Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Active Search for Sparse Signals with Region Sensing

Yifei Ma
Carnegie Mellon University
Pittsburgh PA 15213, US
yifeim@cs.cmu.edu

Roman Garnett
Washington University in St. Louis
St. Louis, MO, USA
garnett@wustl.edu

Jeff Schneider
Carnegie Mellon University
Pittsburgh PA 15213, US
schneide@cs.cmu.edu

Abstract

Autonomous systems can be used to search for sparse signals in a large space; e.g., aerial robots can be deployed to localize threats, detect gas leaks, or respond to distress calls. Intuitively, search algorithms may increase efficiency by collecting aggregate measurements summarizing large contiguous regions. However, most existing search methods either ignore the possibility of such region observations (e.g., Bayesian optimization and multi-armed bandits) or make strong assumptions about the sensing mechanism that allow each measurement to arbitrarily encode all signals in the entire environment (e.g., compressive sensing). We propose an algorithm that actively collects data to search for sparse signals using only noisy measurements of the average values on rectangular regions (including single points), based on the greedy maximization of information gain. We analyze our algorithm in 1d and show that it requires $\widetilde{O}(n/\mu^2 + k^2)$ measurements to recover all of k signal locations with small Bayes error, where μ and n are the signal strength and the size of the search space, respectively. We also show that active designs can be fundamentally more efficient than passive designs with region sensing, contrasting with the results of Arias-Castro, Candes, and Davenport (2013). We demonstrate the empirical performance of our algorithm on a search problem using satellite image data and in high dimensions.

1 Introduction

Active search describes the problem where an agent is given a target to search for in an unknown environment and actively makes data-collection decisions so as to locate the target as quickly as possible. Examples of this setting include using aerial robots to detect gas leaks, radiation sources, and human survivors of disasters. The statistical principles for efficient designs of measurements date back to Gergonne (1815), but the growing trend to apply automated search systems in a variety of environments and with a variety of constraints has drawn much research attention recently, due to the need to address the disparate aspects of new applications.

One possibility in such active search scenarios we aim to explore, inspired by the robotic aerial search setting but with statistical insights that we hope to generalize, is the opportunity to take aggregate measurements that summarize large contiguous regions of space. For example, an aerial robot carrying a radiation sensor will sense a region of space whose area depends on its altitude. How can such a robot dynamically trade off the ability to make noisier observations of larger regions of space against making higher-fidelity measurements of smaller regions?

To simplify the discussion, we will limit such *region sensing* observations to reveal the average value of an underlying function on a rectangular region of space, corrupted by independent observation noise. Noisy binary search is a simple realization of active search using such an observation scheme. This mechanism turns out to be sufficiently informative in the cases that we analyze to offer insights into a variety of search problems.

The ability to make aggregate region measurements in noisy environments has rarely been considered in previous work. *Bayesian optimization,* which has been used for localization of sparse signals (Carpin et al. 2015; Ma et al. 2015; Hernández-Lobato, Hoffman, and Ghahramani 2014; Jones, Schonlau, and Welch 1998), usually considers only point measurements of an objective function. Notice that point observations can be considered in our framework if the allowed region sensing actions are constrained to be arbitrarily small. On the other extreme, *compressive sensing* (Donoho 2006; Candès and Wakin 2008; Wainwright 2009), considers scenarios where every measurement can reveal information about the entire environment through linear projection with arbitrary coefficients. This is not always a realistic assumption, as for example for an aerial robot, which can only sense its immediate vicinity. Between the two extremes, Jedynak, Frazier, and Sznitman (2012); Rajan et al. (2015); Haupt et al. (2009); Carpentier and Munos (2012); Abbasi-Yadkori (2012); Yue and Guestrin (2011) considered policies for search where observations can be made on any arbitrary subset of the search space, including discontiguous subsets, which is also often incompatible with the constraints in physical search systems.

Another assumption we make, common for example in compressive sensing, is *sparsity.* We assume that there are only a small number of strong signals in the environment; our goal is to recover these signals. Sparsity is necessary for the definition of active search problems; otherwise, for dense or weak signals, there is usually no better search approach than simply exhaustively mapping the entire space.

In addition to applicability in real search settings, sparsity has unique mathematical properties when considered alongside region sensing. In unconstrained sensing, Arias-Castro, Candes, and Davenport (2013) discovered a paradox that active compressive sensing (that is, the ability to adaptively select observations based on previously collected data) does not improve detection efficiency beyond logarithmic terms over random compressive sensing. This limitation is seen also when considering theoretical detection rates for active compressive sensing methods (Abbasi-Yadkori 2012; Carpentier and Munos 2012; Haupt et al. 2009). However, we show that active learning can in fact offer significant improvements in detection rates when observations are constrained to contiguous regions.

We propose an algorithm we call *Region Sensing Index* (RSI) that actively collects data to search for sparse signals using only noisy region sensing measurements. RSI is based on greedy maximization of information gain. Although information gain is a classic principle, we believe that its use in the recovery of sparse signals is novel and a good fit for robotic applications. We show that RSI uses $\tilde{O}(n/\mu^2 + k^2)$ measurements to recover all of k true signal locations with small Bayes error, where μ and n are the signal strength and the size of the search space, respectively (Theorem 3). The number of measurements with RSI is comparable with the rates offered by unconstrained compressive sensing, even though our constraints seem strong (i.e., region sensing loses all spatial resolution inside the region of measurement). Furthermore, we show that all passive designs under our contiguous region sensing constraint in $1d$ search spaces are fundamentally worse, with efficiency no better than sequential scanning of every point location, however strong the signals are. These results provide evidence to promote the use of and research into active methods.

