}.$$

Notice that submodularity along with curvature gives a tighter control on f as $f(\{i\}) - f(\emptyset) \geq f(S \cup \{i\}) - f(S) \geq (1 - c(f))\,(f(\{i\}) - f(\emptyset))$, for all $S \subsetneq [n]$ and $i \notin S$. Curvature $c(f) = 0$ means that the function f is linear. Therefore, small $c(f)$ is desirable and easier to handle.

Scaling the matrix up by α makes the new entropy $\log\det(X[S, S]) + \alpha\,|S|$, which also helps reduce the curvature.

Proposition 3. Let $f : 2^{[n]} \to \mathbb{R}$ be a monotone, submodular function, and let $g(S) = f(S) + \alpha\,|S|$, for some fixed $\alpha > 0$. Then g is also a monotone, submodular function with $c(g) < c(f)$.

Proof. Submodularity and monotonicity are easy to verify by observing $g(S \cup \{i\}) - g(S) = f(S \cup \{i\}) + \alpha\,|S \cup \{i\}| -$

$f(S) - \alpha |S| \geq f(S \cup \{i\}) - f(S)$.

$$\begin{aligned} c(g) &= 1 - \min_{i \in [n]} \frac{g([n]) - g([n] \setminus \{i\})}{g(\{i\}) - g(\emptyset)} \\ &= 1 - \min_{i \in [n]} \frac{f([n]) - f([n] \setminus \{i\}) + \alpha}{f(\{i\}) - f(\emptyset) + \alpha} \\ &\leq 1 - \min_{i \in [n]} \frac{f([n]) - f([n] \setminus \{i\})}{f(\{i\}) - f(\emptyset)} \\ &= c(f). \end{aligned}$$

□

The scaling trick does not work for mutual information because after scaling by α we get an additive $\alpha |S| + \alpha \left|\bar{S}\right| = \alpha n$, for all S. Now we try to bound the curvature of $\log\det(X[S,S])$ in terms of the row-lengths of X and $\lambda_{\min}(X)$. The reason being that the minimum eigenvalue and the row-lengths can each be bounded independently of n for Gaussian RBF kernels of well separated points (see Lemma 7 and Lemma 8).

Proposition 4. For any positive semidefinite matrix $X \in \mathbb{R}^{n \times n}$ with $\lambda_{\min}(X) \geq 1$, if $\max_{i \in [n]} \|X[\bar{i}, i]\| \leq \lambda_{\min}(X)$, then the curvature $c(f)$ of the function $f(S) = \log\det(X[S,S])$ can be upper bounded as follows.

$$c(f) \leq \frac{\lambda_{\min}(X)^{-2} \max_{i \in [n]} \|X[\bar{i}, i]\|^2}{\log \lambda_{\min}(X)}.$$

Proof. By the definition of curvature and the well-known identity for determinant of block matrices

$$\det \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \det(A) \det\left(D - CA^{-1}B\right)$$

if A is invertible, we have

$$\begin{aligned} c(f) &= 1 - \min_{i \in [n]} \frac{f([n]) - f([n] \setminus \{i\})}{f(\{i\}) - f(\emptyset)} \\ &= 1 - \min_{i \in [n]} \frac{\log\left(X[i,i] - X[\bar{i}, i]^T X[\bar{i}, \bar{i}]^{-1} X[\bar{i}, i]\right)}{\log X[i,i]} \\ &= \max_{i \in [n]} \frac{-\log\left(1 - \dfrac{X[\bar{i}, i]^T X[\bar{i}, \bar{i}]^{-1} X[\bar{i}, i]}{X[i,i]}\right)}{\log X[i,i]} \\ &\leq \max_{i \in [n]} \frac{-\log\left(1 - \dfrac{\lambda_{\min}(X[\bar{i}, \bar{i}])^{-1} \|X[\bar{i}, i]\|^2}{X[i,i]}\right)}{\log X[i,i]} \\ &\leq \frac{-\log\left(1 - \lambda_{\min}(X)^{-2} \max_{i \in [n]} \|X[\bar{i}, i]\|^2\right)}{\log \lambda_{\min}(X)} \\ &\leq \frac{\lambda_{\min}(X)^{-2} \max_{i \in [n]} \|X[\bar{i}, i]\|^2}{\log \lambda_{\min}(X)}. \end{aligned}$$