1.1 Related Work

Arias-Castro, Candes, and Davenport (2013) proved that the minimax sample complexity[1] for any (i.e., potentially adaptive) algorithm to recover k sparse signal locations is at least $\Omega(n/\mu^2)$, analyzing the problem in terms of the mean-squared error in the recovery of the underlying signal values. The authors also showed that a passive *random* design, combined with a nontrivial inference algorithm, e.g., Lasso (Wainwright 2009) or the Dantzig selector (Candes and Tao 2007), can have similar recovery rates (up to $O(\log n)$ terms). This result was presented as a paradox, suggesting that the folk statement that active methods have better sample complexity is not always true. Here we show that active search can make a substantial difference in recovery rates when the measurements are subject to the physically plausible constraint of region sensing, especially if the physical space has low dimensions.

Malloy and Nowak (2014) presented the first *active* search algorithm that achieves the minimax sample complexity for general k. The algorithm is called Compressive Adaptive Sense and Search (CASS) and it can be adapted to region sensing in one-dimensional physical spaces. CASS directly extends bisection search, by allocating different sensing budgets to measurements at different bisection levels so as to minimize the cumulative error rates. However, CASS may fail if the repeated measurements of the same regions do not contain perfectly independent noise. It also has the limitation that it requires knowledge of the sensing budget *a-priori*, yet produces no signal localization results until the very last measurements at the lowest level. Our paper addresses these practical issues with a redesigned active search algorithm using the Bayesian approach, which compares evidence instead of blindly trust the assumptions, and we use Shannon-information criteria, which implies bisection search in noiseless one-sparse cases.

Braun, Pokutta, and Xie (2015) also used Shannon-information criteria for active search but did not analyze their sample complexity under noisy measurements. Jedynak, Frazier, and Sznitman (2012); Rajan et al. (2015) studied a similar search problem where the "regions" are relaxed to any unions of disjoint subsets.

[1]Sample complexity is equivalent to the number of measurements.

2 Problem Formulation

Consider a discrete space that is the Cartesian product of one-dimensional grids, $\mathcal{X} = \prod_{i=1}^{d}[n_i]$; $[n] = \{1, \ldots, n\}$. Let $n = \prod n_i$ be the total number of points in $\mathcal{X}$ (here the product symbol is the arithmetic rather than the Cartesian product). We presume there is a latent real-valued nonnegative vector $\boldsymbol{\beta} \in \mathbb{R}^n$ that represents the vector of true signals at all locations in $\mathcal{X}$. We further assume that $\boldsymbol{\beta}$ is sparse: it has value $\mu > 0$ on $k \ll n$ locations in $\mathcal{X}$ and has value 0 elsewhere. We consider making observations related to $\boldsymbol{\beta}$ through rectangular region sensing measurements, defined by

$$y_t = \mathbf{x}_t^\top \boldsymbol{\beta} + \varepsilon_t, \text{ s.t. } x_{tj} = w_t 1_{j \in A_t}, \; \varepsilon_t \sim \mathcal{N}(0, \sigma_t^2). \quad (1)$$

Here $\mathbf{x}_t \in \mathbb{R}^n$ is a sensing vector that has support on $A_t \subseteq \mathcal{X}$, a rectangular subset of $\mathcal{X}$. We assume that the sensing vector has equal weight w_t across its support. The resulting measurement, y_t, is equal to the mean value of $\boldsymbol{\beta}$ on A_t corrupted by independent Gaussian noise with variance σ_t^2. Note that selecting A_t suffices to specify the measurement location.

In $1d$ search environments, A_t may be any interval of $[n]$, and the corresponding design takes the form $\mathbf{x}_t = (0, \ldots, 0, w_t, \ldots, w_t, 0, \ldots, 0)^\top$. In higher search dimensions, we consider only regions that are contained in a hierarchical spacial pyramid (Lazebnik, Schmid, and Ponce 2006), i.e., a sequence of increasingly finer grid boxes with dyadic side lengths to cover the space at multiple resolutions.

Our goal is to choose a sequence of designs $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^T$ so as to discover the support of $\boldsymbol{\beta}$ with high confidence. Given a particular confidence, we will measure sample complexity by assuming $\|\mathbf{x}_t\|_2 = 1$ and $\sigma_t \equiv 1$ for each measurement and count the total number of measurements required to achieve that confidence, T. Letting $\|\mathbf{x}_t\|_2 = 1$ implies $w_t = 1/\sqrt{\|\mathbf{x}_t\|_0}$, which can be seen as a *relaxed* notion of the region average, because the signal strength of a region measurement, which is μw_t, still decreases as the region size $\|\mathbf{x}_t\|_0$ increases.

Algorithm 1 Region Sensing Index (RSI)

Require: $\pi_0(k, n, \mu), T$ or ϵ, and the unknown $\boldsymbol{\beta}^*$
Ensure: $\hat{S}_t$ // (5)
for $t = 1, 2, \ldots$ **do**
  pick $\mathbf{x}_t = \arg\max_{\mathbf{x}\in\mathcal{X}} I(\boldsymbol{\beta}; y \mid \mathbf{x}, \pi_{t-1})$ // (3)&(4)
  observe $y_t = \mathbf{x}_t^\top \boldsymbol{\beta}^* + \varepsilon_t$
  update $\pi_t(\boldsymbol{\beta}) \propto \pi_{t-1}(\boldsymbol{\beta}) p(y_t \mid \boldsymbol{\beta}, \mathbf{x}_{t-1})$ // (2)
  find $(\bar{\epsilon}_t, \hat{S}_t) = \arg\min_{|\hat{S}|=k} \frac{1}{k}\mathbb{E}\big[|\hat{S}\Delta S| \mid \pi_t\big]$ // (5)
  break if $t \geq T$ or $\bar{\epsilon}_t < \epsilon$, if either is defined

The measure of T is made to be comparable with another common choice of sample complexity, the Frobenius norm of the entire design $\|\mathbf{X}\|_F^2$, when the rows of $\mathbf{X}$ are normalized (Arias-Castro, Candes, and Davenport 2013). However, the normalization is often overlooked in classical compressive sensing, which allows algorithms to cheat in region sensing by making an enormous number of measurements of small weight and changing the sensing locations frequently. Another measure of complexity is to measure both $\|\mathbf{X}\|_F^2$ and the number of location changes simultaneously (Malloy and Nowak 2014). However, our discretized counting of measurements is conceptually simpler.

Our analysis is Bayesian and we will analyze performance in expectation, with prior $\boldsymbol{\beta} \sim \pi_0(\boldsymbol{\beta})$, a uniform distribution on the model class, $\mathcal{S}_\mu\binom{n}{k}$, which includes all k-sparse models with μ signal strength among n locations (i.e., it has $\binom{n}{k}$ possible outcomes). The Bayes risk will be measured by the expected Delta loss, $\bar{\epsilon}_T = \frac{1}{k}\mathbb{E}|S\Delta\hat{S}_T|$, where $\hat{S}_T$ is the best estimator of the k signal locations after T measurements and Δ is the symmetric difference operator on a pair of sets.

3 Proposed Methods

We note that region sensing loses all spatial resolution inside the region of measurement. Here we borrow ideas from noisy binary search, which has a similar property, and use information gain (IG) to drive the observation process. We name our algorithm *Region Sensing Index* (RSI, Algorithm 1). Like other active learning algorithms, RSI is a combination of an *inference* subroutine that constantly updates the distribution of $\boldsymbol{\beta}$ using the collected data and a *design* subroutine that chooses the next region to sense based on the latest information from the inference subroutine.

The inference subroutine. We use exact Bayesian inference with a uniform prior $\pi_0(\boldsymbol{\beta})$ on the model class $\mathcal{S}_\mu\binom{n}{k}$. Denote the outcome of the first t measurements as $\mathcal{D}_t = \{(\mathbf{x}_\tau, y_\tau) : 1 \leq \tau \leq t\}$. Even though $\mathcal{D}_t$ contains a dependent sequence of data collections, where $\mathbf{x}_\tau$ depends on $\mathcal{D}_{\tau-1}, \forall\tau$, Bayesian inference decomposes into a series of efficient updates:

$$\begin{aligned}\pi(\boldsymbol{\beta} \mid \mathcal{D}_t) &\propto \pi(\boldsymbol{\beta}) p(\mathcal{D}_t \mid \boldsymbol{\beta}) \\ &= \pi_0(\boldsymbol{\beta}) \prod_{\tau=1}^t \big(p(\mathbf{x}_\tau \mid \mathcal{D}_{\tau-1}) p(y_\tau \mid \boldsymbol{\beta}, \mathbf{x}_\tau)\big) \\ &\propto \pi_0(\boldsymbol{\beta}) \prod_{\tau=1}^t p(y_\tau \mid \boldsymbol{\beta}, \mathbf{x}_\tau), \quad (2)\end{aligned}$$

where $p(\mathbf{x}_\tau \mid \mathcal{D}_{\tau-1})$ is the design without knowledge of the true $\boldsymbol{\beta}$ and thus dropped. Define $\pi_t(\boldsymbol{\beta}) = \pi(\boldsymbol{\beta} \mid \mathcal{D}_t)$; the updates have the form $\pi_t(\boldsymbol{\beta}) \propto \pi_{t-1}(\boldsymbol{\beta}) p(y_t \mid \boldsymbol{\beta}, \mathbf{x}_t) = \pi_{t-1}(\boldsymbol{\beta})\phi(y_t - \mathbf{x}_t^\top\boldsymbol{\beta})$, where ϕ is the standard normal pdf.

The design subroutine. The next sensing vector, $\mathbf{x}_{t+1} \in \mathcal{X}$, is chosen to maximize the IG:

$$I(\boldsymbol{\beta}; y \mid \mathbf{x}, \pi_t) = H(y \mid \mathbf{x}, \pi_t) - \mathbb{E}\big[H(y \mid \mathbf{x}, \boldsymbol{\beta}) \mid \pi_t\big], \quad (3)$$

which is the difference between the entropy of the marginal distribution, $p(y \mid \mathbf{x}, \pi_t) = \int \phi(y - \mathbf{x}^\top\boldsymbol{\beta})\pi_t(\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{\beta}$, and the expected entropy of the conditional distribution, $p(y \mid \boldsymbol{\beta}; \mathbf{x}) = \phi(y - \mathbf{x}^\top\boldsymbol{\beta})$. The latter, i.e., the conditional distribution for any realization of $\boldsymbol{\beta}$, has fixed entropy: $\log\sqrt{2\pi e}$. Meanwhile, the marginal entropy has no closed-form solutions; instead, we use numerical integration.