Algorithm 1 Greedy(f, k)

Initialize $S \leftarrow \emptyset$
for $t = 1$ to k **do**
  $i_{\max} \leftarrow \operatorname{argmax}_{i \notin S} f(S \cup \{i\}) - f(S)$
  $S \leftarrow S \cup \{i_{\max}\}$
end for
Output S

We have used $\lambda_{\min}(X[S,S]) \geq \lambda_{\min}(X) \geq 1$, for any $S \subseteq [n]$, which gives $\lambda_{\min}(X[\bar{i}, \bar{i}]) \geq \lambda_{\min}(X)$ as well as $X[i,i] \geq \lambda_{\min}(X)$. We also used $\max_{i \in [n]} \|X[\bar{i}, i]\| \leq \lambda_{\min}(X)$ to ensure that the expression inside log is nonnegative. □

Note that our bound is stronger than the $c(f) \leq 1 - 1/\lambda_{\min}$ bound mentioned in (Sviridenko et al., 2015).

3. Greedy algorithm and its variants

In each step, the greedy algorithm (see Algorithm 1) picks the element that maximizes the marginal gain. The approximation guarantees of $(1 - 1/e)$ by Nemhauser et al. (Nemhauser et al., 1978) and $(1 - e^{-c})/c$ by Conforti and Cornuejols (Conforti & Cornuejols, 1984) discussed in Subsection 1.2 also hold for monotone, submodular maximization over subsets of a predetermined size k.

In practice, it is possible to run a faster version of greedy selection while not losing on the approximation guarantee. Such algorithms include Lazy-Greedy (Krause et al., 2008) and Stochastic-Greedy (Mirzasoleiman et al., 2015). Our analysis of the curvature can be extended to these settings as well, and will be included in the full version.

4. Greedy maximization of entropy

Now we are ready to show that the greedy selection gives close to optimal solution for maximum entropy sampling on Gaussian RBF kernels satisfying a reasonable condition on the bandwidth parameter, the inter-point separation, and the dimension of the underlying space but *independent of the number of points* n.

Theorem 5. Let $X \in \mathbb{R}^{n \times n}$ be a Gaussian RBF kernel matrix, that is, its ij-th entry $X[i,j] = \exp\left(-\gamma \|x_i - x_j\|^2\right)$, for given points $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ and $\gamma > 0$. If the minimum separation $\delta = \min_{i \neq j} \|x_i - x_j\|$, the bandwidth parameter $\gamma > 0$ and the dimension d satisfy

$$d \geq \log\left(\frac{1}{\epsilon}\right) \quad \text{and} \quad \gamma\delta^2 \geq 10d \log d$$

then Greedy(f, k) on $f(S) = \log\det(X[S,S])$ outputs

$$f(S) \geq (1 - \epsilon) f(O) - k\epsilon,$$

where $O = \operatorname{argmax}_{|S|=k} f(S)$ is the optimal solution.

Proof. Consider $Y = 2\lambda_{\min}(X)^{-1} X$. $\lambda_{\min}(Y) = 2$ and $g(S) = \log\det(Y[S,S]) = f(S) + |S|\log(2\lambda_{\min}(X)^{-1})$ is a monotone, submodular function by Proposition 2. For a fixed cardinality constraint $|S| = k$, $g(S)$ is a fixed translation of $f(S)$ by $k\log\left(2\lambda_{\min}(X)^{-1}\right)$, and hence, optimal solutions coincide, giving $g(O) = f(O) + k\log\left(2\lambda_{\min}(X)^{-1}\right)$. Moreover, the choices made by Greedy(f,k) and Greedy(g,k) are the same. Therefore, we analyze Greedy(g,k) instead to infer about Greedy(f,k).