The numerical integration is rather straightforward, because the marginal *density* function is analytical. From now on, we will assume that $(\mathbf{x}, A, a, w_\mathbf{x})$ correspond to the same design (its sensing vector, its locations, its region size, and its sensing weight per coordinate, respectively). Define two new variables, $\lambda = \mu w_\mathbf{x} (= \mu/\sqrt{a})$ and $\gamma = \mathbf{x}^\top\boldsymbol{\beta}/\lambda$, and one new parameter $\mathbf{p} = (p_0, \ldots, p_k)^\top$ in (4). The goal is to change the variable of the integration for the marginal *density* function of y to:

$$\begin{aligned}p(y \mid \mathbf{x}, \pi_t) &= \int \pi_t(\boldsymbol{\beta})\phi(y - \mathbf{x}^\top\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{\beta} \\ &= \sum_{c=0}^k p_c\,\phi(y - c\lambda) = p(y \mid \lambda, \mathbf{p}), \\ \text{where} \quad p_c &= \Pr(\gamma = c) = \sum_{\boldsymbol{\beta}:\mathbf{x}^\top\boldsymbol{\beta}=c\lambda} \pi_t(\boldsymbol{\beta}). \quad (4)\end{aligned}$$

Notice, γ only has a finite number of choices: $\gamma = |A \cap S| \in \{0, \ldots, k\}$, where S is the nonzero support of $\boldsymbol{\beta}$, because both $\mathbf{x}$ and $\boldsymbol{\beta}$ are constant on their respective supports ($x_j = w_\mathbf{x}, \forall j \in A$ and $\beta_j = \mu, \forall j \in S$). We then numerically evaluate $H(y \mid \mathbf{x}, \pi_t) = H(y \mid \lambda, \mathbf{p})$ with the obtained (4).

The Bayes estimator of signal locations. We pick the k-sparse set $\hat{S}_T$ to minimize the posterior risk:

$$\min_{|\hat{S}|=k} \frac{1}{k}\mathbb{E}\big[|\hat{S}\Delta S| \mid \pi_T\big] = \frac{1}{k}\sum_{\hat{\imath}\in\hat{S}} \mathbb{E}\big(1_{\{\beta_{\hat{\imath}}=0\}} \mid \pi_T\big), \quad (5)$$

where $\beta_{\hat{\imath}}$ is the $\hat{\imath}$-th element of $\boldsymbol{\beta}$. In other words, RSI picks the top k locations where the posterior marginal expectation is the largest. When $k = 1$, this is equivalent to picking $\hat{\boldsymbol{\beta}}_T = \arg\max \pi_T(\boldsymbol{\beta})$. Otherwise, (5) yields the smallest Bayes risk $\bar{\epsilon}(\mathcal{D}_T)$ given any collected data $\mathcal{D}_T$.

3.1 Accelerations

In practice, holding $\binom{n}{k}$ models in memory can be infeasible if k is large, we can instead recover the support of $\boldsymbol{\beta}$ element-wise by repeatedly applying RSI assuming $k = 1$. After the posterior distribution $\pi_t(\boldsymbol{\beta}^{(1)})$ converges to a point-mass distribution at the most-likely one-sparse model with sufficient confidence, we report its location and move on by

[2] In real world experiments, we additionally estimate $\hat{\mu}_{\hat{\jmath}}$ using a point measurement on the inferred signal location for better modeling.

Table 1: Conditions and conclusions for sample complexity.

Design Type	Region Sensing	Algorithm	Prior for Bayes Risk	Min T to Guarantee $\bar{\epsilon}_T = \frac{1}{k}\mathbb{E}\lvert S\Delta\hat{S}_T\rvert \le \epsilon$	Sample Complexity*
passive	yes	(any)	π_0 ($\mu \to \infty$)	$T \ge \frac{n}{2}(1 - \frac{n-1}{n-k}\epsilon)$ (Theorem 1)	$\Theta(n)$
		Point sensing		$T \le n(1 - \frac{n-1}{n-k}\epsilon)$ (Corollary A.2)	
active	no	(any)	$\tilde{\pi}_0$	$T \ge \frac{4n}{\mu^2}(1-\epsilon)^2$ (Theorem 2)	$\Omega(\frac{n}{\mu^2})^\dagger$
	yes	CASS [2014]	max risk (incl. π_0)	$T \le 20\frac{n}{\mu^2}\log(\frac{8k}{\epsilon}) + 2k\log_2(\frac{n}{k})$	$\tilde{O}(\frac{n}{\mu^2}+k)^\ddagger$
		RSI (ours)	π_0	$\bar{T}_\epsilon \le 50(\frac{n}{\mu^2}+\frac{k^2}{9})\log_2(\frac{2}{\epsilon})\log(\frac{n}{\epsilon})$ (Theorem 3)	$\tilde{O}(\frac{n}{\mu^2}+k^2)^\ddagger$

* Assume $\epsilon = O(1)$ and $k \ll n$. † Shown for unconstrained sensing; binary search requires $\Omega(\log_2(n)+k)$ additional measurements. ‡ $\log(n)$ terms are left out. $\bar{T}_\epsilon$ is defined differently; see Section 4.2 for details.

Algorithm 2 Region Sensing Index-Any-k **(RSI-A)**

Require: n, μ, ϵ, and the unknown $\boldsymbol{\beta}^*$
Ensure: $\hat{S}$
initialize $\hat{S} = \emptyset$, $\hat{\boldsymbol{\beta}} = \mathbf{0}$
for $k = 1, 2, \ldots,$ **do**
  infer $\pi_0(\boldsymbol{\beta}^{(k)}) \propto \prod_{\tau=1}^{t} p(y_\tau \mid \boldsymbol{\beta}^{(k)} + \hat{\boldsymbol{\beta}}, \mathbf{x}_\tau)$, $\forall \boldsymbol{\beta}^{(k)} \in \{\mu \mathbf{1}_j : j \notin \hat{S}\}$
  call $\hat{S}^{(k)} = \text{RSI}\ (\pi_0, \epsilon, \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})$
  aggregate $\hat{S} = \cup_{c \le k} \hat{S}^{(c)}$ and $\hat{\boldsymbol{\beta}} = \sum_{\hat{j} \in \hat{S}} \hat{\mu}_{\hat{j}} \mathbf{1}_{\hat{j}}$.[2]

removing the reported point from the search and recomputing the posterior distributions using the uniform prior, $\pi_0(\boldsymbol{\beta}^{(2)})$, on the new class, $\mathcal{S}_\mu\binom{n-1}{1}$.