This helps in two ways. Firstly, even though f is not monotone and does not have a well-defined curvature in $[0,1]$, g is monotone with curvature $c(g) \in [0,1]$. Moreover, $c(g)$ can be bounded using Proposition 4 as follows.

$$\begin{aligned} c(g) &\le \frac{\lambda_{\min}(Y)^{-2}\max_{i\in[n]}\|Y[\bar{i},i]\|^2}{\log\lambda_{\min}(Y)} \\ &= \frac{\max_{i\in[n]}\|Y[\bar{i},i]\|^2}{4\log 2} \\ &= \frac{\lambda_{\min}(X)^{-2}\max_{i\in[n]}\|X[\bar{i},i]\|^2}{\log 2} \\ &\le C\exp\left(d\log(d\gamma\delta^2) - \gamma\delta^2\right), \end{aligned}$$

by Lemma 7 and Lemma 8, and the condition $\max_{i\in[n]}\|Y[\bar{i},i]\| = \lambda_{\min}(X)^{-1}\max_{i\in[n]}\|X[\bar{i},i]\| \le 2 = \lambda_{\min}(Y)$ of Proposition 4 is also satisfied. Here C is the constant from the big-Oh notation in Lemma 8. Notice that since $\gamma\delta^2 \ge 10d\log d$, we can bound $c(g) \le C\exp\left(d\log(d\gamma\delta^2) - \gamma\delta^2\right) \le \exp(-d\log d)$.

Secondly, adding $\alpha|S|$ to a monotone submodular function, for a fixed $\alpha > 0$, decreases its curvature (see Proposition 3). Therefore, if we could control the curvature in terms of α, we may be able to exploit better approximation guarantees for smaller curvature. That is exactly our strategy. Notice that we actually scale X up slightly beyond the $\lambda_{\min}(Y) \ge 1$ condition for monotonicity. Thus, if S is the output of Greedy(f,k) then

$$\begin{aligned} f(S) &= g(S) - k\log(2\lambda_{\min}(X)^{-1}) \\ &\ge \frac{1-e^{-c(g)}}{c(g)}\, g(O) - k\log(2\lambda_{\min}(X)^{-1}) \\ &= \frac{1-e^{-c(g)}}{c(g)}\left(f(O) + k\log(2\lambda_{\min}(X)^{-1})\right) \\ &\qquad - k\log(2\lambda_{\min}(X)^{-1}) \\ &= \frac{1-e^{-c(g)}}{c(g)}\, f(O) \\ &\qquad + \frac{1-c(g)-e^{-c(g)}}{c(g)}\, k\log(2\lambda_{\min}(X)^{-1}) \\ &\ge \frac{1-e^{-c(g)}}{c(g)}\, f(O) - c(g)\, k\log(2\lambda_{\min}(X)^{-1}) \\ &\ge \left(1 - \frac{c(g)}{2}\right) f(O) - c(g)kd\log d \\ &\ge (1-\exp(-d\log d))\, f(O) - \exp(-d\log d)kd\log d \\ &\ge (1-\epsilon)f(O) - \epsilon k, \end{aligned}$$

for large enough $d \ge \log(1/\epsilon)$. □

The main point here is that we can get an approximation guarantee very close to 1, and the approximation guarantee and the required bandwidth parameter do not depend on the total number of points n at all. It may be noted that the constant 10 in the requirement $\gamma\delta^2 \ge 10d\log d$ can be made close to 1 with the same analysis but done more carefully, which gives interesting values for minimum separation ($\exp\left(-\gamma\delta^2\right) \approx \exp\left(-d\log d\right)$) for the case of small d (for $d = 2$, the value is 0.25).

4.1. Condition numbers of RBF kernels

Schoenberg (Schoenberg, 1937) showed a striking result that if $x_1, x_2, \ldots, x_n$ are distinct points in a Hilbert space then the matrix $\left(\|x_i - x_j\|\right)_{ij}$ is invertible, which gave rise to radial basis interpolation methods. To implement such methods, it is important that this matrix be well-conditioned, in particular its eigenvalues be bounded away from 0. Keith Ball (Ball, 1992) showed such a bound, *independent of* n.

Proposition 6. For any points $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ with minimum separation δ, all the eigenvalues of the matrix $\left(\|x_i - x_j\|\right)_{ij}$ have absolute value at least $\Omega\left(\delta/\sqrt{d}\right)$.