We call this alternative algorithm *Region Sensing Index-Any-k* (RSI-A, Algorithm 2) and use it in our simulations so that the computational cost is no longer exponential in k. Notice, our analysis is for the unmodified RSI; the statistical disadvantage of RSI-A is no more than $O(k)$, multiplicatively.

When implementing RSI-A, we also avoid unnecessary numerical integration (3), if the region is guaranteed to have inferior IG, indicated by its $\mathbf{p}$ vector (4), which is easier to compute. We use the fact that $I(\gamma; y \mid \mathbf{p}, \lambda)$ with fixed $\lambda > 0$ is concave in the probability simplex $\Delta^k = \{\mathbf{p} \in [0,1]^{k+1} : \mathbf{p}^\top \mathbf{1} = 1\}$. Under $k = 1$ approximation, the region whose marginal probability $p_1 = \sum_{\mathbf{x}^\top \boldsymbol{\beta} > 0} \pi(\boldsymbol{\beta})$ is closest to 0.5 will provably have the largest IG among all regions of the same size. Thus, we find the region with the highest IG in two steps: (1) compare the p_1 value for all regions for every region size and (2) evaluate the IG of only these regions with the best p_1 values (closest to 0.5) in their region sizes.

4 Theoretical Analysis in 1D

The analysis is cleanest when the search space is 1d, where the regions can be any integer intervals that subset $[1, n]$. Without loss of generality (WLOG), assume n is a multiple of k and $n \ge 2k$. Our goal is to find the smallest number of measurements, T, to guarantee a small Bayes risk $\bar{\epsilon}_T = \frac{1}{k}\mathbb{E}|S\Delta\hat{S}_T| \le \epsilon$. Table 1 summarizes our analysis. The sample complexity is best appreciated assuming $\mu \gg 1$, $k \ll n$, and $\epsilon = \mathcal{O}(1)$. A typical choice is $\epsilon = 1/2$, i.e., the number of measurements to guarantee that half of the signal support can be recovered on average.

4.1 Baseline Results

Here we provide lower bounds on sample complexity. We show that under region-sensing constraints, all passive methods require $T \ge \Omega(n)$ measurements and active methods require $T \ge \Omega(n/\mu^2 + k)$. When $\mu \gg 1$, active methods have significant potential for improvement over passive methods using region sensing, which contradicts with the view in unconstrained compressive sensing by Arias-Castro, Candes, and Davenport (2013); Soni and Haupt (2014).

Theorem 1 (Limits of any passive methods using region sensing)**.** *Assume $\boldsymbol{\beta}$ has prior π_0 (uniform random on $\mathcal{S}_\mu\binom{n}{k}$). Any passive method with T noiseless region measurements on 1d must incur Bayes risk $\bar{\epsilon}_T \ge \frac{n-k}{n-1}(1 - \frac{2T}{n})$. To guarantee $\bar{\epsilon}_T \le \epsilon$, $T \ge \frac{n}{2}(1 - \frac{n-1}{n-k}\epsilon)$ is required.*

The proof is due to model identifiability, neglecting observation noise. More details can be found in the appendix. It applies to any $\mu \ge 0$ and particularly $\mu \to \infty$.

Theorem 2 (Limits of any methods, (Arias-Castro, Candes, and Davenport 2013))**.** *Assume $\boldsymbol{\beta}$ has a slightly different prior, $\tilde{\pi}_0$, that includes each location in $\mathcal{X}$ in the support of $\boldsymbol{\beta}$ independently with probability k/n. Any method (including active and non-region-sensing) must have $\bar{\epsilon}_T \ge 1 - \frac{\mu}{2}\sqrt{T/n}$. To guarantee $\bar{\epsilon}_T \le \epsilon$, $T \ge \frac{4n}{\mu^2}(1-\epsilon)^2$ is required.*

The proof can be found under Theorem 3 of (Arias-Castro, Candes, and Davenport 2013). Arias-Castro, Candes, and Davenport (2013) gave a minimax risk with similar terms by modifying $\tilde{\pi}_0$ to a *least favorable prior* on all models that are at most k-sparse. However, we only study Bayes risk for technical convenience.

When using Theorem 2 for reference, notice the difference between $\tilde{\pi}_0$ and π_0 that the former additionally treats the sparsity to be a random variable $\tilde{k}$ with expectation k. From concentration inequalities, $|\tilde{k} - k| \leq O(\sqrt{k})$, with high probability. While $\tilde{k}$ and k are not directly comparable, Theorem 2 is still a useful baseline. Under region-sensing constraints, the number of measurements must also be at least $\Omega(k)$ to allow visits to most of the nonzero locations at least once, in a nontrivial draw of S where the signals are separated.

With respect to Theorem 1, the point sensing or any non-repeating region sensing will achieve the optimal sample complexity (up to constant factors, see Appendix A for more details). For Theorem 2, the CASS method published by Malloy and Nowak (2014) for active sensing with region constraints[3] acheives a nearly optimal rate in theory. Table 1 contains a detailed summary of the sample complexities of several algorithms, including our own.

4.2 Main Result

For technical convenience, we directly express our main result in terms of the expected number of measurement that are actually taken so as to realize $\bar{\epsilon}(D_{\mathcal{T}}) \leq \epsilon$ for a given threshold ϵ in an experiment. Taking $\mathcal{T} = \mathcal{T}_\epsilon$ as a random variable, the expected number of actual measurements is different from the pre-determined sampling budget that an algorithm fully consumes to guarantee a desirable averaged risk (see Section 4.1). However, it is a comparable alternative in Bayesian analysis, used by e.g., Lai and Robbins (1985); Kaufmann, Korda, and Munos (2012). When the objective is constant $\epsilon = \mathcal{O}(1)$, our result implies a deterministic budget requirement of the same order of complexity, $T \leq \epsilon_2^{-1}\mathbb{E}\mathcal{T}_{\epsilon_2}$, where $\epsilon_2 = \frac{\epsilon}{2}$, by direct application of Markov's inequality.