This idea was later generalized to various RBF kernel matrices by Narcowich and Ward (Narcowich & Ward, 1992). The case of interest to us is that of Gaussian RBF kernels, although similar results hold for other RBF kernels too, e.g., exponential RBF kernel. Here we state a simple corollary of Theorem 2.3 from (Narcowich & Ward, 1992).

Lemma 7. Let $X \in \mathbb{R}^{n\times n}$ be a Gaussian RBF kernel matrix, that is, its ij-th entry $X[i,j] = \exp\left(-\gamma\|x_i - x_j\|^2\right)$, for given points $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ and $\gamma > 0$. If the minimum separation $\delta = \min_{i\ne j}\|x_i - x_j\|$ and satisfies $\gamma\delta^2 \ge 6d$ then $\lambda_{\min}(X) = \Omega\left(\exp(-\frac{d}{2}\log(d\gamma\delta^2))\right)$.

Proof. By Theorem 2.3 of Narcowich-Ward (Narcowich & Ward, 1992) mentioned above,

$$\lambda_{\min}(X) \ge C_d\,\gamma^{-d/2}\left(\frac{\delta}{2}\right)^{-d} e^{-\alpha^2\left(\frac{\delta}{2}\right)^{-2}\gamma^{-1}},$$

where

$$\alpha = 12\left(\frac{\pi\Gamma^2\left(1+\frac{d}{2}\right)}{9}\right)^{\frac{1}{d+1}} \approx d^{\frac{1}{d+1}}\left(\frac{d}{2e}\right)^{\frac{d}{d+1}} \approx d,$$

and

$$C_d = \frac{\alpha^2}{2^{d+1}\Gamma\left(1+\frac{d}{2}\right)} \approx d^{3/2}\left(\frac{2d}{e}\right)^{-\frac{d}{2}},$$

up to constants. Plugging in and simplifying gives

$$\lambda_{\min}(X) = \Omega\left(d^{3/2}\exp\left(-\frac{d}{2}\log\left(\frac{d\gamma\delta^2}{4}\right) - \frac{4d^2}{\gamma\delta^2}\right)\right),$$

which becomes $\Omega\left(\exp(-\frac{d}{2}\log(d\gamma\delta^2))\right)$ if $\gamma\delta^2 \geq 6d$. □

4.2. Off-diagonal row-lengths in Gaussian RBF kernels

Now we show that the off-diagonal entries of Gaussian RBF kernels decay rapidly if the points satisfy a minimum separation condition. This helps us bound the row-lengths of Gaussian RBF kernels *independent of* n.

Lemma 8. Let $X \in \mathbb{R}^{n\times n}$ be a Gaussian RBF kernel matrix, that is, its ij-th entry $X[i,j] = \exp\left(-\gamma\left\|x_i - x_j\right\|^2\right)$, for given points $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ and $\gamma > 0$. If the minimum separation $\delta = \min_{i\neq j}\|x_i - x_j\|$ and $\gamma\delta^2 \geq d$ then

$$\left\|X[\bar{i}, i]\right\|^2 = O\left(\exp(-\gamma\delta^2)\right), \quad \text{for all } i \in [n].$$

Proof. For any fixed i, define

$$C_t(i) = \left\{j \;:\; \frac{t\delta}{2} \leq \|x_i - x_j\| \leq \frac{(t+1)\delta}{2}\right\}.$$

Let $\mathbb{B}(x,r)$ denote the ball of radius r centered at x. For any $j \in C_j$, the ball $\mathbb{B}(x_j, \delta/2)$ lies outside $\mathbb{B}(x_i, (t-1)\delta/2)$ and inside $\mathbb{B}(x_i, (t+2)\delta/2)$. Moreover, the balls $\mathbb{B}(x_j, \delta/2)$ are all disjoint because δ is the minimum separation between all pairs. Thus,

$$\begin{aligned}|C_t(i)| &\leq \frac{\operatorname{vol}\left(\mathbb{B}\left(x_i, \frac{(t+2)\delta}{2}\right)\right) - \operatorname{vol}\left(\mathbb{B}\left(x_i, \frac{(t-1)\delta}{2}\right)\right)}{\operatorname{vol}\left(\mathbb{B}\left(x_j, \delta/2\right)\right)} \\ &\leq (t+2)^d - (t-1)^d \\ &\leq \exp(d\log t).\end{aligned}$$