Theorem 3 (Sample complexity of RSI)**.** *In active search for k sparse signals with strength μ in 1d physical space of size $n \geq 2k$ (WLOG, assume n is a multiple of k), given any $\epsilon > 0$ as tolerance of posterior Bayes risk,* RSI *using region sensing has bounded expected number of actual measurements,*

$$\bar{T}_\epsilon = \mathbb{E}[\min\{\mathcal{T} : \bar{\epsilon}(D_{\mathcal{T}}) \leq \epsilon\}]$$
$$\leq 50\Big(\frac{n}{\mu^2} + \frac{k^2}{9}\Big)\log_2\Big(\frac{2}{\epsilon}\Big)\log\Big(\frac{n}{\epsilon}\Big) = \tilde{O}\Big(\frac{n}{\mu^2} + k^2\Big), \quad (6)$$

where the expectation is taken over the prior distribution and sensing outcomes.

4.3 Proof Sketch

The proof for Theorem 3 hinges on an observation that the information gain (IG) where RSI makes measurements is consistently large, before active search terminates with minimal Bayes risk. For example, the IG of any measurement in binary search with $k = 1$ and noiseless observations is always $O(\log(2))$. However, IG is harder to approximate when the observations are noisy. Therefore, we first show an intuitive lower bound for IG. Recall notations from (4).

[3]The original result in Malloy and Nowak (2014) is stronger; it considers the maximum probability of support recovery mistakes, $P(S \neq \hat{S}) \leq \delta$, for any S that are k-sparse and any signals with at least μ strength.

Proposition 4. *The IG score of a region sensing design has lower bounds with respect to its design parameters $(\lambda, \mathbf{p})$, as*

$$I(\gamma; y \mid \lambda, \mathbf{p}) \geq 2q_c\bar{q}_c\Big(2\Phi\Big(\frac{\lambda}{2}\Big) - 1\Big)^2$$
$$\geq \frac{1}{12}\min\{q_c, \bar{q}_c\}\min\{\lambda^2, 3^2\}, \quad \forall 1 \leq c \leq k, \quad (7)$$

where $q_c = \Pr(\gamma \geq c) = \sum_{\kappa \geq c} p_\kappa$, $\bar{q}_c = 1 - q_c$, and $\Phi(u)$ is the standard normal cdf.

The proof uses Pinsker's inequality and is given in Section B in the appendix. Notice using the common choice of Jensen's inequality will give bounds in the opposite direction. To formalizes our observation that the IG is bounded:

Lemma 5. *WLOG, assume n is a multiple of k and $n \geq 2k$. At any step, if the current Bayes risk $\bar{\epsilon}(D) > \epsilon$, we can always find a region A of size at most $\frac{n}{k}$, such that $\lambda^2 \geq \frac{\mu^2}{a} = \frac{k\mu^2}{n}$ and $\frac{\epsilon}{2} \leq \mathbb{E}[\gamma \mid D] \leq 1 - \frac{\epsilon}{2}$ (we call this* Condition E*), which further yields*

$$I(\gamma; y \mid \lambda, \mathbf{p}) \geq I^*_\epsilon = \frac{\epsilon}{25k}\min\Big\{\frac{k^2\mu^2}{n}, 3^2\Big\}. \quad (8)$$

The way to find the region A that satisfies *Condition E* is given in Lemma B.5 in the appendix. The reason that Condition E is sufficient for (8) can be derived from Proposition 4 for $k = 1$ and Lemma B.6 in the appendix for $k > 1$. □

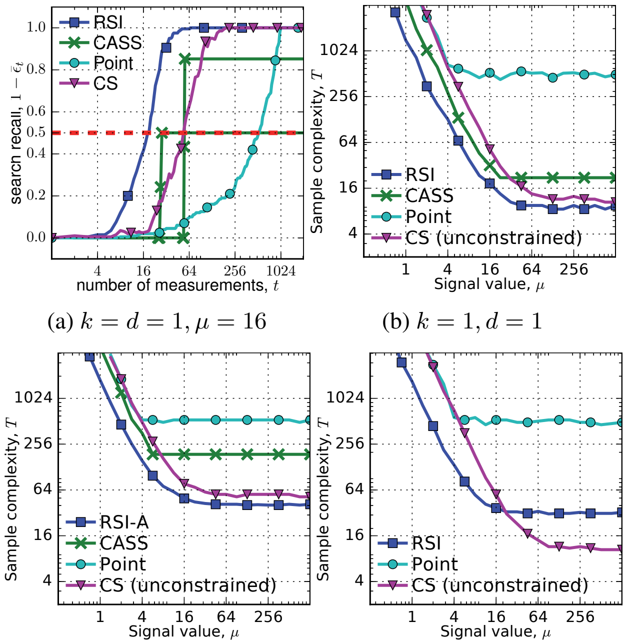
Eq (8) shows the minimum decrease in the model entropy in expectation after each measurement, starting from the maximum entropy of a uniform prior distribution, $k\log(n)$. However, the posterior entropy can never be negative, which implies a bound on the expected number of times that (8) can be applied, i.e. the expected number of measurements to reach ϵ Bayes risk is $\frac{25\log(n)}{\epsilon}(\frac{n}{\mu^2} + \frac{k^2}{9})$. Lemma D.5 in the appendix shows some additional improvements to obtain the logarithmic dependency of ϵ in Theorem 3.

5 Simulation Studies

We evaluated **RSI** or its approximation **RSI-A** when $k > 1$. Other baseline algorithms include:

- **CASS** (compressive adaptive sense and search) (Malloy and Nowak 2014): a branch-and-bound algorithm that traverses the region hierarchy from top to down using pre-allocated budgets per level. We count each $\mathbf{x}_i$ as $\|\mathbf{x}_i\|_2^2$ region sensing measurements (rounded up to the next integer).
- **Point** sensing: a passive design that uses exhaustive point measurements on all locations.
- **CS** (compressive sensing) (Donoho 2006): a non-region-sensing design that draws $\mathbf{x}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and rescales $\|\mathbf{x}_t\|_2^2$ to 1. CS then solves a convex optimization problem to infer the nonzero signals, by minimizing $\sum_t \|y_t - \mathbf{x}_t^\top\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$ s.t. $\boldsymbol{\beta} \geq 0$, where λ is chosen to produce exactly k nonzero coefficients using the Lasso regularization path.

We picked $n = 1024$ and various k (sparsity) and d (the dimension of the physical space) annotated below the plots. In the $d = 5$ case, we chose the region space to be the Cartesian product of $[4]^5$ and allowed regions from a spatial pyramid (Lazebnik, Schmid, and Ponce 2006) of granularity



(a) $k = d = 1, \mu = 16$ (b) $k = 1, d = 1$

(c) $k = 10, d = 1$ (d) $k = 1, d = 5$

Figure 1: Sensing efficiency. (a) Average search progresses as more measurements are taken. (b-d) Minimum sample size T in different SNR scenarios to guarantee $\bar{\epsilon}_T < 0.5$.

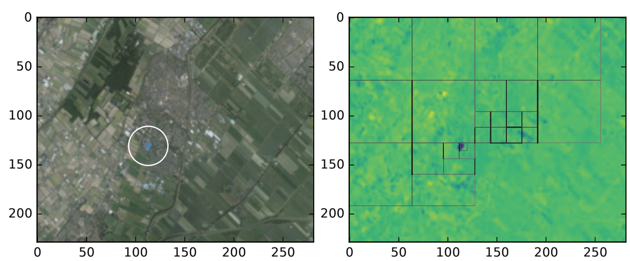
4^5, 2^5, and 1^5. Each method was run with 200 repetitions to find its average performance.

Figure 1(a) compares the recall rates of the algorithms as they progressed in a 1d search for a single true signal of strength $\mu = 16$. RSI was the most efficient, finding the correct location in 50% of the cases with as few as $T = 20$ measurements. CASS was comparable only at the step points when all the allocated budgets were used, due to its rather rigid designs. We drew multiple curves for CASS to reflect this fact; the turning points were at $T = 28$ and 56 for $\epsilon = 0.5$ and 0.85, respectively. CS was less effective compared with CASS with equal budgets (e.g., $\|\mathbf{X}\|_F^2 = 52 > 28$ for $\epsilon = 0.5$) which agrees with the analysis in Arias-Castro, Candes, and Davenport (2013). Point sensing was the least efficient, using $T = n/2 = 512$ measurements, which was worse than the other methods by a factor of $\tilde{\Omega}(\mu^2)$ (ignoring logarithmic terms). Notice, due to non-identifiability, any passive designs would have equal or worse rates.

Figure 1(b) extends the comparison on the full spectrum of SNR, $1/4 < \mu < 1024$, showing the minimum number of measurements T to guarantee constant Bayes risk $\bar{\epsilon}_T < 0.5$. RSI led the comparison, showing a sample complexity of $\tilde{O}(n/\mu^2)$ when μ is small and $\tilde{O}(1)$ when μ is large. CASS also had a similar trend. CS ignores the region sensing constraints and was inferior to RSI. Notice CS also has a minimum sample complexity, but in order to meet the incoherence conditions for Lasso sparsistency (Candes and Tao 2007; Wainwright 2009; Raskutti, Wainwright, and Yu 2010), the rank of the covariance matrix of the measurements $\mathbf{X}_S^\top \mathbf{X}_S$ must be at least k. Point sensing and other passive region sensing would always require at least $\Omega(n)$ measurements regardless of μ. Figure 1(c-d) show similar conclusions with other choices of k and d. The number of measurements was largely unaffected by $k > 1$ if μ is low, which supports the first term of Theorem 3, which is $\tilde{O}(n/\mu^2)$. Comparisons between CS and RSI in high dimensions ($d > 1$) depend on how region constraints are defined. In our high-dimensional simulations, the region choices were rather limiting for RSI, giving more advantage to the unconstrained CS when μ is large.

6 Real World Datasets

Region sensing was intended to address the problems of real robotic search. Here, we took satellite images like Figure 2a and used natural blue pixels, e.g., the blue roof which we circled near the lower left of the center of the image, as a simulated target of interest. These experiments directly simulate search and rescue in open areas based on life jacket colors or communication signals and also share similarities with gas leaks or radiation detection, where real data is usually sensitive or expensive. Many assumptions were violated in these experiments, e.g., the noise was not iid and the target was a collection of neighboring pixels. For the purpose of more accurate modeling of the actual measurement powers at each region level, we used statistics from the real data to model $\mu w(a)$ and $\sigma(a)$ as functions of the region size $a (= \|\mathbf{x}\|_0)$.



(a) Target blue pixels (circled) (b) RSI search pattern

Figure 2: Demo active search on real images.

Figure 2b shows in the background the actual scalar observations, affinely transformed from the original RGB values to filter out the target blue color. The foreground contains the rectangular regions of measurement, sequentially decided by RSI after observing the average values in previous regions. Feasible region choices were contained in a spatial pyramid (Lazebnik, Schmid, and Ponce 2006). RSI behaved similarly to sequential scanning at the optimal altitude except for occasional bisections into subregions.