We can bound $\left\|X[\bar{i}, i]\right\|^2$ as

$$\begin{aligned}\left\|X[\bar{i}, i]\right\|^2 &\leq \sum_{t=1}^{\infty}|C_t(i)|\exp\left(\frac{-\gamma t^2\delta^2}{4}\right) \\ &\leq \sum_{t=1}^{\infty}\exp\left(d\log t - \frac{\gamma t^2\delta^2}{4}\right) \\ &\leq \sum_{t=1}^{\infty}\exp\left(\gamma\delta^2\left(\log t - \frac{t^2}{4}\right)\right),\end{aligned}$$

using $\gamma\delta^2 \geq d$. Notice that the terms decay rapidly with t and can be upper bounded by a geometric progression $\exp\left(-t\gamma\delta^2\right)$, giving the final $\exp(-\gamma\delta^2)$ upper bound (up to constants). □

5. Extending to mutual information

Instead of using entropy for sensor placement, Guestrin-Krause-Singh (Krause et al., 2008) use mutual information, which is another submodular function.

Proposition 9. Given any symmetric, positive semidefinite matrix $X \in \mathbb{R}^{n\times n}$ the *mutual information* $F(S) = \log\det\left(X[S,S]\right) + \log\det\left(X[\bar{S},\bar{S}]\right)$ is submodular, where $X[S,S]$ denotes the $|S|\times|S|$ principal submatrix of X with row and column indices in S, and $\bar{S}$ denotes the complement $[n]\setminus S$.

Proof. See (Krause et al., 2008). □

Now we show that mutual information is monotone over sets of small size for Gaussian RBF kernels satisfying the same conditions we used in the previous section about entropy.

Proposition 10. Let $X \in \mathbb{R}^{n\times n}$ be a Gaussian RBF kernel matrix satisfying the conditions as in Theorem 5. Then the mutual information $F(S) = \log\det\left(X[S,S]\right) + \log\det\left(X[\bar{S},\bar{S}]\right)$ is monotone over sets of size $k \ll n$.

Proof. For any $S \subseteq [n]$ and $i \notin S$,

$$\begin{aligned}F(S\cup\{i\}) - F(S) = &\log\left(\frac{\det\left(X[S\cup\{i\}, S\cup\{i\}]\right)}{\det\left(X[S,S]\right)}\right) \\ &- \log\left(\frac{\det\left(X[\bar{S},\bar{S}]\right)}{\det\left(X[\bar{S}\setminus\{i\}]\right)}\right)\end{aligned}$$

However, we can show

$$\begin{aligned}&\frac{\det\left(X[S\cup\{i\}, S\cup\{i\}]\right)}{\det\left(X[S,S]\right)} \\ &= X[i,i] - X[S,i]^T X[S,S]^{-1} X[S,i] \\ &\geq 1 - \lambda_{\max}\left(X[S,S]^{-1}\right)\left\|X[S,i]\right\|^2 \\ &\geq 1 - \lambda_{\min}(X)^{-1}\left\|X[S,i]\right\|^2 \\ &\geq 1 - O\left(\exp\left(-\gamma\delta^2\right)\right).\end{aligned}$$

and

$$\begin{aligned}&\frac{\det\left(X[\bar{S},\bar{S}]\right)}{\det\left(X[\bar{S}\setminus\{i\}]\right)} = X[i,i] - \\ &\qquad X[\bar{S}\setminus\{i\}, i]^T X[\bar{S}\setminus\{i\}, \bar{S}\setminus\{i\}]^{-1} X[\bar{S}\setminus\{i\}, i] \\ &\leq 1 - \lambda_{\min}\left(X[\bar{S}\setminus\{i\}, \bar{S}\setminus\{i\}]^{-1}\right)\left\|X[\bar{S}\setminus\{i\}, i]\right\|^2 \\ &\leq 1 - \lambda_{\max}(X)^{-1}\,(n-k-1)\,\exp\left(-\gamma\Delta^2\right) \\ &\leq 1 - \exp\left(-\gamma\Delta^2\right),\end{aligned}$$

using $\lambda_{\max}(X) \leq \mathrm{tr}\,(X) = n$ and $k \ll n$, where $\Delta = \max_{i \neq j} \|x_i - x_j\|$ is the maximum separation.