By comparing the IG of all feasible regions, RSI usually decides to (a) sense the next region in space when the previous outcome is low, (b) investigate the subregions when the last parent region yields a large outcome (we disallow repeated actions for the lack of noise-independence,) or (c) back out from an investigation if the subsequent measurements yield low outcomes. Option (c) demonstrates the ability of error recovery, which is our advantage to CASS thanks to Bayesian modeling. The search in Figure 2b ended after 36 measurements, whereas the image contained 36 000 pixel points.

Figure 3 compares the performances on 221 image patches of 512×512 pixels, cropped from National Agriculture Imagery Program (NAIP).[4] The other algorithms for comparison include random (point), CS, and CASS*. Here, CASS* is a modified CASS method where each measurement can only be taken once, because repeated measurements yield the same outcome. To fully represent CASS*, in addition to choosing k by the true sparsity, we added fixed choices of $k = 64$ and 512, yielding three different curves.

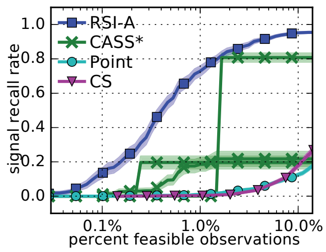


Figure 3: Performances on 221 NAIP image crops.

RSI achieved the best performance, finding on average 60% blue pixels with as few as 1700 measurements (0.5% of the total number of feasible observations). CASS* performance highly depended on the parameter choices and produced results only near the end of the experiment. CS did poorly, probably due to the fact the signals were not iid (a blue object can contain multiple pixels).

7 Discussions

Region sensing is a new setting motivated by robotic search operations where we also found statistical insights to contrast with the unconstrained sensing in Arias-Castro, Candes, and Davenport (2013). RSI performs near-optimally in 1d search domains and fundamentally faster than passive sensing. In higher dimensions, the analysis may be harder, especially for passive baselines. The number of subregions generated by intersecting the measurement regions may be harder to count, unless measurement regions are restricted to grid regions in a spatial pyramid (such that any pair of regions is either nested or disjoint). We also want to establish frequentist analysis in the future. Finally, it is interesting to generalize the measurement model beyond taking the average value of a single region at a time.

Acknowledgments

This work is partially supported by the DARPA grant FA87501220324, National Science Foundation under Award Number IIA-1355406, and ARPA-E TERRA-REF award DE-AR0000594. We also appreciate suggestions and discussions from Aarti Singh, Ying Yang, and Yining Wang.

References

Abbasi-Yadkori, Y. 2012. Online-to-confidence-set conversions and application to sparse stochastic bandits.

Arias-Castro, E.; Candes, E. J.; and Davenport, M. 2013. On the fundamental limits of adaptive sensing. *Information Theory, IEEE Transactions on*.

Braun, G.; Pokutta, S.; and Xie, Y. 2015. Info-greedy sequential adaptive compressed sensing. *IEEE Journal of Selected Topics in Signal Processing* 9(4):601–611.

Candes, E., and Tao, T. 2007. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*.

Candès, E. J., and Wakin, M. B. 2008. An introduction to compressive sampling. *Signal Processing Magazine, IEEE* 25(2):21–30.

Carpentier, A., and Munos, R. 2012. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In *AISTATS*, volume 22, 190–198.

Carpin, M.; Rosati, S.; Khan, M. E.; and Rimoldi, B. 2015. Uavs using bayesian optimization to locate wifi devices. *arXiv preprint arXiv:1510.03592*.

Donoho, D. L. 2006. Compressed sensing. *Information Theory, IEEE Transactions on* 52(4):1289–1306.

Gergonne, J. D. 1815. Application de la méthode des moindre quarrés a l'interpolation des suites. *Annales des Math Pures et Appl*.

Haupt, J. D.; Baraniuk, R. G.; Castro, R. M.; and Nowak, R. D. 2009. Compressive distilled sensing: Sparse recovery using adaptivity in compressive measurements. In *Signals, Systems and Computers*. IEEE.

Hernández-Lobato, J. M.; Hoffman, M. W.; and Ghahramani, Z. 2014. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in Neural Information Processing Systems*.

Jedynak, B.; Frazier, P. I.; and Sznitman, R. 2012. Twenty questions with noise: Bayes optimal policies for entropy loss. *Journal of Applied Probability* 49(1):114–136.

Jones, D. R.; Schonlau, M.; and Welch, W. J. 1998. Efficient global optimization of expensive black-box functions. *Journal of Global optimization* 13(4):455–492.

Kaufmann, E.; Korda, N.; and Munos, R. 2012. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory*, 199–213. Springer.

Lai, T. L., and Robbins, H. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* 6(1):4–22.

Lazebnik, S.; Schmid, C.; and Ponce, J. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition*. IEEE.

Ma, Y.; Sutherland, D.; Garnett, R.; and Schneider, J. 2015. Active pointillistic pattern search. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*.

Malloy, M. L., and Nowak, R. D. 2014. Near-optimal adaptive compressed sensing. *Information Theory, IEEE Transactions on*.

Rajan, P.; Han, W.; Sznitman, R.; Frazier, P.; and Jedynak, B. 2015. Bayesian multiple target localization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*.

Raskutti, G.; Wainwright, M. J.; and Yu, B. 2010. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*.

Soni, A., and Haupt, J. 2014. On the fundamental limits of recovering tree sparse vectors from noisy linear measurements. *Information Theory, IEEE Transactions on*.

Wainwright, M. J. 2009. Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on*.

Yue, Y., and Guestrin, C. 2011. Linear submodular bandits and their application to diversified retrieval. In *Advances in Neural Information Processing Systems*, 2483–2491.

[4] https://lta.cr.usgs.gov/node/300