Therefore,

$$F(S \cup \{i\}) - F(S) \geq \log\left(\frac{1 - O\left(\exp\left(-\gamma\delta^2\right)\right)}{1 - \exp\left(-\gamma\Delta^2\right)}\right) \geq 0.$$

□

The main difficulty in obtaining improved performance bounds for mutual information based greedy algorithm is the lack of monotonicity which makes it impossible to use the notion of curvature here. Our result on monotonicity for small k is a first step in removing this difficulty. Empirically mutual information is also known to exhibit the near-optimal performance of the greedy approach, and it should be interesting to theoretically establish the same under reasonable assumptions.

Table 1. Data sets for multivariate classification with real attributes, chosen to demonstrate our analysis.

DATA SET	# INSTANCES	# ATTRIBUTES
IRIS	147	4
SONAR MINES	111	60
CLOUD	1024	10

6. Experiments

In this section, we empirically verify the applicability of our analysis on three real world data sets as tabulated in Table 1. The data sets have been chosen to capture variation of both number of examples and number of features. They have been used to generate Gaussian kernels of different sizes and in different dimensions, over which our results have been studied. The experiments essentially consist of construction of these Gaussian kernels with appropriately chosen parameters and comparison of entropy of the greedy algorithm with the entropy of the optimal subset, computed for some small value of k. The experiments were repeated for mutual information based optimization, with exact optimality of the greedy algorithm for large range of parameters and small values of k.

6.1. Pre-processing

Data sets were first cleaned to remove duplicate instances[1] as presence of duplicates makes the smallest eigenvalue of the Gaussian kernel zero and our results cannot be studied ($\lambda_{min} = \delta = 0$). This step only reduced the size of the Iris data set by 3, and the other data sets remained unchanged. The features were then normalized to lie in the interval $[0, 1]$. These normalized features were finally used to generate Gaussian kernels with carefully chosen bandwidth parameter, γ.

[1]It is easy to argue that introduction of duplicates does not change either the greedy set or the optimal set of sensors, so long as total number of chosen sensors, k is less than the total number of distinct instances.

6.2. Variation of approximation ratio with bandwidth parameter

As in Lemma 7, $\gamma \geq \gamma_0 = \frac{d \log d}{\delta^2}$ suffices for the greedy algorithm to have near optimal approximation ratio. In the following we experimentally observe the transition of the greedy algorithm's approximation ratio to near optimal values as the bandwidth parameter is incremented to γ_0 from below. Figure 2(a) depicts these plots for the three chosen data sets. Note that in order to overlay the plots for different data sets on the same figure, the horizontal scale has been normalized for the different data sets to range over the interval $[\log \gamma_0 - 9, \log \gamma_0]$.

An interesting observation from the plot is that the transition to near-optimality is rather steep and occurs at approximately the same value of $\log \gamma$ for different data sets, of about $\log \gamma_0 - 4$. This indicates that our results qualitatively capture the requirement for near-optimal ratios quite accurately, and are probably tight up to constant factors in the worst case.

6.3. Variation of approximation ratio with scaling

Scaling increases the curvature, and hence the multiplicative term in the approximation inequality of Theorem 5, but also increases the negative additive term. The net effect can be seen as a sub-logarithmic increase in the approximation ratio, as in Figure 2(b). For the plots, we fix the value of $\log \gamma$ at $\log \gamma_0 - 6$, i.e. slightly before the transition to optimal. Two useful observations can be made from the plot. Scaling improves the approximation ratio of the Gaussian kernel, although rather slowly. Even an exponential increase in scaling seems to improve the ratio only in a sub-logarithmic fashion. Also, the qualitative nature of effect of scaling the matrix on the approximation ratio does not seem to vary significantly over the different data sets or choices of the bandwidth parameter.

We repeated the same sets of experiments for the mutual information criterion and interestingly observed that the greedy algorithm is able to find the optimal subset for the entire range of bandwidth parameter and the scaling parameter for each of the data sets.

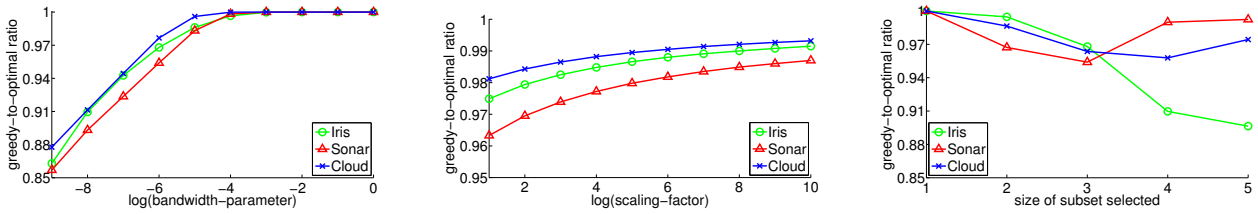

Figure 2. **(a)** Plot of approximation ratio of the greedy algorithm against logarithm of Gaussian kernel bandwidth parameter. Zero reference for the x-axis corresponds to the critical bandwidth ($\log \gamma_0$) for each data set. Size of subset selected, $k = 3$. **(b)** Plot of approximation ratio of the greedy algorithm against logarithm of factor with which the kernel is scaled. The bandwidth parameter of the kernel is fixed at $\log \gamma = \log \gamma_0 - 6$ for each data set. Size of subset selected, $k = 3$. **(c)** Plot of approximation ratio of the greedy algorithm against the size k of the subset selected. The bandwidth parameter is the same as in (b), and no scaling is used.

6.4. Variation of approximation ratio with k

The variation of the approximation ratio for the greedy algorithm with k, the size of subset to be selected, is relatively more complex. The decrease due to the small negative additive term in Theorem 5 is countered by increase in $f(O)$ with k to an indeterminate extent. Thus, for reasonably small k the approximation ratio does not decrease or vary significantly with increase in the number of sensors, as in Figure 2(c). For the experiments, the value of $\log \gamma$ was fixed at $\log \gamma_0 - 6$ and the scaling was fixed at unity.

6.5. Variation of $\lambda_{\min}$ with d

Finally we evaluate the lower bound of $\Omega\left(\exp\left(-d \log d\right)\right)$ on the minimum eigenvalue (Lemma 7) used in our analysis by using the Sonar data set. To generate the plot we randomly sample d features from the data set and determine the minimum eigenvalue for the corresponding Gaussian kernel. We observe that the blue curve representing our bound is a rather pessimistic bound for the minimum eigenvalue (shown in red in Figure 3), even though it suffices in our analysis to establish near-optimality.

This indicates that it might be possible to obtain significantly better lower bounds for decrease in the minimum eigenvalue with increasing dimensions for real world data sets. Plugging them in our analysis will give stronger approximation guarantees on the greedy performance, by allowing for weaker assumptions than those used here in order to establish near-optimality. it should be interesting to obtain an understanding of the properties that the data points satisfy which could be exploited to get such an improvement.

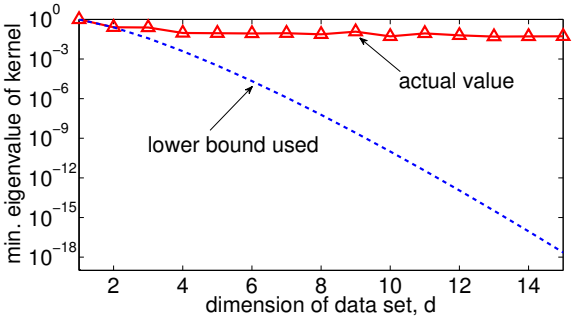

Figure 3. Plot of minimum eigenvalue of the Gaussian kernel obtained using subset of Sonar data set by randomly sampling d features and taking a minimum over a few iterations for smoothness. A logarithmic scale is used on the vertical axis for sake of clarity.

7. Conclusion and Future Work

The main result of this paper is to establish a theoretical basis for the empirically observed near optimal performance of the greedy algorithm for maximum entropy sampling, which has important applications in sensor placement. This is the first improvement over the general $(1 - 1/e)$ bound for submodular optimization, and holds for the extremely common case of Gaussian RBF kernels.

There is great scope for extension of this result to similar results for other kernels and also to other optimization criteria like mutual information. In fact, it seems that greedy performs close to the optimal result even with random kernels, with overwhelming probability.

References

Ball, Keith. Eigenvalues of euclidean distance matrices. *Journal of Approximation Theory*, 68:74–82, 1992.

Conforti, Michele and Cornuejols, Gerard. Submodular set functions, matroids and the greedy algorithm: Tight worst-case bounds and some generalizations of the rado-edmonds theorem. *Discrete Applied Mathematics*, 7(3): 251–274, 1984.

Cressie, Noel. *Statistics for spatial data*. John Wiley and Sons Inc., New York, NY, 1991.

Isler, Volkan and Bajcsy, Ruzena. The sensor selection problem for bounded uncertainty sensing models. In *Proceedings of the 4th international symposium on Information processing in sensor networks*, pp. 20. IEEE Press, 2005.

Kapoor, Ashish and Horvitz, Eric. Breaking boundaries: Active information acquisition across learning and diagnosis. *Advances in Neural Information Processing Systems (NIPS)*, 2009.

Ko, C., Lee, J., and Queyranne, M. An exact algorithm for maximum entropy sampling. *Operations Research*, 43 (4):684–691, 1995.

Krause, Andreas and Guestrin, Carlos E. Near-optimal nonmyopic value of information in graphical models. *arXiv preprint arXiv:1207.1394*, 2012.

Krause, Andreas, Singh, Ajit, and Guestrin, Carlos. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *The Journal of Machine Learning Research*, 9:235–284, 2008.

Mirzasoleiman, Baharan, Badanidiyuru, Ashwinkumar, Karbasi, Amin, Vondrak, Jan, and Krause, Andreas. Lazier than lazy greedy. In *Proc. Conference on Artificial Intelligence (AAAI)*, 2015.

Narcowich, Francis J. and Ward, Joseph D. Norm estimates for the inverses of a general class of scattered-data radial-function interpolation matrices. *Journal of Approximation Theory*, 69:84–109, 1992.

Nemhauser, George L, Wolsey, Laurence A, and Fisher, Marshall L. An analysis of approximations for maximizing submodular set functionsi. *Mathematical Programming*, 14(1):265–294, 1978.

Rasmussen, Carl Edward and Williams, Christopher K. I. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Rowaihy, Hosam, Eswaran, Sharanya, Johnson, Matthew, Verma, Dinesh, Bar-Noy, Amotz, Brown, Theodore, and La Porta, Thomas. A survey of sensor selection schemes in wireless sensor networks. In *Defense and Security Symposium*, pp. 65621A–65621A. International Society for Optics and Photonics, 2007.

Schoenberg, I. J. On certain metric spaces arising from euclidean space by a change of metric and their imbedding in hilbert space. *Annals of Math*, 38:787–793, 1937.

Settles, Burr. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010.

Shamaiah, Manohar, Banerjee, Siddhartha, and Vikalo, Haris. Greedy sensor selection: Leveraging submodularity. In *Decision and Control (CDC), 2010 49th IEEE Conference on*, pp. 2572–2577. IEEE, 2010.

Shewry, Michael C and Wynn, Henry P. Maximum entropy sampling. *Journal of Applied Statistics*, 14(2):165–170, 1987.

Strang, Gilbert. *Introduction to Linear Algebra (4th edition)*. Wellesley-Cambridge Press, 2009.

Sviridenko, Maxim, Vondrák, Jan, and Ward, Justin. Optimal approximation for submodular and supermodular optimization with bounded curvature. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1134–1148, 2015.

Wang, Hanbiao, Yao, Kung, Pottie, Greg, and Estrin, Deborah. Entropy-based sensor selection heuristic for target localization. In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, pp. 36–45. ACM, 2004.

Wang, Hanbiao, Yao, Kung, and Estrin, Deborah. Information-theoretic approaches for sensor selection and placement in sensor networks for target localization and tracking. *Communications and Networks, Journal of*, 7(4):438–449, 2005.